UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
CURSO ENGENHARIA DE PRODUÇÃO ELÉTRICA

Lucas Bergo Dias

Implementação de um guia para aplicação de *machine learning* em séries temporais multivariadas

Florianópolis
2022

Lucas Bergo Dias

Implementação de um guia para aplicação de *machine learning* em séries temporais multivariadas

Trabalho Conclusão do Curso de Graduação em Engenharia de Produção Elétrica do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito para a obtenção do título em Engenharia, área Eletricidade, habilitação Engenharia de Produção Elétrica.
Orientador: Prof. Eduardo Ferreira da Silva, Dr.

Florianópolis
2022

Ficha de identificação da obra

Dias, Lucas Bergo
Implementação de um guia para aplicação de machine learning em séries temporais multivariadas / Lucas Bergo Dias ; orientador, Eduardo Ferreira da Silva, 2022.
100 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Elétrica, Florianópolis, 2022.

Inclui referências.

1. Engenharia de Produção Elétrica. 2. Séries temporais. 3. Aprendizado de máquina. 4. Previsão de dados. I. Silva, Eduardo Ferreira da. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Elétrica. III. Título.

Lucas Bergo Dias

Título: Implementação de um guia para aplicação de *machine learning* em séries temporais multivariadas

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de "Engenheiro Eletricista com habilitação em Engenharia de Produção" e aprovado em sua forma final pelo Curso de Engenharia de Produção Elétrica.

Florianópolis, 24 de fevereiro de 2022.

Documento assinado digitalmente
Monica Maria Mendes Luna
Data: 22/03/2022 11:58:13-0300
CPF: 314.104.993-91
Verifique as assinaturas em https://v.ufsc.br

_____

Prof.(a) Mônica Maria Mendes Luna, Dr.(a)
Coordenadora do Curso

Banca Examinadora:

Documento assinado digitalmente
Eduardo Ferreira da Silva
Data: 21/03/2022 15:00:54-0300
CPF: 790.231.347-53
Verifique as assinaturas em https://v.ufsc.br

_____

Prof. Eduardo Ferreira da Silva, Dr.
Orientador
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Guilherme Ernani Vieira
Data: 22/03/2022 10:00:19-0300
CPF: 888.311.759-04
Verifique as assinaturas em https://v.ufsc.br

_____

Prof. Guilherme Ernani Vieira, Dr.
Avaliador
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Mauricio Uriona Maldonado
Data: 21/03/2022 16:20:36-0300
CPF: 010.972.939-04
Verifique as assinaturas em https://v.ufsc.br

_____

Prof. Maurício Uriona Maldonado, Dr.
Avaliador
Universidade Federal de Santa Catarina

Este trabalho é dedicado aos meus pais, por sempre me proporcionarem a melhor educação possível e incentivarem meus estudos, ao meu irmão por sempre me passar seus conhecimentos, e ao meu orientador Professor Eduardo, por toda dedicação e apoio no desenvolvimento deste trabalho.

AGRADECIMENTOS

Agradeço aos meus pais, Nilton e Maria Elizabeth, que sempre me incentivaram e me forneceram todas as condições para que eu me dedicasse aos estudos. Agradeço ao meu irmão, Igor, que desde pequeno nunca mediu esforços para me ensinar e passar seus conhecimentos.

Ao meu orientador, Professor Eduardo Ferreira da Silva, por me apresentar e incentivar o estudo sobre o tema abordado no trabalho e abrir meus olhos para a importância desse tema no contexto global de hoje. Também por me apoiar e ter paciência, pela participação ativa e suporte fornecidos durante o desenvolvimento.

Um agradecimento especial a todos os professores que me ensinaram no decorrer da graduação, que me incentivaram e permitiram que extraísse o melhor desempenho e potencial no meu processo de formação ao longo do curso.

Aos meus amigos e colegas de curso, com quem convivi intensamente meus anos de graduação, em especial, Erik Spanhol Lind e Álisson Carlos da Silva, obrigado pelo companheirismo e pela troca de experiências que me permitiram crescer não só como pessoa, mas também como graduando.

Aos meus amigos, Diogo Norio Romão Iwata, Bruno Pires, Matheus Berkenbrock Nedel e Lyon Sojfer, por todos os momentos compartilhados durante esse período e toda parceria. Também gostaria de agradecer a Ana Carla Reisdorfer pelo suporte e companheirismo nesses anos de graduação.

À Universidade Federal de Santa Catarina, que foi essencial no meu processo de formação profissional, por proporcionar todo o ambiente necessário para meu desenvolvimento e oportunidades oferecidas.

Por fim, agradeço a todos que de alguma forma contribuíram direta ou indiretamente para a minha pesquisa e para a minha formação. Muito obrigado!

"O caminho mais certo para o sucesso é tentar apenas uma vez mais." (Thomas Edison)

RESUMO

Com o avanço da tecnologia e automação dos processos, a quantidade de dados armazenados aumentou significativamente, contribuindo para que a tomada de decisão em empresas possa ser apoiada por previsões cada vez melhores. Algumas dessas previsões usam séries históricas de dados, também conhecidas como séries temporais, que requerem modelos adequados para capturar padrões de comportamento associados ao momento em que cada evento ocorre, como por exemplo, um ciclo sazonal de alta em um dado mês ou uma tendência de queda ao passar dos anos. Contudo, apesar do avanço na área de aprendizado de máquina (ML) ter popularizado o uso de modelos de previsão cada vez mais sofisticados, poucos destes modelos se adequam ao contexto das séries temporais. E aqueles que se adequam são altamente customizados como as Redes Neurais Recorrentes, cuja aplicação e configuração não é trivial. Essa é uma das razões pelas quais modelos tradicionais são largamente escolhidos, como por exemplo os modelos de Holt, Holt-Winters e ARIMA, mesmo quando o resultado não impressiona. Esta simplicidade na sua aplicação faz com que a utilização desses modelos seja bem mais frequente do que a maioria dos modelos de ML. Uma das grandes desvantagens dos modelos tradicionais é a dificuldade em explorar a correlação da série com outras variáveis independentes, justamente o ponto forte de grande parte dos algoritmos de ML. Modelos de ML possuem uma excelente capacidade de extrair relações de dependência entre variáveis independentes e a variável dependente. Dito isso, este trabalho visa apresentar um guia para realização de previsão de séries temporais usando modelos genéricos de ML multivariados, que permitirá aos modelos de ML identificar relações temporais na variável de interesse, além da inclusão de novas variáveis independentes, o que não é possível nos modelos clássicos. Assim, uma gama de modelos de ML que incorporam o estado da arte em previsão poderão ser melhor aplicados no contexto de previsões em séries temporais. O desenvolvimento é apresentado na forma de um guia que pode servir de referência para o leitor na preparação dos dados de uma série temporal a serem inseridos em modelos de ML, fazendo com que o modelo reconheça as características típicas de uma série temporal e com isso realize previsões mais assertivas. Finalmente, os resultados apresentados deixam claro como a criação de *features* específicas em séries temporais pode viabilizar a aplicação de algoritmos de ML e influenciar significativamente sua performance.

Palavras-chave: Séries Temporais. Aprendizado de Máquina. Previsão de Dados.

ABSTRACT

With the advancement of technology and process automation, the amount of stored data has increased significantly, contributing to the decision making in companies to be supported by increasingly better forecasts. Some of these forecasts use historical data series, also known as time series, which require adequate models to capture patterns of behavior associated with the moment when each event occurs, such as a seasonal upward cycle in a given month or a trend of fall over the years. However, despite advances in machine learning (ML) having popularized the use of increasingly sophisticated forecasting models, few of these models are suitable for the context of time series. And those that fit are highly customized such as recurrent neural networks, whose application and configuration is not trivial. This is one of the reasons why traditional models are widely chosen, such as Holt, Holt-Winters and ARIMA models. In addition to the simplicity in its application, even when the result is not impressive, the use of these models is more frequent than most ML models. One of the major disadvantages of traditional models is the difficulty in exploring the correlation of the series with other independent variables, precisely the strength of most ML algorithms. ML models have an excellent ability to extract dependency relationships between independent variables and the dependent variable. That said, this work aims to present a guide for performing time series forecasting using generic multivariate ML models, which will allow ML models to identify temporal relationships in the variable of interest, in addition to the inclusion of new independent variables, which is not possible on classic models. Thus, a range of ML models that incorporate the state of the art in forecasting can be better applied in the context of time series forecasts. The development is presented in the form of a guide that can serve as a reference for the reader in the preparation of data from a time series to be inserted into ML models, making the model recognize the characteristics of this series and thus make more assertive predictions. Finally, the results presented make it clear how the creation of specific features in time series can enable the application of ML algorithms and significantly influence their performance.

Keywords: Time Series. Machine Learning. Data Forecast.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT Associação Brasileira de Normas Técnicas

ABEPRO Associação Brasileira de Engenharia de Produção

ML Machine Learning ou Aprendizado Máquina

XGBoost Extreme Gradient Boosting

ARIMA Autoregressive Integrated Moving Average

RMSE Root Mean Squared Error

SUMÁRIO

1 INTRODUÇÃO

Os objetivos deste capítulo são de contextualizar o tema do presente trabalho, justificar a escolha e importância deste, definir o objetivo geral, definir os objetivos específicos e apresentar sua delimitação.

1.1 CONTEXTUALIZAÇÃO

As técnicas de previsão tornaram-se grandes aliadas para garantir a estabilidade dos negócios, fazendo com que a coleta e análise de dados seja um assunto muito estudado pelas organizações. Como resultado, é notável que grande parte dos planejamentos de uma empresa competitiva são desenvolvidos a partir de métodos de previsões. Dessa forma, a previsão é crucial para a estabilidade dos negócios, pois auxilia no planejamento e desenvolvimento das estratégias organizacionais, ajuda na identificação de prioridades, auxilia também nas decisões de alocação de recursos, expansões de capacidade, na redução dos estoques e no planejamento da produção (SOUZA; CAMARGO, 2004).

Com o aumento da competitividade, a necessidade de se antecipar das diversidades e cenários ameaçadores tornou-se ainda mais crucial a utilização de métodos de previsões como ferramenta para auxiliar os colaboradores tomar decisões, aumentando assim as chances das organizações se manterem bem posicionadas no mercado (MANLY, 1997).

Juntamente com o aumento na busca por técnicas de previsões mais eficientes, houve um aumento na complexidade dos problemas de previsão, trazendo assim a necessidade de modelos preditivos cada vez mais complexos e robustos (JOHNSON; WICHERN, 1999). De maneira geral, modelos mais complexos trazem consigo mais dados e necessitam de mais ajustes para serem treinados. Tendo isso em vista, há um correspondente aumento do custo computacional exigido para que se realize essas análises de dados e previsões assertivas (BITTENCOURT, 2006).

Considerando a crescente variedade e complexidade dos problemas de previsão, um tema muito discutido para resolver problemas dessa natureza é a inteligência artificial, que vem sendo cada vez mais usada para a solução de problemas em variados setores da economia, tecnologia e até mesmo na área da pesquisa científica. Nessa temática, uma área de estudo com destaque especial é o *machine learning* (ML), ou em português, aprendizado de máquina (CARVALHO, 2021).

O aprendizado de máquina é uma subárea de atuação da inteligência artificial que tem por objetivo o desenvolvimento de programas computacionais com a capacidade de aprender a executar tarefas com sua própria experiência, utilizando-se de um conjunto de dados passados (FACELI; LORENA; GAMA; DE CARVALHO, 2019). Trata-se de uma área de pesquisa que envolve as áreas de: probabilidade e estatística, inteligência artificial, tecnologia da informação, etc.

Modelos de *machine learning* estão cada vez mais sendo testados em previsões, em especial com séries temporais, sendo mais uma opção para se avaliar o comportamento de uma variável ao longo do tempo. É esperado que os modelos usem dados passados de uma variável para capturar elementos desse tipo de série, tais como: tendências, sazonalidades, padrões de repetição e aleatoriedade (WEIGEND, GERSHENFELD, 1994).

Entretanto, a aplicação dos modelos de *machine learning* com séries temporais pode ser bastante desafiadora, principalmente, no que diz respeito a preparação dos dados para que o modelo consiga reconhecer e levar em consideração as características desse tipo de série, observando as relações temporais dos dados ao realizar previsões (SOUZA, CAMARGO, 2004).

Dito isso, esse trabalho desenvolverá um guia para que o leitor possa ter uma visão geral dos passos a serem seguidos no que diz respeito a aplicação de modelos de *machine learning*, principalmente no tratamento e preparação dos dados, para tornar possível a utilização desses modelos em previsão de séries temporais.

1.2 OBJETIVOS

Nas seções abaixo estão descritos o objetivo geral e os objetivos específicos deste estudo.

1.2.1 Objetivo geral

- Realizar um estudo de caso que sirva de guia na realização de previsões de séries temporais com base em modelos genéricos de *machine learning*.

1.2.2 Objetivos específicos

- Escolher o conjunto de dados que será utilizado no estudo de caso;
- Apresentar os modelos clássicos de previsão;
- Descrever os passos na forma de um guia de como os dados devem ser preparados para serem submetidos a um modelo genérico de *machine learning*;
- Selecionar pelo menos dois modelos de *machine learning*;
- Analisar os resultados e comparar com os modelos tradicionais de previsão;
- Destacar as vantagens da abordagem apresentada.

1.3 DELIMITAÇÕES

O assunto previsão de séries temporais é extenso, e existem várias abordagens para se realizar previsão a partir deste tipo de dados. Contudo, o presente estudo abordará a comparação entre os modelos clássicos, que utilizam somente os dados da série, e os modelos de *machine learning* (ML); e, mais especificamente, como preparar os dados de uma série temporal para que possam ser inseridos nos diversos modelos multivariados de *machine learning*, e como estas etapas são realizadas. Dessa forma, vários modelos de ML poderosos que antes eram ineficientes para este tipo de dado, agora serão passíveis de serem empregados, mantendo sua boa performance.

Com a infinidade de modelos de *machine learning* existentes na atualidade, vários modelos não serão contemplados pelo guia aqui apresentado. Entretanto, a sua adaptação poderá ser realizada sem maiores dificuldades.

1.4 LIMITAÇÕES

Pelos mesmos motivos expostos no item 1.3, serão escolhidos somente dois modelos de ML para exemplificar os passos do guia, um mais simples e outro de estrutura mais complexa. Desse modo, não postulando que esses serão os melhores modelos de ML a serem usados com séries temporais.

1.5 JUSTIFICATIVA

Quando se fala em previsão de séries temporais usando modelos de *machine learning*, é observado na literatura que o modelo de redes neurais recorrentes é uma das melhores opções a ser utilizada (HOCHREITER S, SCHMIDHUBER, 1997). Isso acontece, pois grande parte dos modelos multivariados de *machine learning* recebem os dados de entrada associados a um

único registro (por exemplo, uma linha de uma tabela). Essas informações são tratadas pelo modelo que procura mapear relações entre as variáveis independentes e a variável dependente. Já no caso de séries temporais, os dados estão dispostos em ordem cronológica, e é a sequência dos dados que possui as relações a serem mapeadas. No caso das redes neurais recorrentes, simplificadamente, os registros são "acumulados dentro da rede" em sequência e processados em lotes. Dessa maneira, os dados próximos se comunicam, e relações associadas à posição relativa entre os dados podem ser identificadas.

Baseado no exposto acima, boa parte dos modelos de *machine learning* que possuem uma excelente capacidade de extrair relações de dependência entre variáveis independentes e a variável dependente, aparentemente possuem uma performance abaixo do razoável quando aplicados a séries temporais. Por outro lado, tem-se o fato dos modelos tradicionais aplicados a séries temporais, como por exemplo Holt, Holt-Winters e ARIMA, não serem multivariados. E, mesmo quando os dados possuem variáveis independentes associadas a série, normalmente são poucos dados, o que inviabiliza a utilização de modelos de ML.

Este trabalho visa, portanto, apresentar um guia para realização de previsão de séries temporais usando modelos genéricos de ML multivariados, que permitirá a esses modelos identificar relações temporais na variável de interesse, além da inclusão de novas variáveis independentes, o que não é possível nos modelos clássicos. Com isso, uma gama de modelos de ML que incorporam o estado da arte em previsão estarão disponíveis para o contexto de previsões em séries temporais.

Vale ainda frisar que as considerações supracitadas também contribuem para que os modelos tradicionais sejam ainda largamente utilizados em modelos de previsão de séries temporais no mercado.

1.6 ESTRUTURA DO TRABALHO

Inicialmente, no primeiro capítulo introdutório, apresenta-se uma contextualização dos assuntos abordados, bem como as delimitações e limitações do estudo

Seguindo pelo segundo capítulo, onde serão aprofundados os tópicos do estudo, baseando-se em revisões bibliográficas e referenciais teóricos.

O capítulo terceiro é trata dos procedimentos metodológicos, que são as técnicas e processos para desenvolver o trabalho, e onde se definem os passos a serem seguidos para a realização do desenvolvimento do mesmo.

O quarto capítulo centra no desenvolvimento do estudo em si, seguindo o que foi definido no capítulo terceiro. É um dos capítulos mais importantes e extensos, pois é onde se aplicam todos os conhecimentos absorvidos no referencial teórico e é onde se buscam as respostas e conclusões.

O quinto capítulo é referente aos resultados obtidos no capítulo anterior. É onde se avalia se o que foi proposto no início do trabalho foi realmente atingido, ou o quão perto se chegou dos resultados esperados. Também onde se confirma a teoria apresentada no referencial teórico.

Por fim, no sexto e último capítulo apresentam-se as conclusões do estudo, os resultados obtidos, o que foi verificado de pertinente e quais poderiam ser as melhorias e sugestões para futuro.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 A IMPORTÂNCIA DA PREVISÃO EM SÉRIES HISTÓRICAS

Hoje em dia, com a globalização da economia e, por conseguinte, um acréscimo de concorrência entre as organizações, as vantagens competitivas tornaram-se um diferencial de sobrevivência para as empresas. Contudo, inovar num ambiente de incertezas pode ser desastroso, e é natural que as decisões estratégicas sejam tomadas com base em diversas previsões.

Este desejo de compreender o passado e prever o futuro impulsiona a procura de leis que expliquem o comportamento de dados, fenômenos ou acontecimentos. No entanto, na ausência de regras que definem o comportamento de um sistema, procura-se determinar o seu comportamento futuro a partir de observações do passado (WEIGEND, GERSHENFELD, 1994).

As previsões são um exercício que fornece às organizações informações necessárias a respeito do mercado em que estão inseridas. Trata-se de um instrumento estratégico, pois permite evidenciar, por exemplo, o padrão de consumo dos clientes, estimativa da demanda de sangue em um hospital, previsão do clima, previsão de rotas rodoviárias e aeroespaciais, estimativas para renovação de estoque, reconhecimento de voz e frases, cálculos matemáticos e do mercado financeiro, entre outras aplicações (ARMSTRONG, 2001).

As empresas que se antecipam às possibilidades que o mercado está trazendo, aproveitam melhor o cenário e o "*timing*" em que se encontra o setor. Dito isso, as decisões que são fundamentadas nas previsões, influenciam o desenvolvimento interno de recursos e, com isso, na ascensão da empresa (MOON, MENTZER, SMITH, GARVER, 1998).

A escolha do modelo de previsão a ser implementado é um dos grandes desafios relacionados às previsões. Empresas, na maioria dos casos, têm urgências e atributos diferentes em relação aos seus produtos e serviços, fazendo da escolha do modelo mais adequado, uma das primeiras e mais importantes decisões a serem tomadas. Essa divergência dos cenários precisa de diferentes métodos de previsão e, sabendo disso, cada tipo de empresa necessitará analisar suas necessidades para, então, escolher o modelo que melhor se adequa à sua realidade (CECATTO, BELFIORE, 2015).

Além disso, a previsão pode ser feita com diversos tipos de dados. No contexto deste trabalho o foco será em séries temporais, ou seja, um registro de cronológico de dados, como demanda e vendas, número de visitantes de um site, preço das ações, etc.

2.2 SÉRIES TEMPORAIS

Uma série temporal é, portanto, um conjunto sequencial de pontos de dados, medidos normalmente em tempos sucessivos. Isto é matematicamente definido como um conjunto de vetores x(t), t = 0,1,2, ... onde t representa o tempo decorrido. A variável x(t) é tratada como uma variável aleatória. As medições tomadas durante um evento em uma série de tempo são organizadas em uma ordem cronológica adequada (MONTGOMERY, JOHNSON, GARDINER, 1990).

Uma série temporal contendo registros de uma única variável é denominada uni variada. Mas se os registros de mais de uma variável são considerados, é denominado como série multivariada. Ademais, uma série temporal pode ser contínua ou discreta. Em uma série de tempo contínua, as observações são medidas em cada instância de tempo, enquanto uma série de tempo discreta contém observações medidas em pontos discretos de tempo. Por exemplo, leituras de temperatura, fluxo de um rio, concentração de um processo químico etc., podem ser gravadas como uma série de tempo contínua. Por outro lado, a população de uma determinada cidade, produção de uma empresa, taxas de câmbio entre duas moedas diferentes podem representar séries temporais discretas. Normalmente, em uma série temporal discreta, as observações consecutivas são registradas em intervalos de tempo igualmente espaçados, como horário, diário, semanal, mensal ou anual separações. Além disso, séries temporais contínuas podem ser facilmente transformadas em discretas mesclando dados juntos ao longo de um intervalo de tempo especificado (HAMZACEBI, 2008).

Os objetivos de analisar uma série temporal são observar a tendência e sazonalidade da série, identificar padrões não aleatórios de uma variável de interesse e a observação deste comportamento passado pode permitir fazer previsões sobre o futuro, orientando a tomada de decisões. (SOUZA, CAMARGO, 2004).

Os conjuntos de dados registrados com base no tempo não contêm só uma variável como, por exemplo, a venda e a data em que essa venda foi realizada. Existem outras variáveis associadas que são registradas e influenciam na previsão, por isso essas outras variáveis deverão ser consideradas para se ter uma previsão mais fiel à realidade (HAMZACEBI, 2008).

Uma das formas de conseguir analisar mais de uma variável é o *machine learning*, que é uma ferramenta muito eficaz para análise de modelos multivariados. Modelos multivariados e *machine learning* serão mais bem detalhados nas seções posteriores.

2.2.1 Componentes da série temporal

Uma série temporal em geral poderá ser afetada por três componentes principais. Essas componentes são: tendência, sazonalidade e componentes aleatórias.

A tendência de uma série temporal indica um movimento aumento ou diminuição de valor no longo prazo. Por exemplo, séries relacionadas ao crescimento populacional e número de casas em uma cidade mostram tendência de alta, enquanto a tendência de queda pode ser observada em série relacionadas a taxas de mortalidade, epidemias, etc. (SHUMWAY, STOFFER, 2006).

Remover a tendência de uma série auxilia na identificação de outros elementos e permite observar uma faixa de valores base. Há diferentes formas de se extrair a tendência de uma série temporal, como por exemplo usando o método de médias móveis, o modelo de regressão, ou o ajuste exponencial (HAMZACEBI, 2008).

As variações sazonais em uma série temporal são flutuações periódicas. Um exemplo de fatores importantes que podem causar variações sazonais são o clima e condições meteorológicas, costumes, hábitos tradicionais. Outros exemplos clássicos quando se fala de sazonalidade são as vendas de sorvete, que aumentam no verão e as vendas de lã e panos que aumentam no inverno. A variação sazonal é um fator importante a ser considerado por empresários, lojistas e produtores de alimentos.

Variações irregulares ou aleatórias em uma série temporal são causadas por influências imprevisíveis, que não se repetem dentro de um determinado padrão. Essas variações são causadas por incidências esporádicos como: greve, terremoto, inundação, revolução etc. Não há estatísticas confiáveis para se medir flutuações aleatórias em uma série de temporal (SOUZA, CAMARGO, 2004).

2.3 MODELOS TRADICIONAIS PARA PREVISÃO DE SÉRIES TEMPORAIS

Neste tópico serão abordados os principais modelos tradicionais utilizados para realizar a previsão de séries temporais.

2.3.1 Modelo de Holt

A técnica de suavização de Holt, também conhecida como suavização exponencial linear, é um modelo de suavização amplamente conhecido para previsão de dados com tendência. A suavização exponencial dupla é uma ferramenta de previsão para dados de séries temporais que exibem uma tendência linear (RAGSDALE, 2012).

Após observar o valor da série temporal no período t (Y_t), o modelo de Holt calcula uma estimativa do nível base, ou esperado, da série temporal (E_t) e a taxa esperada de aumento ou diminuição (tendência) por período (T_t). A função de previsão no modelo de Holt é representada por:

$$\hat{Y}_{t+n} = E_t + nT_t \quad (1)$$

$$E_t = \alpha Y_t + (1 - \alpha)(E_{t-1} + T_{t-1}) \quad (2)$$

$$T_t = \beta(E_t - E_{t-1}) + (1 - \beta)T_{t-1} \quad (3)$$

onde,

Y_t = valor observado em uma série temporal;

$\hat{Y}_{t+n}$ = valor estimado para o período n, a partir de um valor observado;

E_t = valor do nível base excluído da tendência no período t;

T_t = coeficiente tendência para o período t;

α = parâmetros de suavização exponencial doe nível base;

β = parâmetros de suavização exponencial da tendência.

A equação (1) é usada para obter previsões no próximos n períodos de tempo no futuro.

Os parâmetros de suavização e nas equações (2) e (3) pode assumir qualquer valor entre 0 e 1. Se houver uma tendência de alta nos dados, E_t tende a ser maior que E_{t-1}. Isso tende a aumentar o valor do fator de ajuste de tendência T_t. Alternativamente, se houver uma queda tendência nos dados, Et tende a ser menor que E_{t-1}, isso tende a diminuir o valor do fator de ajuste de tendência T_t (RAGSDALE, 2012).

Embora o método de Holt possa parecer complicado, é um processo simples de três etapas:

1. Calcule o nível base E_t para o período de tempo t usando a equação (2).
2. Calcule o valor de tendência T_t para o período de tempo t usando a equação (3).
3. Calcule a previsão final $\hat{Y}_{t+n}$ para o período de tempo t n usando a equação (1).

Em resumo, primeiro é calculado o nível base da série para que possa ser calculado o valor da tendência, e por fim com o valor dessas variáveis, calcula-se o valor da previsão.

2.3.2 Modelo de Holt-Winters

A suavização exponencial de Holt-Winters, assim chamada em homenagem a seus dois colaboradores: Charles Holt e Peter Winters, é uma das técnicas de análise de séries temporais mais antigas que leva em consideração a tendência e a sazonalidade ao fazer a previsão (HYNDMAN, MAKRIDAKIS, WHEELWRIGHT, 1998). Assim, é uma maneira de modelar os três aspectos da série temporal: um valor base, uma tendência e um padrão de repetição cíclico (sazonalidade). Os três aspectos são representados pelos três parâmetros de suavização exponencial e, portanto, o modelo de Holt-Winters também é conhecido como suavização exponencial tripla (MORETTIN, TOLOI, 2006).

A função da previsão é então dada por:

$$\hat{Y}_{t+n} = E_t + nT_t + S_{t+n-p} \quad (4)$$

$$E_t = \alpha(Y_t - S_{t-p}) + (1 - \alpha)(E_{t-1} + T_{t-1}) \quad (5)$$

$$T_t = \beta(E_t - E_{t-1}) + (1 - \beta)T_{t-1} \quad (6)$$

$$S_t = \gamma(Y_t - E_t) + (1 - \gamma)S_{t-p} \quad (7)$$

onde,

Y_t = valor observado em uma série temporal;
$\hat{Y}_{t+n}$ = valor estimado para o período n, a partir de um valor observado;
E_t = valor do nível base excluído da tendência;
T_t = coeficiente tendência para o período t;
S_t = coeficiente de sazonalidade (fator sazonal);
p = tamanho do período sazonal;

α = parâmetros de suavização exponencial de nível;
β = parâmetros de suavização exponencial de tendência;
γ = parâmetros de suavização exponencial de sazonalidade.

Seguindo a mesma lógica apresentada anteriormente, a equação (4) será usada para realizar a previsão dos n períodos a frente. Da mesma forma que os parâmetros de suavização são calculados utilizando as equações (5), (6) e (7). Esses parâmetros podem assumir valore entre 0 e 1 (RAGSDALE, 2012).

O modelo de Holt-Winter é basicamente um processo de quatro etapas:

1. Calcule o nível base E_t para o período de tempo t usando a equação (5).
2. Calcule o valor de tendência T_t para o período de tempo t usando a equação (6).
3. Calcule o fator sazonal S_t estimado para o período de tempo t usando a equação (7).
4. Calcule a previsão final $\hat{Y}_{t+n}$ para o período de tempo t n usando a equação (4).

A parte mais importante no processo de modelagem usando o modelo Holt ou Holt-Winters é a escolha de bons parâmetros. Para isso, normalmente usa-se o solver para buscar o valor dos parâmetros que minimizem um dado erro de previsão (RAGSDALE, 2012).

Outro ponto que merece destaque no modelo Holt-Winters é que o período da sazonalidade deve ser definido pelo analista previamente. Geralmente o período é obtido inspecionando manualmente os dados (HYNDMAN, MAKRIDAKIS, WHEELWRIGHT, 1998).

2.3.3 ARIMA

Os modelos ARIMA são modelos estatísticos lineares para análise de séries temporais (BAKAR, ROSBI, 2017). A abreviação em língua inglesa refere-se a “Auto-Regressive Integrated Moving Average model”, ou seja, um modelo autorregressivo integrado de médias móveis. Esses modelos são amplamente utilizados, mas funcionam como uma caixa-preta, dificultando a análise das características da série que levaram a uma dada previsão.

Em um modelo autorregressivo as previsões correspondem a uma combinação linear de valores passados da variável. Em um modelo de média móvel, as previsões correspondem a uma combinação linear de erros de previsão anteriores. Basicamente, os modelos ARIMA

combinam essas duas abordagens. Como eles exigem que as séries temporais sejam estacionárias, diferenciar (integrar) as séries temporais pode ser um passo necessário, ou seja, considerar a série temporal das diferenças ao invés da original (SOUZA, CAMARGO, 2004).

O modelo tem como premissa básica que a série temporal é gerada por um processo estocástico cuja natureza pode ser representada através de um modelo. A notação empregada para designação do modelo é normalmente ARIMA (p,d,q) onde p representa o número de parâmetros auto-regressivos, d o número de diferenciações para que a série torne-se estacionaria e q o número de parâmetros de médias móveis (SATRIO, et al., 2020).

Nesse estudo não será utilizado o modelo ARIMA, pois a análise dos resultados do modelo não é tão clara e intuitiva quanto os modelos de Holt e Holt-Winter. Dessa forma, o modelo foi apresentado, porém não será objeto de comparação nos capítulos seguintes.

2.4 MODELOS DE *MACHINE LEARNING*

Aprendizado de máquina, como também chamado de *machine learning*, é composto por modelos em seu processo de otimização em que "ensina" computadores a decodificar informações mais complexas em grande escala de maneira quase que independente por meio da análise de dados e de algoritmos previamente programados (FACELLI, LORENA, GAMA, de CARVALHO, 2019).

Esses modelos de aprendizado de máquina, quando colocados em contato com novas entradas (*inputs*), conseguem identificar padrões que se repetem e, até mesmo, podem tomar providências utilizando pouca ou nenhuma ação do homem. Permitem alterar o comportamento do modelo, apenas se baseando em dados anteriores, produzindo resultados (*outputs*) ainda mais precisos (LUDERMIR, 2021).

Modelos que utilizam de aprendizado de máquina têm uma alta maleabilidade e mutabilidade, pois cada conjunto de dados apresentam um formato específico e diferentes variáveis, tornando essa flexibilidade um instrumento de grande valia para sua aplicação em diversos tipos de organizações (Ertel, 2017).

Uma vez que um modelo de *machine learning* é treinado, estará apto a realizar funções complexas e dinâmicas, o que o torna um forte aliado na análise com diversos tipos de banco de dados (CARVALHO, 2021).

Para cada um dos problemas de *machine learning*, existem diferentes algoritmos que podem ser utilizados, é recomendado realizar testes para verificação do modelo mais adequado para a situação (CARVALHO, 2021).

Muitas vezes, a melhor alternativa para solucionar um problema não consiste na aplicação de somente uma técnica de ML, mas na combinação de mais de uma técnica na mesma análise (HAIR, BLACK, BABIN, ANDERSON, TATHAM, 2009).

Existem quatro tipos de algoritmos de aprendizado de máquina: supervisionados, semi-supervisionados, não supervisionados e de reforço.

Aprendizado supervisionado: No aprendizado supervisionado, a máquina é ensinada por exemplo. O operador fornece ao algoritmo de aprendizado de máquina um conjunto de dados conhecido que inclui entradas e saídas desejadas, e o algoritmo deve encontrar um método para determinar como chegar a essas entradas e saídas (MAHESH, 2018). Enquanto o operador sabe as respostas corretas para o problema, o algoritmo identifica padrões nos dados, aprende com as observações e faz previsões. O algoritmo faz previsões e é corrigido pelo operador e esse processo continua até que o algoritmo atinja um alto nível de precisão/desempenho. Sob o guarda-chuva da aprendizagem supervisionada caem: Classificação, Regressão e Previsão (MAHESH, 2018).

- Classificação: Em tarefas de classificação, o programa de aprendizado de máquina deve tirar uma conclusão dos valores observados e determinar a qual categoria as novas observações pertencem. Por exemplo, ao filtrar e-mails como 'spam' ou 'não spam', o programa deve analisar os dados observacionais existentes e filtrar os e-mails de acordo.
- Regressão: Em tarefas de regressão, o programa de aprendizado de máquina deve estimar – e entender – as relações entre as variáveis. A análise de regressão concentra-se em uma variável dependente e uma série de outras variáveis – tornando-a particularmente útil para previsão e previsão.
- Previsão: A previsão é o processo de fazer previsões sobre o futuro com base nos dados passados e presentes e é comumente usada para analisar tendências.

Aprendizagem semi-supervisionada: O aprendizado semi-supervisionado é semelhante ao aprendizado supervisionado, mas usa dados rotulados e não rotulados. Dados rotulados são essencialmente informações que possuem *tags* significativas para que o algoritmo possa entender os dados, enquanto os dados não rotulados não possuem essa informação. Ao usar essa combinação, os algoritmos de aprendizado de máquina podem aprender a rotular dados não rotulados (MAHESH, 2018).

Aprendizado não supervisionado: Aqui, o algoritmo de aprendizado de máquina estuda dados para identificar padrões. Não há chave de resposta ou operador humano para fornecer instruções. Em vez disso, a máquina determina as correlações e relacionamentos analisando os dados disponíveis. Em um processo de aprendizado não supervisionado, o algoritmo de aprendizado de máquina é deixado para interpretar grandes conjuntos de dados e endereçar esses dados de acordo (MAHESH, 2018). O algoritmo tenta organizar esses dados de alguma forma para descrever sua estrutura. Isso pode significar agrupar os dados em clusters ou organizá-los de uma maneira que pareça mais organizada. À medida que avalia mais dados, sua capacidade de tomar decisões sobre esses dados melhora gradualmente e se torna mais refinada.

Sob o tópico do aprendizado não supervisionado, estão:

- Clustering: Clustering envolve agrupar conjuntos de dados semelhantes (com base em critérios definidos). É útil para segmentar dados em vários grupos e realizar análises em cada conjunto de dados para encontrar padrões.
- Redução de dimensão: A redução de dimensão reduz o número de variáveis consideradas para encontrar as informações exatas necessárias.

Aprendizado por reforço: O aprendizado por reforço se concentra em processos de aprendizado regimentados, em que um algoritmo de aprendizado de máquina é fornecido com um conjunto de ações, parâmetros e valores finais (MAHESH, 2018). Ao definir as regras, o algoritmo de aprendizado de máquina tenta explorar diferentes opções e possibilidades, monitorando e avaliando cada resultado para determinar qual é o ideal. O aprendizado por reforço ensina a máquina por tentativa e erro. Ele aprende com as experiências passadas e começa a adaptar sua abordagem em resposta à situação para alcançar o melhor resultado possível (MAHESH, 2018).

A escolha do algoritmo de aprendizado de máquina certo depende de vários fatores, incluindo, mas não limitado a: tamanho, qualidade e diversidade dos dados, bem como quais respostas as empresas desejam obter desses dados. Considerações adicionais incluem precisão, tempo de treinamento, parâmetros, pontos de dados e muito mais. Portanto, escolher o algoritmo certo é uma combinação de necessidade de negócios, especificação, experimentação e tempo disponível (MAHESH, 2018).

Dito isso, o foco do presente trabalho será nos algoritmos de aprendizado de máquina supervisionado. A seguir serão apresentados os modelos de *machine learning* escolhidos para serem usados neste trabalho.

2.4.1 Regressão linear

A regressão linear é um algoritmo utilizado apenas para problemas de regressão e é, talvez um dos algoritmos mais conhecidos e bem compreendidos em estatística e aprendizado de máquina (DOBRA, 2002). Ele assume uma relação linear entre as variáveis de entrada (variáveis independentes) e a variável de saída única (variável dependente).

Quando há uma única variável de entrada, o método é chamado de regressão linear simples. Quando há múltiplas variáveis de entrada, a literatura de estatística geralmente se refere ao método como regressão linear múltipla.

A regressão linear simples pode ser verificada pela equação (8) abaixo.

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (8)$$

onde,

Y_i = variável dependente;
α = coeficiente de interceptação da reta com o eixo vertical;
β = coeficiente angular;
X_i = variável independente;
ε_i = Representa todos os fatores residuais mais os possíveis erros de medição.

A regressão linear simples usa a forma tradicional de interceptação de inclinação, onde α e β são as variáveis que o algoritmo tentará "aprender" a produzir as previsões mais precisas.

Já a regressão linear múltipla pode ser observada na equação (9) a seguir.

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \varepsilon_i$$

onde,

Y_i = variável dependente;
α = coeficiente de interceptação da reta com o eixo vertical;
βi = coeficiente angular da variável i;

Xi = variável independente;

ε_i = Representa todos os fatores residuais mais os possíveis erros de medição.

A regressão linear múltipla segue o mesmo princípio da regressão linear simples, somente adicionando mais variáveis independentes ao modelo.

Diferentes técnicas podem ser usadas para preparar ou treinar a equação de regressão linear a partir de dados, a mais comum é chamada de mínimos quadrados ordinários. É comum, portanto, referir-se a um modelo elaborado desta forma como regressão linear de mínimos quadrados ordinários ou apenas regressão de mínimos quadrados (SEBER, LEE, 2012).

Em *machine learning* a regressão linear é usada para estimar os valores dos coeficientes que melhor representam dados que se tem disponíveis.

2.4.2 XGBoost

XGBoost, que significa *eXtreme Gradient Boosting*, é um modelo que utiliza de um conjunto de algoritmos baseados em modelos de árvores de decisão, onde novas árvores corrigem erros daquelas árvores que já fazem parte do modelo. As árvores são adicionadas até que não seja possível fazer mais melhorias no modelo (CHEN; GUESTRIN, 2016).

O fator mais importante por trás do sucesso do XGBoost é sua escalabilidade e flexibilidade em todos os cenários. O sistema é executado mais de dez vezes mais rápido do que as soluções populares existentes em uma única máquina e é dimensionado para bilhões de exemplos em configurações distribuídas ou com limitação de memória (DAWOOD, 2019). O modelo realiza a combinação de diversos modelos de previsão fracos para elaboração de um modelo principal mais forte.

O XGBoost é projetado para classificação e regressão em conjuntos de dados tabulares, mas podendo ser adaptado e usado para realizar a previsão de séries temporais. Para isso, exige-se que o conjunto de dados de seja transformado primeiro em um problema de aprendizado supervisionado (CHEN; GUESTRIN, 2016).

O ponto abordado acima será o foco principal do trabalho, explicar e mostrar passo a passo como realizar o tratamento e preparação do conjunto de dados para que o XGboost reconheça a série temporal. Lembrando que esses passos poderão ser replicados para outros modelos de *machine learning*.

O XGBoost tem parâmetros internos para serem definidos. Os parâmetros mais comumente usados para configuração do modelo são:

- Objective: Parâmetro utilizado para definir qual modelo que o XGBoost irá utilizar.
- Booster: Parâmetro que define o tipo de booster (gbtree, gblinear ou dart) a ser usado. Para problemas de classificação, pode-se usar gbtree, dart. Para regressão, pode ser usado as três possibilidades.
- Eta: Esse parâmetro controla a taxa de aprendizado, ou seja, a taxa na qual o modelo aprende padrões nos dados. Após cada rodada, o modelo reduz os pesos dos recursos para alcançar o melhor resultado. Menor eta leva a computação mais lenta, pois eleva o grau de exigência nos resultados.
- Max_depth: Controla a profundidade da árvore. Quanto maior a profundidade, mais complexo o modelo; maiores chances de *overfitting*. Não há valor padrão para max_depth. Conjuntos de dados maiores exigem árvores profundas para aprender as regras dos dados. Dessa forma, foi escolhido um valor elevado, na tentativa de obter melhores resultados.
- Subsample: Parâmentro que controla o número de amostras (observações) fornecidas a uma árvore. Normalmente, seus valores ficam entre (0,5-0,8).
- Colsample_bytree: Controla o número de recursos (variáveis) fornecidos a uma árvore. Normalmente, seus valores ficam entre (0,5,0,9).
- Gamma: Controla a regularização (ou previne o *overfitting*). O valor ideal de gama depende do conjunto de dados e de outros valores de parâmetros. Quanto maior o valor, maior a regularização. Regularização significa penalizar grandes coeficientes que não melhoram o desempenho do modelo. default = 0 significa não regularização.

Usualmente são sugeridos alguns parâmetros base para cada modelo. A partir desses valores de parâmetros base, realiza-se o treino e verifica qual o resultado do modelo com os parâmetros escolhidos, se os valores não forem satisfatórios, o recomendado é que altere os valores dos parâmetros e refaça o treino até obter melhores resultados (CHEN; GUESTRIN, 2016).

3 PROCEDIMENTOS METODOLÓGICOS

O procedimento metodológico do presente trabalho é feito prioritariamente por pesquisa de natureza aplicada, com abordagem quantitativa, com propósito de caráter explicativo e de método experimental, que está inserido na área de pesquisa operacional na engenharia de produção, na subárea de inteligência computacional (ABEPRO, [s.d.]).

3.1 DADOS DA PESQUISA

O conjunto de dados (*dataset*) escolhido é conhecido na literatura por 'Seatbelts'. Trata-se de um conjunto de dados que foi encomendado pelo departamento de transportes em 1984 para medir as diferenças nas mortes antes e depois que a legislação do cinto de segurança dianteiro entrou em vigor em 31 de janeiro de 1983 na Grã-Bretanha (HARVEY, DURBIN, 1986).

Os dados fornecem informações relacionadas as mortes ocorridas em acidentes de trânsito e foram colhidos de janeiro de 1969 até dezembro de 1984. Durante este período, uma lei tornou obrigatório o uso de cintos de segurança a partir de 31 de janeiro de 1983.

O conjunto de dados original apresenta oito colunas descritas a seguir (identificadas com o nome no dataset):

- DriversKilled - número de motoristas que morreram em acidente de carro;
- drivers front - passageiros do banco dianteiro mortos ou gravemente feridos;
- rear - passageiros do banco traseiro mortos ou gravemente feridos;
- kms - distância percorrida em quilómetros;
- PetrolPrice - preço da gasolina;
- VanKilled - número de condutores de van que morreram em acidentes;
- law - 0/1: indica se a lei estava ou não em vigor naquele mês;
- Date - Coluna que representa a data das observações, contendo os números de 1 até 192, que representam os meses de janeiro de 1969 até dezembro de 1984.

Nesse estudo serão utilizadas as colunas, Date, DriversKilled, kms, PetrolPrice e law.

O tamanho da base de dados também foi um fator importante para a sua escolha. São 192 registros, uma quantidade que não é muito pequena para ser submetida aos modelos de *machine learning*, mas também não é muito grande para ser analisada pelos modelos tradicionais, Holt e Holt Winters, em planilhas como Excel.

3.2 ETAPAS DA PESQUISA

As etapas deste presente trabalho visam guiar o leitor passo a passo em como trabalhar com séries temporais e prepará-las para serem inseridas nos modelos de *machine learning*. Mas antes, para efeito de comparação, os dados foram analisados com modelos tradicionais de séries temporais. Desta forma, é possível perceber melhor a diferença entre as duas abordagens.

A etapas da pesquisa podem ser verificadas a seguir.

- Análise exploratória dos dados;
- Aplicação dos modelos tradicionais Holt e Holt-Winters;
- Aplicação do modelo de regressão múltipla acrescentando aos dados novas *features* baseadas na data;
- Aplicação do modelo XGBoost usando os mesmos dados submetidos ao modelo de regressão;
- Aplicação do modelo XGBoost acrescentando aos dados novas *features*, chamadas de "*features* adicionais";
- Análise de resultados e conclusão.

A primeira etapa da pesquisa consistiu em uma análise exploratória inicial dos dados, onde foram visualizados os dados do conjunto escolhido, observando alguns fatores importantes para a previsão, tais como: a presença de valores faltantes no conjunto de dados, verificação da frequência em que os dados foram registrados, necessidade de tratamento dos dados.

Em seguida, foi iniciada a segunda etapa, onde os dados foram submetidos aos modelos tradicionais, primeiramente no modelo de Holt, seguido pelo modelo de Holt-Winters.

Feito a aplicação dos modelos tradicionais, fez-se a decomposição da coluna temporal, ou seja, baseado na coluna "Date" foram criadas duas novas colunas, uma coluna com o ano, outra coluna com o mês, a coluna do dia não foi criada pois os dados foram registrados mensalmente.

Com as *features* baseadas na decomposição da data, foi aplicado esse conjunto de dados aos dois modelos de *machine learning* escolhidos. Primeiramente foi aplicado o modelo de regressão múltipla e em seguida ao modelo XGBoost.

Após essa primeira previsão realizada com os modelos de acima, foi criado um novo conjunto de novas *features* (78), chamadas de "*features* adicionais". Em seguida os dados foram

submetidos novamente ao modelo XGBoost, com os mesmos parâmetros usados anteriormente, de forma a avaliar o impacto das novas features na performance do modelo. A última etapa foi a análise comparativa dos resultados obtidos com todos os modelos, nessa etapa verificou-se os resultados e foram feitas as considerações finais.

4 DESENVOLVIMENTO

O presente capítulo representa o desenvolvimento e aplicação do roteiro definido no Capítulo 3. Inicia com os modelos tradicionais e em seguida apresenta as etapas de preparação dos dados e realização da previsão com séries temporais usando modelos de *Machine Learning*. Dessa forma, mostra as operações necessárias para submeter dados de séries temporais aos modelos de Machine Learning genéricos, salientando o impacto da criação de novas *features* na performance dos modelos de ML.

Para todos os modelos abaixo, os dados foram divididos em um conjunto de treinamento e outro conjunto de teste, para manter um padrão que possibilite a comparação dos resultados.

O conjunto de treinamento cobre os dados do início das observações até dezembro de 1983 e o conjunto de teste os dados de janeiro até dezembro de 1984 (12 últimos períodos).

Um ponto importante para ser observado em relação a separação dos dados é que ela não é feita de forma aleatória, pois o conjunto de dados é uma série temporal, ou seja, deve-se respeitar a ordem cronológica das observações.

Para análise dos resultados dos modelos em relação a eficácia, foi utilizada a raiz quadrada do erro quadrático médio (RMSE). O erro quadrático médio é calculado a partir da diferença entre o valor previsto e o valor real, o resultado dessa diferença é elevado ao quadrado e por fim se faz a média de todos esses valores para obter o valor do MSE. Já o RMSE, é obtido com a raiz quadrada do erro quadrático médio.

4.1 ANÁLISE EXPLORATÓRIA DOS DADOS

Esta etapa inicial é obrigatória para qualquer tipo de análise de dados, sejam ou não dados provenientes de séries temporais. Desta forma, o analista consegue um entendimento básico de seus dados e das relações existentes entre as variáveis analisadas, possibilitando observar as características da série para se ter uma ideia dos modelos elegíveis.

4.1.1 Leitura e visualização inicial dos dados

Primeiramente, para começar a manipular, tratar e posteriormente submeter os dados aos modelos, é feita a leitura e visualização do *dataset*. Essa etapa inicial pode ser feita através de várias ferramentas como Excel, GoogleSheets, Tableau, Microsoft Power Bi, linguagens estatísticas como R/RStudio. No caso do presente trabalho, foram utilizados o Excel e o programa computacional "RStudio", que utiliza a linguagem de programação R.

O passo inicial foi analisar a dimensão do *dataset*, quais são as colunas, sua quantidade e o número de observações (linhas). Seguindo essa primeira análise, procura-se visualizar os dados por meio de gráficos como forma de ter uma melhor visão geral de como os mesmos se comportam ao longo do tempo.

A Tabela 1 apresenta um extrato dos dados. A tabela completa está no Apêndice A.

Tabela 1- Apresentação do conjunto de dados.

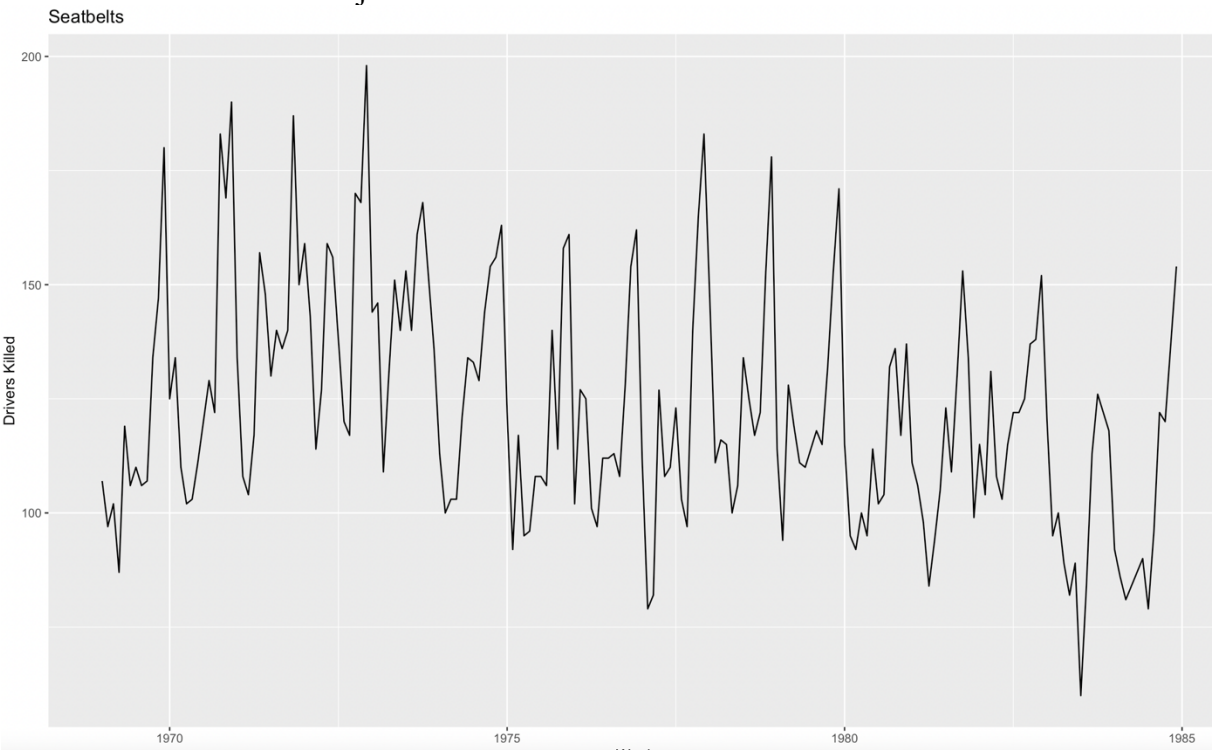
Date	DriversKilled	kms	PetrolPrice	law
1969-01-01	107	9059	0,103	0
1969-02-01	97	7685	0,102	0
1969-03-01	102	9963	0,102	0
1969-04-01	87	10955	0,101	0
1969-05-01	119	11823	0,101	0
1969-06-01	106	12391	0,101	0
1969-07-01	110	13460	0,104	0
1969-08-01	106	14055	0,104	0
1969-09-01	107	12106	0,104	0
1969-10-01	134	11372	0,103	0
1969-11-01	147	9834	0,103	0
1969-12-01	180	9267	0,102	0
1970-01-01	125	9130	0,101	0
1970-02-01	134	8933	0,101	0
1970-03-01	110	11000	0,100	0
1970-04-01	102	10733	0,099	0
1970-05-01	103	12912	0,098	0
1970-06-01	111	12926	0,098	0
1970-07-01	120	13990	0,097	0
1970-08-01	129	14926	0,097	0
1970-09-01	122	12900	0,097	0
1970-10-01	183	12034	0,096	0
1970-11-01	169	10643	0,096	0
1970-12-01	190	10742	0,095	0
1971-01-01	134	10266	0,097	0
1971-02-01	108	10281	0,096	0
1971-03-01	104	11527	0,095	0
1971-04-01	117	12281	0,095	0
1971-05-01	157	13587	0,094	0
1971-06-01	148	13049	0,094	0
1971-07-01	130	16055	0,093	0
1971-08-01	140	15220	0,093	0
1971-09-01	136	13824	0,093	0
1971-10-01	140	12729	0,092	0
1971-11-01	187	11467	0,092	0

Fonte: Autor (2021).

Como pode-se perceber, o conjunto de dados contém as quatro colunas que foram escolhidas que foram descritas no Capítulo 3 e o conjunto de dados é composto por 192 registros.

Na Figura 1 abaixo, foi feita a visualização desses dados para tornar mais fácil a observação das características de série temporal. O gráfico é uma das principais ferramentas que auxiliam nessa observação.

Figura 1 – Número de mortes registradas causadas por acidente de carro na Grã-Bretanha de janeiro de 1969 até dezembro de 1984.



Fonte: Autor (2021).

Foram analisadas a presença das principais características de séries temporais, tais como a sazonalidade, tendência, ciclos e padrões de repetição.

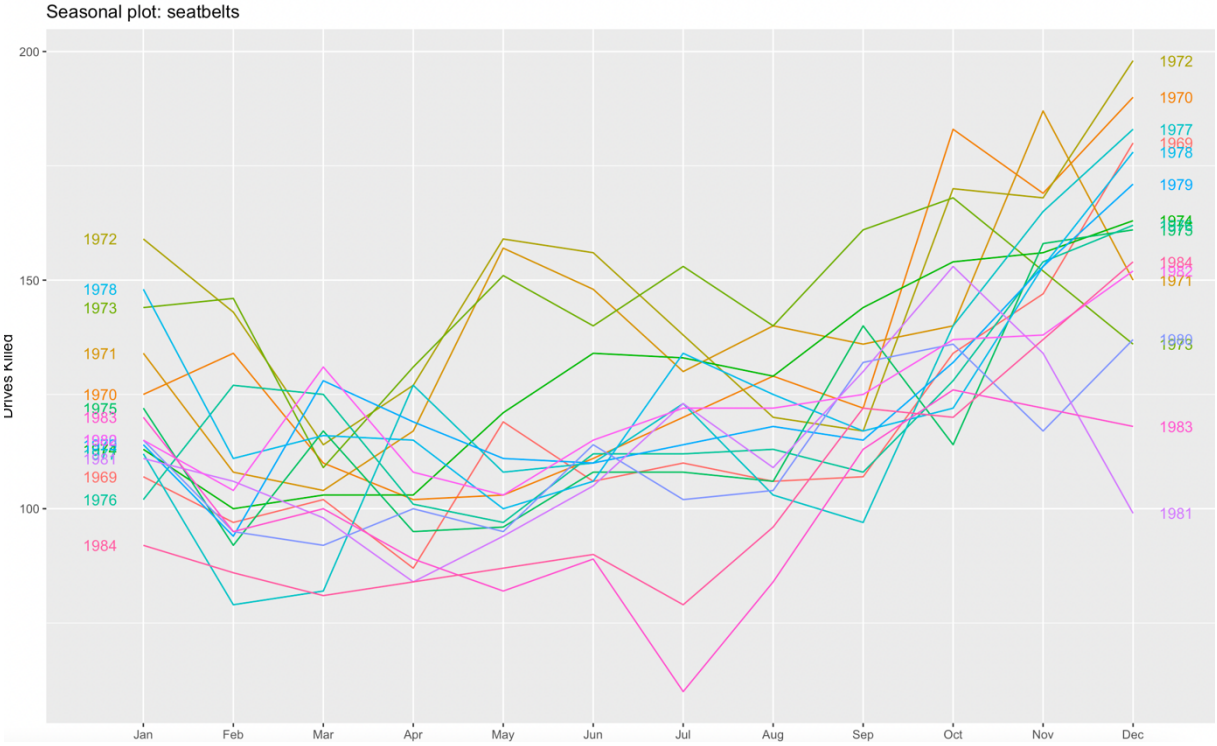
Notou-se a presença de uma leve tendência de queda, já que os dados parecem estar diminuindo de valor de uma forma geral com o passar do tempo.

Juntamente com a tendência, pode-se perceber a sazonalidade. Uma série temporal sazonal (periódica) é caracterizada por fenômenos que se repetem por períodos idênticos. Por exemplo, fenômenos que ocorrem diariamente em uma certa hora, todos os dias, ou em um certo mês em todos os anos.

Essa é uma tentativa de projetar o que os modelos teriam de capturar em termos de padrões apresentados.

Para melhorar as suposições acima, decidiu-se visualizar a série separada pelos diferentes anos (legenda), pelos meses do ano (eixo horizontal) que está representado por 12 meses, correspondente a um ano, a fim de ter uma visualização dos diferentes anos separados por meses, como mostra o Figura 2 abaixo.

Figura 2 - Sazonalidade por ano.



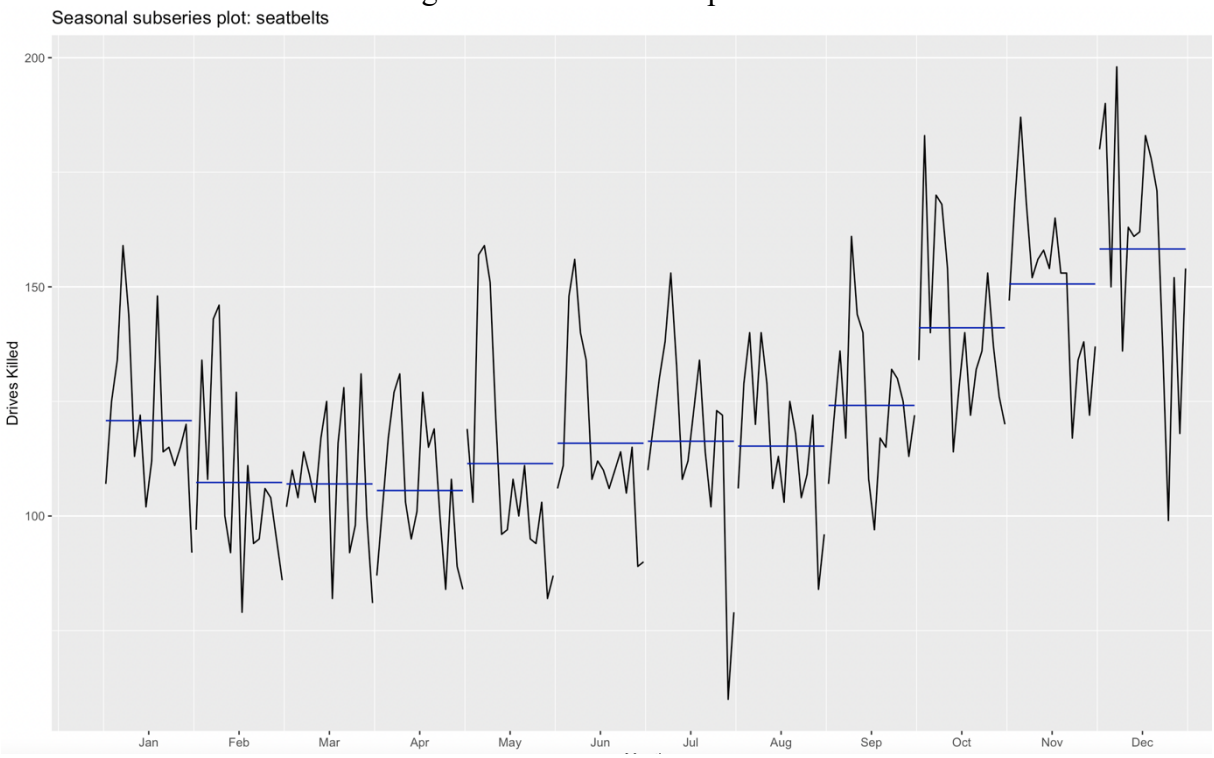
Fonte: Autor (2021).

Percebe-se que independente do ano analisado, o número de acidentes aumenta significativamente nos meses de outubro, novembro e dezembro e os meses que apresentam uma média dos menores valores são fevereiro e março. Esses fatores observados são muito relevantes para a característica de sazonalidade.

Outra análise interessante, é a observação de uma leve tendência de queda do número de mortes por acidentes ao passar dos anos, principalmente após entrar em vigor a lei que exige o uso de cintos de segurança (1983). Os anos de 1983 e 1984 tem os menores valores em de maio até agosto, ou seja, um terço do ano.

Feito isso, o próximo passo foi visualizar a média por mês de todos os anos. A sazonalidade pode ser melhor observada separando-se os meses, dividindo em subséries e também com a representação do ponto médio de cada mês, como mostrado na Figura 3.

Figura 3 - Sazonalidade por mês.



Fonte: Autor (2021).

Notou-se que o ponto médio variou na faixa entre 100 até aproximadamente 160, quando analisado todo período. Com essa visualização da distribuição ao longo do tempo fica nítida a presença da sazonalidade. Percebe-se que o número de mortes começa aumentar nos meses de setembro até dezembro e segue quase que estável na faixa de 100 até 125 nos outros meses.

4.1.2 Limpeza e tratamento dos dados

Para realizar a limpeza e tratamento dos dados, a primeira verificação foi a existência de dados faltantes no *dataset*. É muito comum a presença de dados faltantes em um *dataset*, e esses dados podem aparecer por diversas razões. Se houver muitos dados faltantes no conjunto

de dados escolhido e esses dados não forem tratados, a análise e a posterior previsão ficam prejudicadas.

Feito essa verificação, constatou-se que no conjunto de dados escolhido, não haviam dados faltantes, dessa forma, não foi necessário realizar tratamento para a ausência de dados no dataset.

O tratamento de dados faltantes, é uma abordagem de grande relevância para ajuste do *dataset*. Esses valores faltantes podem interferir de forma negativa ao modelo. O importante é que seja analisado para cada caso, qual abordagem que mais se encaixa para realização do tratamento desses dados.

Para outros conjuntos de dados que apresentarem dados faltantes, o leitor poderá adotar algumas formas de tratamento, tais como: analisar o valor da média dos últimos três períodos e inserir o resultado no local com a ausência do valor; observar o valor da moda na coluna onde tem-se dados faltantes e preencher com esse valor etc.

No tratamento dos dados é importante preencher os dados faltantes de forma que não interfira na veracidade dos valores do dataset, pois se preencher os dados com valores aleatórios ou com valores nulos, trará informações não realísticas para o modelo e consequentemente prejudicará o resultado.

O tratamento de dados também pode ser feito com colunas que apresentam valores muito pequenos e/ou que apresentem muitas casas decimais, tornando a análise mais complexa. Observou-se que a coluna “PetrolPrice” possuíam valores com muitas casas decimais e optou-se por reduzir os valores em até 3 casas depois da vírgula, mas não foi alterado o valor em relação a ordem de grandeza.

Para coluna de dados que apresentarem uma ordem de grandeza muito pequena, é recomendado que faça a multiplicação por um número que não causará influência nos valores, mas que simplifique a análise. Um multiplicador utilizado para esses casos são os valores múltiplos de 10, pois dessa forma aumentará a ordem de grandeza dos valores, sem alterar a análise.

4.2 APLICAÇÃO DOS MODELOS TRADICIONAIS

Como explicado na metodologia, nesta seção serão aplicados os modelos tradicionais de Holt e Holt-Winters (Aditivo).

Para aplicação desses dois modelos tradicionais, como são modelos de análise univariada, foi utilizado somente a coluna “Driverskilled” e a coluna referente a data das observações “Date”, para realização da previsão.

4.2.1 Modelo Holt

Foi aplicado o modelo tradicional de Holt para demonstrar a aplicação do modelo e análise dos resultados obtidos. A explicação do modelo de Holt, juntamente com as fórmulas e demonstração das variáveis envolvidas na aplicação, pode ser observado na seção 2.3.1 deste trabalho.

Os valores dos parâmetros α e β foram encontrados usando o Solver do Excel para minimizar o erro médio quadrático (MSE) da previsão nos valores de treinamento. Foram encontrados os valores mostrados no Quadro 1 a seguir.

Quadro 1 - Parâmetros do modelo Holt.

alpha	**0,828**
beta	**0,000**

Fonte: Autor (2021).

Nota-se que o valor de α encontrado é igual a 0,828 que diz respeito a componente de amortecimento base da série e o valor de β nulo ao amortecimento da tendência. Neste caso, o valor de β nulo indica que a série não apresenta tendência.

Na Tabela 2 abaixo, é mostrado a aplicação do modelo e os valores obtidos. A tabela completa com os valores e aplicação do modelo de Holt pode ser encontrada no Apêndice B.

Tabela 2 - Aplicação modelo Holt.

Ano	Mês	Período	DrivesKilled	Nível Base	Tendência	Previsão
1969	1	1	107	107,00	0,00	--
	2	2	97	98,72	0,00	107,00
	3	3	102	101,44	0,00	98,72
	4	4	87	89,49	0,00	101,44
	5	5	119	113,92	0,00	89,49
	6	6	106	107,36	0,00	113,92
	7	7	110	109,55	0,00	107,36
	8	8	106	106,61	0,00	109,55
	9	9	107	106,93	0,00	106,61
	10	10	134	129,34	0,00	106,93
	11	11	147	143,96	0,00	129,34
	12	12	180	173,79	0,00	143,96
1970	1	13	125	133,40	0,00	173,79
	2	14	134	133,90	0,00	133,40
	3	15	110	114,12	0,00	133,90
	4	16	102	104,09	0,00	114,12
	5	17	103	103,19	0,00	104,09
1983	1	169	120	125,08	0,00	149,49
	2	170	95	100,18	0,00	125,08
	3	171	100	100,03	0,00	100,18
	4	172	89	90,90	0,00	100,03
	5	173	82	83,53	0,00	90,90
	6	174	89	88,06	0,00	83,53
	7	175	60	64,83	0,00	88,06
	8	176	84	80,70	0,00	64,83
	9	177	113	107,44	0,00	80,70
	10	178	126	122,80	0,00	107,44
	11	179	122	122,14	0,00	122,80
	12	180	118	118,71	0,00	122,14
1984	1	181	92			118,71
	2	182	86			118,71
	3	183	81			118,71
	4	184	84			118,71
	5	185	87			118,71
	6	186	90			118,71
	7	187	79			118,71
	8	188	96			118,71
	9	189	122			118,71
	10	190	120			118,71
	11	191	137			118,71
	12	192	154			118,71

Fonte: Autor (2021).

Para o primeiro valor do nível base, foi repetido o valor original da primeira observação da coluna “DriversKilled” e para os demais aplicou-se a fórmula para o cálculo.

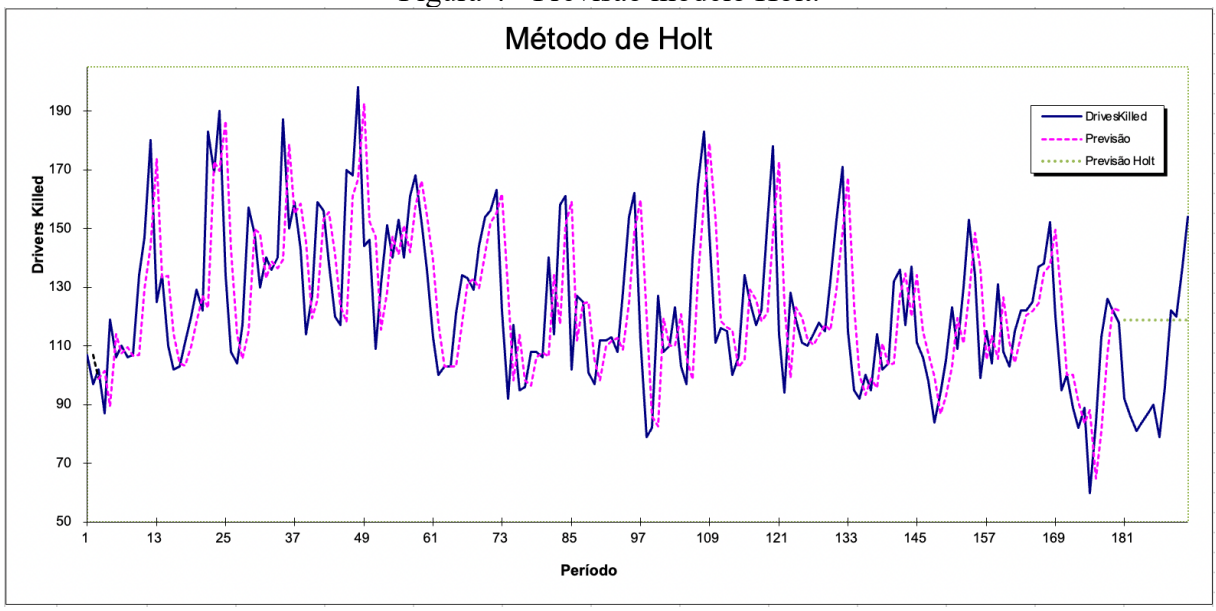
Como dito anteriormente, a coluna da tendência ficou com o valor zero, por causa do β calculado pela ferramenta Solver ter apresentado valor nulo. A coluna do nível base e previsão foram calculadas de acordo com as fórmulas mostradas na seção 2.2.2.1.

Como explicado anteriormente, o conjunto de dados foi separado em 2 conjuntos, sendo que o primeiro deles, reúne os valores desde o primeiro período de 1969 (início das observações), até o último período de 1983. Estes foram utilizados para o cálculo dos parâmetros do modelo. O segundo conjunto de dados consta os valores do último ano (1984). Os últimos 12 períodos foram utilizados para realizar a previsão e análise dos resultados do modelo.

4.2.2 Resultados do modelo Holt

Na Figura 4 a seguir mostra os valores de previsão obtidos com esse modelo.

Figura 4 - Previsão modelo Holt.



Fonte: Autor (2021).

Como pode ser observado nos valores previstos do conjunto de teste (linha tracejada na cor cinza), os 12 períodos finais, a previsão foi um valor único de 118,71.

A eficácia do modelo foi calculada utilizando-se o parâmetro RMSE. O resultado pode ser observado na Quadro 2 na página seguinte, que apresenta o RMSE dos dados usados no "treinamento" do modelo, e em seguida os mesmos cálculos foram feitos para os dados previsto considerando os próximos 3,6 e 12 meses dos dados de teste.

Quadro 2 - Valores RMSE após a aplicação do modelo Holt.

RMSE Conjunto Treino	**22,1**
RMSE Conjunto Teste 3 Meses	**32,7**
RMSE Conjunto Teste 6 Meses	**32,2**
RMSE Conjunto Teste 12 Meses	**28,8**

Fonte: Autor (2021).

Quanto maior o RMSE menos eficiente foi a previsão, ou seja, como visto no Quadro 2 a previsão dos dados do conjunto treino estão mais próximo dos valores reais e os dados do conjunto teste ficaram mais discrepantes.

4.2.3 Modelo Holt-Winters *(Aditivo)*

O modelo de Holt-Winters é um modelo tradicional com capacidade de identificar sazonalidade e tendência nos dados. A explicação deste modelo e forma de aplicação, foi descrito na seção 2.3.2 deste presente trabalho.

Seguindo a mesma lógica do modelo de Holt, os resultados obtidos estão expostos no Quadro 3.

Quadro 3 - Parâmetros do modelo Holt-Winters.

alpha	**0,042**
beta	**0,668**
gamma	**0,662**

Fonte: Autor (2021).

Vale frisar que devido a não linearidade da equação de otimização do MSE, normalmente existem outros conjuntos de valores para os parâmetros acima com o mesmo resultado. Esses valores são chamados também de “ótimos locais”.

Na Tabela 3 abaixo, é exposto a aplicação do modelo de Holt-Winters. A tabela completa com os valores e aplicação do modelo pode ser encontrada no Apêndice C.

Tabela 3 – Aplicação modelo Holt-Winters.

Ano	Mês	Período	DriversKilled	Nível Base	Tendência	Fator Sazonal	Previsão
1969	1	1	107	--	--	-9,833	--
	2	2	97	--	--	-19,833	--
	3	3	102	--	--	-14,833	--
	4	4	87	--	--	-29,833	--
	5	5	119	--	--	2,167	--
	6	6	106	--	--	-10,833	--
	7	7	110	--	--	-6,833	--
	8	8	106	--	--	-10,833	--
	9	9	107	--	--	-9,833	--
	10	10	134	--	--	17,167	--
	11	11	147	--	--	30,167	--
	12	12	180	116,8	0,0	63,167	--
1970	1	13	125	117,58	0,50	1,59	107,00
	2	14	134	119,56	1,49	2,86	98,25
	3	15	110	121,21	1,60	-12,43	106,22
	4	16	102	123,18	1,85	-24,10	92,97
	5	17	103	124,02	1,17	-13,19	127,19
1983	1	169	120	126,02	0,78	-3,79	126,88
	2	170	95	125,91	0,19	-23,99	116,40
	3	171	100	124,79	-0,69	-14,57	131,56
	4	172	89	123,20	-1,29	-27,13	110,83
	5	173	82	120,91	-1,96	-31,15	105,97
	6	174	89	117,92	-2,65	-20,80	114,07
	7	175	60	112,87	-4,25	-34,21	117,65
	8	176	84	107,67	-4,89	-16,21	107,03
	9	177	113	102,94	-4,78	8,81	109,14
	10	178	126	98,49	-4,56	24,93	118,04
	11	179	122	94,46	-4,21	23,42	109,26
	12	180	118	90,64	-3,95	24,33	108,63
1984	1	181	92	--	--	--	82,90
	2	182	86	--	--	--	58,76
	3	183	81	--	--	--	64,22
	4	184	84	--	--	--	47,72
	5	185	87	--	--	--	39,75
	6	186	90	--	--	--	46,16
	7	187	79	--	--	--	28,80
	8	188	96	--	--	--	42,85
	9	189	122	--	--	--	63,93
	10	190	120	--	--	--	76,11
	11	191	137	--	--	--	70,64
	12	192	154	--	--	--	67,61

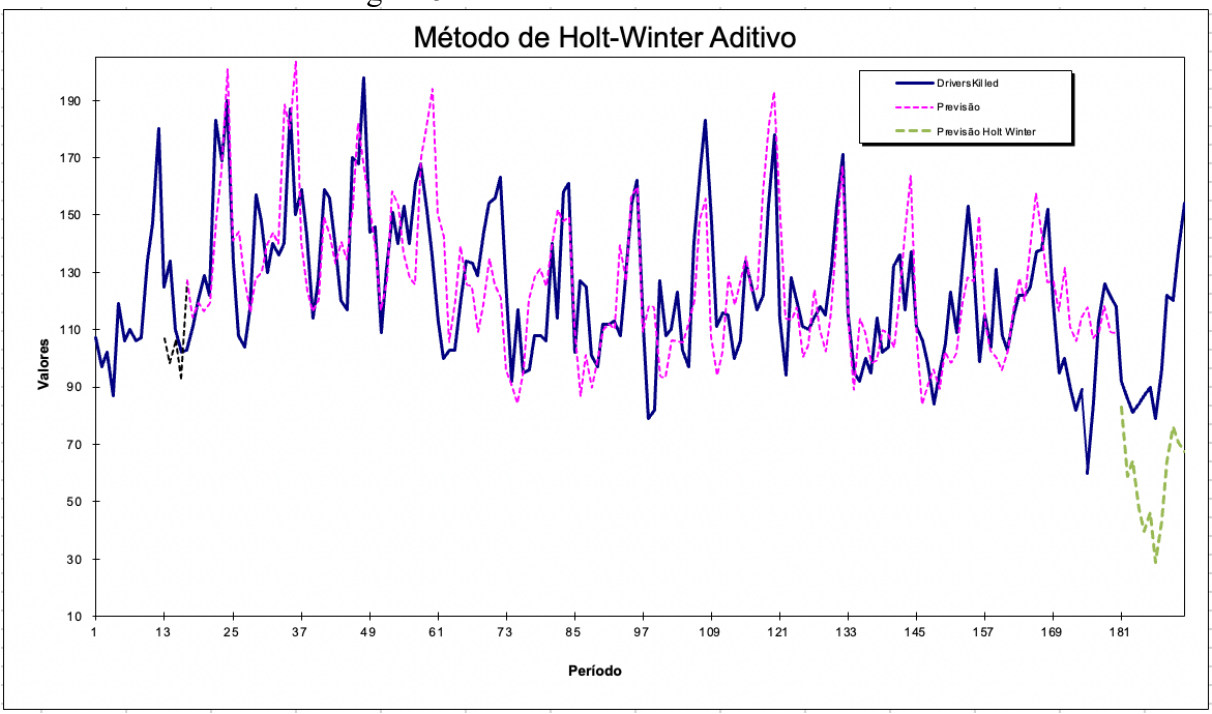
Fonte: Autor (2021).

Percebe-se que no primeiro ano (12 primeiros períodos) não foram calculados os valores para o nível base e nem para tendência, pois esses 12 primeiros períodos foram utilizados para o cálculo dos fatores sazonais iniciais.

4.2.4 Resultados do modelo Holt-Winters

A Figura 5 a seguir mostra os valores de previsão obtidos com esse modelo.

Figura 5 - Previsão modelo Holt-Winters.



Fonte: Autor (2021).

Como pode ser observado, o resultado do modelo de Holt-Winters nos períodos do primeiro conjunto de dados (treino) se mostrou mais fiel aos dados originais, conseguindo seguir e acompanhar os fatores de sazonalidade e tendência. Contudo, para os 12 períodos finais que foram separados para validação, a previsão se distanciou um pouco dos valores originais, porém é observado que o modelo seguiu uma certa sazonalidade.

A eficácia do modelo foi calculada utilizando-se o mesmo parâmetro descrito anteriormente. O resultado pode ser observado no Quadro 4 abaixo.

Quadro 4 - Valores RMSE da aplicação do modelo Holt-Winters.

RMSE Conjunto Treino	**20,6**
RMSE Conjunto Teste 3 Meses	**19,2**
RMSE Conjunto Teste 6 Meses	**33,1**
RMSE Conjunto Teste 12 Meses	**49,2**

Fonte: Autor (2021).

Apesar desse modelo conseguir captar as características de sazonalidade e apresentar isso nos resultados, o modelo não se mostrou tão eficiente, pois quanto menor o valor do RMSE, mais eficiente foi o modelo.

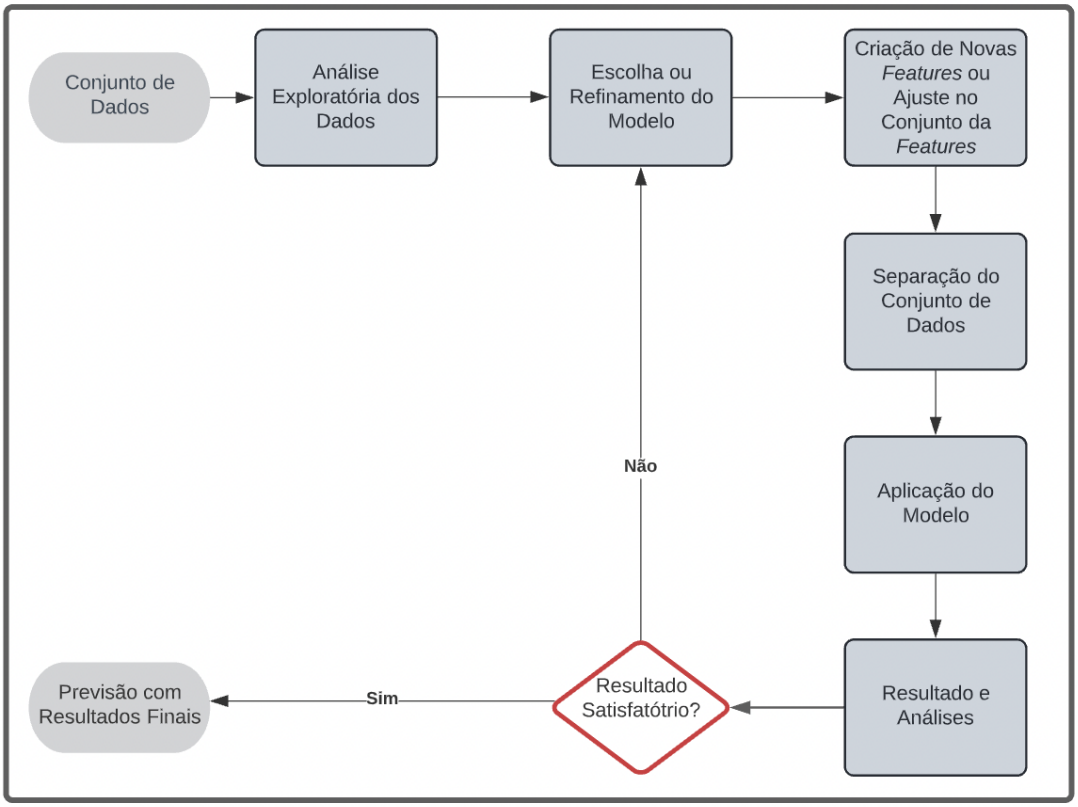
À medida que o horizonte de previsão aumenta, é natural que os erros sejam mais elevados para todos os modelos. Por isso, o RMSE considerando apenas os primeiros 3 e 6 meses do conjunto de testes também foi apresentado.

4.3 IMPLEMENTAÇÃO DO GUIA PARA APLICAÇAO DE MODELOS DE *MACHINE LEARNING*

Nesta seção serão executados os passos necessários para realizar a previsão e como preparar os dados para serem aplicados nos modelos de ML. Esses passos podem servir de um guia para auxiliar o leitor em outras aplicações semelhantes.

O guia com todas as etapas pode ser observado na Figura 6 abaixo.

Figura 6 - Etapas do guia para aplicação de modelos de ML.



Fonte: Autor (2021).

Como mostrado na Figura 6 acima, se os resultados obtidos não forem satisfatórios, o leitor deverá retornar para a etapa de escolha do modelo, podendo utilizar outro modelo para comparar os resultados ou se optar por utilizar o mesmo modelo, poderá seguir para etapas

seguintes fazendo ajustes no modelo, por exemplo, criar novas *features*, separar o conjunto de dados utilizando outra proporção, alterar os parâmetros do modelo, etc.

4.3.1 Aplicação do guia para modelo de regressão

4.3.1.1 Análise Exploratória dos Dados

A análise exploratória dos dados já foi realizada, e consta aqui somente para enfatizar sua presença no guia.

4.3.1.2 Escolha do modelo

Nos dias atuais, existem inúmeras opções disponíveis de modelos. Para escolher um modelo de *machine learning* é recomendado que seja usada a análise exploratória para estimar a complexidade do conjunto de dados, e verificar as características da série.

Esse é um processo iterativo, onde o modelo poderá ser refinado, outros modelos poderão ser testados, e as *features* poderão ser criadas e ajustadas.

Primeiramente, foi escolhido o modelo de regressão. O modelo de regressão será usado aqui para mostrar como um modelo simples de *machine learning* pode usar as *features* criadas pela decomposição da data.

4.3.1.3 Criação de novas features: com base na data

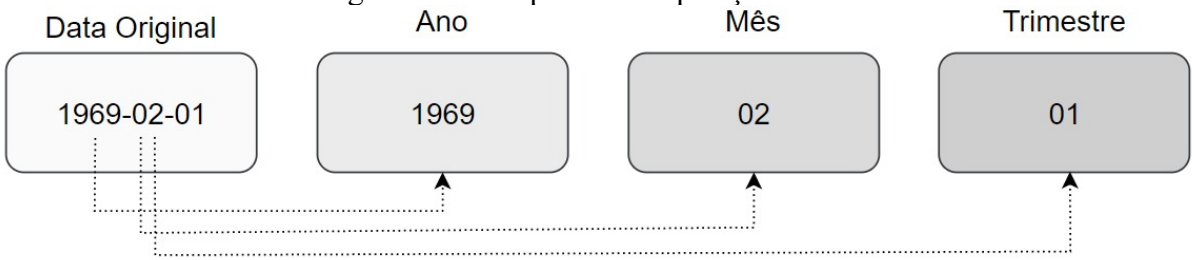
A primeira vantagem em usar os modelos de *machine learning* é a possibilidade de se incorporar outras variáveis independentes que podem contribuir para a previsão. Mas é importante salientar que outras colunas (“f*eatures*”) também podem ser incorporadas ao dataset, sendo essas novas colunas geradas pela combinação das colunas pré-existentes. As colunas geradas desta forma também são chamadas de novas “*features*”, e existe uma área de estudo deste tópico chamada de “*feature engineering*”.

Nesta seção e nas seções a seguir é mostrado as novas *features* criadas, começando pela decomposição da coluna de data e em seguida será mostrado as outras *features* criadas, explicando o motivo de cada coluna nova e como que essas colunas poderão auxiliar os modelos de *machine learning.*

A decomposição da data é o método de decomposição mais tradicional e é facilmente normalmente encontrado na literatura. Ele consiste em decompor a data em várias colunas, como por exemplo dia, mês, trimestre, quadrimestre, dia da semana, e etc.

A decomposição da coluna data em novas colunas foi feita separando os meses, os anos e o trimestre, em três novas colunas. A coluna de tempo (“Date”) contém o ano, mês e dia da observação, porém como o dia não varia de um mês para outro, não traria informações relevantes para o modelo, dessa forma foram criadas três novas colunas utilizando a coluna Date como base, como ilustrado na Figura 7 a seguir.

Figura 7 - Exemplos decomposição da data.



Fonte: Autor (2021).

A porção superior da base de dados com a decomposição da data pode ser vista na Tabela 4 na página seguinte. A tabela completa pode ser vista no Apêndice D.

Tabela 4 - Decomposição da data em ano e mês e trimestre.

Date	DriversKilled	kms	PetrolPrice	law	Ano	Mês	Trimestre
1969-01-01	107	9059	0,103	0	69	1	1
1969-02-01	97	7685	0,102	0	69	2	1
1969-03-01	102	9963	0,102	0	69	3	1
1969-04-01	87	10955	0,101	0	69	4	2
1969-05-01	119	11823	0,101	0	69	5	2
1969-06-01	106	12391	0,101	0	69	6	2
1969-07-01	110	13460	0,104	0	69	7	3
1969-08-01	106	14055	0,104	0	69	8	3
1969-09-01	107	12106	0,104	0	69	9	3
1969-10-01	134	11372	0,103	0	69	10	4
1969-11-01	147	9834	0,103	0	69	11	4
1969-12-01	180	9267	0,102	0	69	12	4
1970-01-01	125	9130	0,101	0	70	1	1
1970-02-01	134	8933	0,101	0	70	2	1
1970-03-01	110	11000	0,100	0	70	3	1
1970-04-01	102	10733	0,099	0	70	4	2
1970-05-01	103	12912	0,098	0	70	5	2
1970-06-01	111	12926	0,098	0	70	6	2
1970-07-01	120	13990	0,097	0	70	7	3
1970-08-01	129	14926	0,097	0	70	8	3
1970-09-01	122	12900	0,097	0	70	9	3
1970-10-01	183	12034	0,096	0	70	10	4
1970-11-01	169	10643	0,096	0	70	11	4
1970-12-01	190	10742	0,095	0	70	12	4
1971-01-01	134	10266	0,097	0	71	1	1
1971-02-01	108	10281	0,096	0	71	2	1
1971-03-01	104	11527	0,095	0	71	3	1
1971-04-01	117	12281	0,095	0	71	4	2
1971-05-01	157	13587	0,094	0	71	5	2
1971-06-01	148	13049	0,094	0	71	6	2
1971-07-01	130	16055	0,093	0	71	7	3
1971-08-01	140	15220	0,093	0	71	8	3
1971-09-01	136	13824	0,093	0	71	9	3
1971-10-01	140	12729	0,092	0	71	10	4
1971-11-01	187	11467	0,092	0	71	11	4

Fonte: Autor (2021).

Essa ação foi feita para separar e analisar qual a influência os meses do ano, os anos e os diferentes trimestres, têm para com a sazonalidade, tendência e os ciclos de repetição. Uma observação importante em relação a esse conjunto de dados em específico, é que os dados são mensais e, portanto, os dias são sempre 1, representando a coleta dos dados sempre no mesmo dia do mês, porém para outras bases de dados com valores diários diferentes a separação e análise dos dias será importante.

Essa decomposição, por exemplo, permite que o modelo de *machine learning* tenha maior chance de encontrar outros períodos ou condições específicas que influenciariam na sazonalidade, além daquele presumido pelo analista.

4.3.1.4 Separação dos dados em conjunto de treino e conjunto de teste

Com todas as *features* criadas e o conjunto de dados preparado, foi usado o mesmo critério anterior para separar o conjunto de dados em conjunto treino e conjunto teste.

4.3.1.5 Aplicação do modelo de regressão

Na Tabela 5 é mostrado a aplicação do modelo. A tabela completa com todos os valores e resultados da aplicação do modelo de regressão, pode ser encontrada no Apêndice E.

Tabela 5 - Aplicação modelo de regressão.

Ano	Período	Mês	kms	PetrolPrice	Ano	Trimestre	DriversKilled	Valor Previsto
1969	1	1	9059	0,103	69	1	107	116,64
	2	2	7685	0,102	69	1	97	124,65
	3	3	9963	0,102	69	1	102	118,69
	4	4	10955	0,101	69	2	87	124,98
	5	5	11823	0,101	69	2	119	124,19
	6	6	12391	0,101	69	2	106	124,50
	7	7	13460	0,104	69	3	110	128,15
	8	8	14055	0,104	69	3	106	128,36
	9	9	12106	0,104	69	3	107	137,89
	10	10	11372	0,103	69	4	134	150,51
	11	11	9834	0,103	69	4	147	158,53
	12	12	9267	0,103	69	4	180	163,00
1970	13	1	9130	0,101	70	1	125	118,21
	14	2	8933	0,101	70	1	134	121,32
	15	3	11000	0,1	70	1	110	116,72
	16	4	10733	0,099	70	2	102	127,63
	17	5	12912	0,098	70	2	103	122,62
1983	169	1	16231	0,113	83	1	120	93,62
	170	2	15511	0,114	83	1	95	98,05
	171	3	18308	0,113	83	1	100	90,78
	172	4	17793	0,118	83	2	89	99,06
	173	5	19205	0,118	83	2	82	96,27
	174	6	19162	0,118	83	2	89	98,82
	175	7	20997	0,12	83	3	60	100,25
	176	8	20705	0,119	83	3	84	104,30
	177	9	18759	0,119	83	3	113	113,82
	178	10	19240	0,118	83	4	126	121,99
	179	11	17504	0,118	83	4	122	130,74
	180	12	16591	0,118	83	4	118	136,47
1984	181	1	16224	0,118	84	1	92	91,35
	182	2	16670	0,115	84	1	86	93,87
	183	3	18539	0,116	84	1	81	88,82
	184	4	19759	0,115	84	2	84	94,28
	185	5	19584	0,115	84	2	87	97,31
	186	6	19976	0,115	84	2	90	98,26
	187	7	21486	0,115	84	3	79	102,06
	188	8	21626	0,115	84	3	96	103,94
	189	9	20195	0,114	84	3	122	112,16
	190	10	19928	0,116	84	4	120	121,30
	191	11	18564	0,116	84	4	137	128,69
	192	12	18149	0,116	84	4	154	132,60

Fonte: Autor (2021).

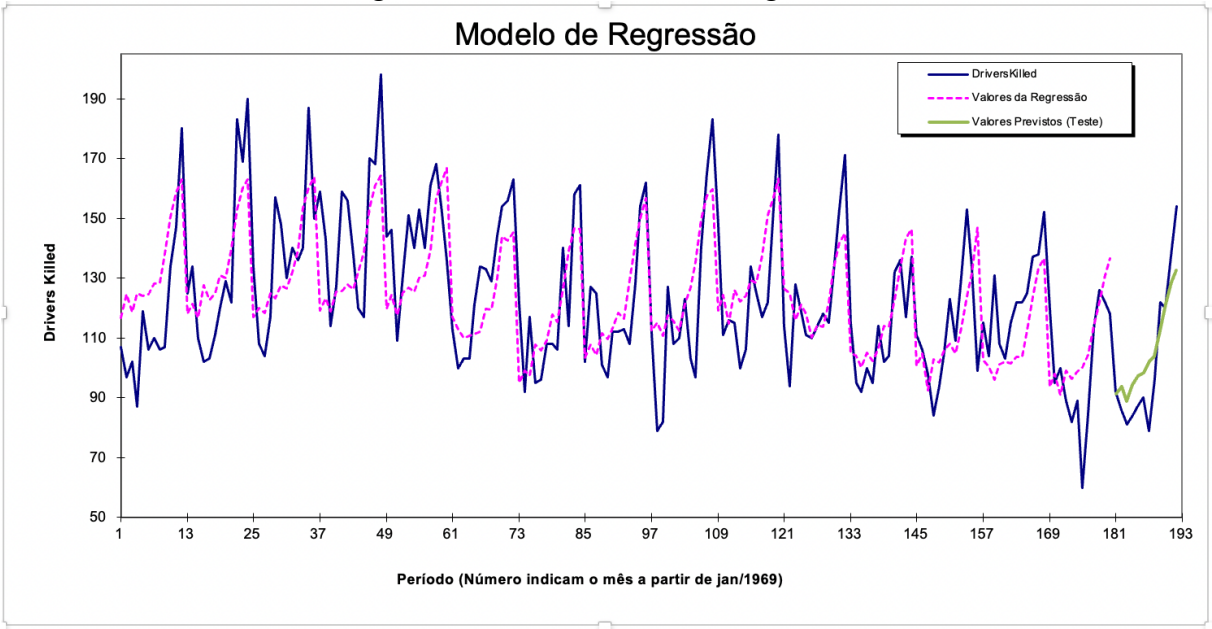
Primeiramente, foi feita a decomposição da data, criando uma coluna para o mês, uma coluna para o ano e uma coluna para o trimestre. Por fim, a coluna “Valor Previsto”, representa a previsão. A previsão de 1969 foi feita com base nos valores do conjunto treino e a previsão dos últimos 12 períodos (ano de 1984) foi feita com os valores do conjunto teste.

Como feito nos modelos tradicionais, foi separado os últimos 12 períodos para o conjunto teste e o períodos iniciais para o conjunto treino, a fim de aplicar a mesma divisão para todos os modelos.

4.3.1.6 Resultados do modelo de regressão

O resultado deste modelo de regressão pode ser observado na Figura 8 a seguir.

Figura 8 - Previsão modelo de regressão.



Fonte: Autor (2021).

Verifica-se que este modelo realizou uma previsão mais próxima dos valores originais, conseguindo seguir a tendência da série e sazonalidade.

O RMSE obtido com a aplicação desse modelo é apresentado no Quadro 5 seguir.

Quadro 5 - Valores RMSE da aplicação do modelo de regressão.

RMSE Conjunto Treino	**16,9**
RMSE Conjunto Teste 3 Meses	**4,5**
RMSE Conjunto Teste 6 Meses	**3,7**
RMSE Conjunto Teste 12 Meses	**11,9**

Fonte: Autor (2021).

Com o resultado obtido do RMSE, constata-se o que foi analisado no Quadro 5. Os valores previstos para o conjunto treino e teste, ficaram bem próximos aos valores originais. Também pode ser observado que esse modelo seguiu e acompanhou as características da série.

Na seção seguinte será aplicado o modelo XGBoost usando os mesmos dados submetidos ao modelo de regressão. Isso permitirá analisar a performance de modelos mais poderosos em contextos de ML, mas que podem precisar de um número maior de *features* para melhorar seu desempenho.

4.3.2 Aplicação do guia para modelo XGBoost

4.3.2.1 Análise Exploratória dos Dados

A análise exploratória dos dados já foi realizada, e consta aqui somente para enfatizar sua presença no guia.

4.3.2.2 Escolha do modelo

Nessa etapa foi escolhido outro modelo de *machine learning*, o XGBoost. Esse modelo foi escolhido pela sua boa capacidade de lidar com um número grande de variáveis independentes no formato numérico. Como o número de colunas no dataset irá aumentar, é interessante migrar para um modelo que lida melhor com as correlações entre um número grande variáveis independentes. Outra vantagem é a possibilidade se analisar a importância das variáveis no resultado da previsão realizada pelo modelo.

4.3.2.3 Criação de novas features

Essa etapa já foi realizada anteriormente, e para esse modelo foi utilizado as mesmas *features* criadas para o modelo de regressão, porém para efeitos de entendimento do guia está explicita aqui essa etapa.

4.3.2.4 Separação dos dados em conjunto de treino e conjunto de teste

Com todas as *features* criadas e o conjunto de dados preparado, foi usado o mesmo critério anterior para separar o conjunto de dados em conjunto treino e conjunto teste. A separação da base de dados é mostrada na Figura 9 abaixo.

Figura 9 - Separação do conjunto de dados.

	Date	DriversKilled	kms	PetrolPrice	law	Ano	Mês	Trimestre
LINHA 1	1970-01-01	125	9130	0,1010	0	70	1	1
LINHA 2	1970-02-01	134	8933	0,1010	0	70	2	1
LINHA 3	1970-03-01	110	11000	0,1000	0	70	3	1
LINHA 4	1970-04-01	102	10733	0,0990	0	70	4	2
LINHA 5	1970-05-01	103	12912	0,0980	0	70	5	2
LINHA 6	1970-06-01	111	12926	0,0980	0	70	6	2
LINHA 7	1970-07-01	120	13990	0,0970	0	70	7	3
LINHA 165	1983-09-01	113	18759	0,1190	1	83	9	3
LINHA 166	1983-10-01	126	19240	0,1180	1	83	10	4
LINHA 167	1983-11-01	122	17504	0,1180	1	83	11	4
LINHA 168	1983-12-01	118	16591	0,1180	1	83	12	4
LINHA 169	1984-01-01	92	16224	0,1180	1	84	1	1
LINHA 170	1984-02-01	86	16670	0,1150	1	84	2	1
LINHA 171	1984-03-01	81	18539	0,1160	1	84	3	1
LINHA 172	1984-04-01	84	19759	0,1150	1	84	4	2
LINHA 173	1984-05-01	87	19584	0,1150	1	84	5	2
LINHA 174	1984-06-01	90	19976	0,1150	1	84	6	2
LINHA 175	1984-07-01	79	21486	0,1150	1	84	7	3
LINHA 176	1984-08-01	96	21626	0,1150	1	84	8	3
LINHA 177	1984-09-01	122	20195	0,1140	1	84	9	3
LINHA 178	1984-10-01	120	19928	0,1160	1	84	10	4
LINHA 179	1984-11-01	137	18564	0,1160	1	84	11	4
LINHA 180	1984-12-01	154	18149	0,1160	1	84	12	4

Conjunto Treino

Conjunto Teste

Fonte: Autor (2021).

Observar-se a divisão do conjunto de dados com os valores referentes ao último ano (1984) para o conjunto teste e os valores iniciais para o conjunto treino, conforme feito nos modelos anteriores.

4.3.2.5 Aplicação do Modelo XGBoost

Para aplicação do modelo XGBoost primeiramente definiu-se os parâmetros. Foram explorados conjuntos de parâmetros que continham pequenas variações, e o melhor conjunto é descrito abaixo:

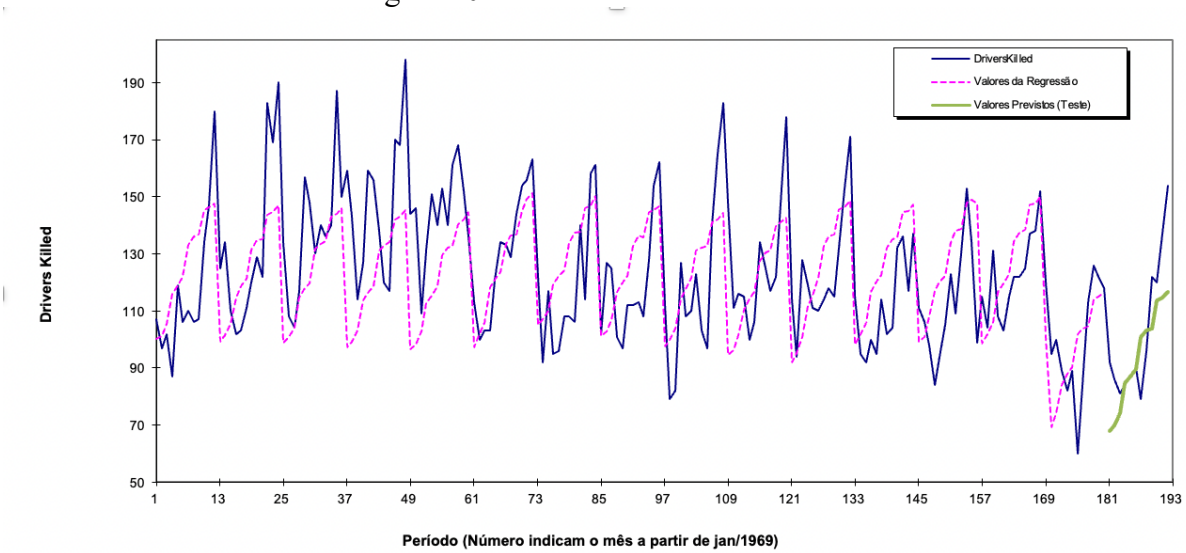
- booster = "gblinear"
- eta = 0.01
- max_depth = 10
- subsample = 0.7
- colsample_bytree = 0.7
- gamma = 0.2

Após a definição dos parâmetros, os dados foram submetidos ao modelo para realizar o treinamento e validação.

4.3.2.6 Resultados XGBoost

Os resultados obtidos com o XGBoost na configuração explicada anteriormente podem ser vistos na Figura 10 abaixo.

Figura 10 - Previsão modelo XGBoost.



Fonte: Autor (2021).

Observa-se que este modelo realizou uma previsão, conseguindo seguir a tendência da série e sazonalidade.

O RMSE com a aplicação do XGBoost é apresentado no Quadro 6 a seguir.

Quadro 6 - Valores RMSE da aplicação do modelo XGBoost com dados usados na regressão.

RMSE Conjunto Treino	**21,3**
RMSE Conjunto Teste 3 Meses	**17,2**
RMSE Conjunto Teste 6 Meses	**12,2**
RMSE Conjunto Teste 12 Meses	**17,4**

Fonte: Autor (2021).

Com base nos resultados encontrados, optou-se por refinar o modelo para melhorar o resultado obtido, visto que o modelo XGBoost foi projetado para trabalhar com grande quantidade de dados.

A seguir foi feito o refinamento do modelo, retornando à segunda etapa do guia.

4.3.2.7 Refinamento do modelo XGBoost

O modelo XGBoost foi mantido, bem como os parâmetros anteriores, mas agora foi feita a inclusão de *features* adicionais. O objetivo é mostrar que modelos mais complexos podem precisar de novas *features* para melhorar seu desempenho.

4.3.2.8 Criação e inserção de features adicionais

Algumas vezes as *features* geradas a partir da data não são suficientes para que o modelo consiga capturar a magnitude das correlações com a sequência histórica dos dados. Para tentar melhorar o conjunto de dados, foram criadas mais *features*, dessa vez baseadas nos dados de interesse.

O passo realizado nesta seção foi de criar novas colunas de dados utilizando como base os dados já existentes, o que é usualmente é chamado de produzir novas *features*, ou traduzido do inglês para o português, produzir novas colunas de variáveis.

Essa atitude foi feita com o objetivo de trazer mais variáveis para auxiliar o modelo de *machine learning* a detectar as características das séries temporais para atingir melhores resultados, lembrando que os modelos tradicionais normalmente só consideram uma variável, por isso essa etapa não foi feita para os modelos tradicionais. A seguir será explicado como criar essas novas colunas de dados.

Como nos modelos anteriores, serão previstos os doze últimos períodos da série. Dito isso, um ponto importante na criação das *features* é cria-las com base no horizonte de valores que se deseja prever, que neste caso serão doze períodos. Dito isso, nas seções seguintes foram criadas as novas *features* levando em consideração doze períodos de análise.

4.3.2.8.1 Criação das colunas defasadas no tempo

A criação de variáveis defasadas no tempo é um artifício utilizado para tentar trazer ao modelo de ML uma correlação entre os valores das linhas, reconhecendo padrões existentes entre os valores da série temporal.

É necessário antes de submeter os dados ao modelo de *machine learning*, fazer esse ajuste criando novas variáveis defasadas no tempo, para trazer uma análise temporal dos dados, modelando o problema para poder aplicar regressão. Essa estratégia traz muitas vantagens porque possibilita utilizar algoritmos mais complexos de *machine learning* em séries temporais.

A estratégia consiste em trazer os dados de períodos posteriores (linhas diferentes) para uma mesma linha, transformando assim o conjunto de dados. Desta forma, foram criadas 78 novas *features*, para que o leitor pudesse ver de forma aplicada essa abordagem, a importância e relevância para os modelos de *machine learning*.

Inicialmente, foram criadas novas colunas denominadas de "*lag*", traduzido do inglês essa palavra significa "atraso". Esta abordagem foi feita a fim de defasar cada nova coluna em uma unidade de tempo em relação às anteriores.

O valor das colunas defasadas é igualmente equivalente aos valores da coluna "DriversKilled" (coluna que contém os valores a serem previstos).

Essa técnica de defasar os valores no tempo é para investigar uma possível dependência da série em relação ao tempo (como ciclos) em uma série temporal, a sazonalidade e a tendência da série. Para isso, criam-se cópias "defasadas" da série. Atrasar uma série temporal significa deslocar seus valores uma ou mais etapas para trás.

O efeito é que as observações na série defasada parecerão ter acontecido mais tarde, dessa forma, cada linha conterá dados sobre uma observação que inclui todas as ocorrências anteriores dessa observação.

Ao atrasar uma série temporal é feito com que seus valores passados sejam trazidos para mesma linha, isso torna as séries defasadas úteis como recursos para modelar a dependência temporal da série.

Observa-se o resultado, após feita a criação das novas colunas na Tabela 6 abaixo.

Tabela 6 – Conjunto de dados com a criação das *lags*.

Date	DriversKilled	lag_01	lag_02	lag_03	lag_04	lag_05	lag_06	lag_07
1969-01-01	107							
1969-02-01	97	107						
1969-03-01	102	97	107					
1969-04-01	87	102	97	107				
1969-05-01	119	87	102	97	107			
1969-06-01	106	119	87	102	97	107		
1969-07-01	110	106	119	87	102	97	107	
1969-08-01	106	110	106	119	87	102	97	107
1969-09-01	107	106	110	106	119	87	102	97
1969-10-01	134	107	106	110	106	119	87	102
1969-11-01	147	134	107	106	110	106	119	87
1969-12-01	180	147	134	107	106	110	106	119
1970-01-01	125	180	147	134	107	106	110	106
1970-02-01	134	125	180	147	134	107	106	110
1970-03-01	110	134	125	180	147	134	107	106
1970-04-01	102	110	134	125	180	147	134	107
1970-05-01	103	102	110	134	125	180	147	134
1970-06-01	111	103	102	110	134	125	180	147
1970-07-01	120	111	103	102	110	134	125	180
1970-08-01	129	120	111	103	102	110	134	125
1970-09-01	122	129	120	111	103	102	110	134

Fonte: Autor (2021).

As novas *features* criadas podem ser observadas nas colunas desde a coluna “lag_1” até a coluna “lag_12”.

Como dito anteriormente, a coluna “lag_1” estão com os valores iguais os da coluna “DriversKilled”, porém com uma unidade de tempo defasados, que no caso estão com os dados dos períodos anteriores. Seguindo a mesma lógica até a última coluna, pode ser observado que os valores estão defasados em doze períodos.

Foi escolhido doze períodos, pois foi o período de tempo utilizado nos modelos anteriores para realizar a previsão, dessa forma, continuou-se adotando o mesmo padrão neste modelo.

4.3.2.8.2 Criação das colunas defasadas no tempo com valores médios

Seguindo, o próximo passo foi a criação de outra *feature* que foi denominada de “*rollmean*”, que “mean” significa média e “roll” pode ser traduzido como rolar. Essa nova *feature* nada mais é do que a média dos valores da coluna “DriversKilled” que, seguindo a

mesma lógica do item anterior, também foram defasados no tempo, porém com os valores da média dos períodos anteriores, como pode-se observar na Tabela 7 a seguir.

Tabela 7 – Conjunto de dados com a criação das colunas rollmean.

Date	DriversKilled	rollmean_02	rollmean_03	rollmean_04
1969-01-01	107,00			
1969-02-01	97,00			
1969-03-01	102,00	102,00		
1969-04-01	87,00	99,50	102,00	
1969-05-01	119,00	94,50	95,33	98,25
1969-06-01	106,00	103,00	102,67	101,25
1969-07-01	110,00	112,50	104,00	103,50
1969-08-01	106,00	108,00	111,67	105,50
1969-09-01	107,00	108,00	107,33	110,25
1969-10-01	134,00	106,50	107,67	107,25
1969-11-01	147,00	120,50	115,67	114,25
1969-12-01	180,00	140,50	129,33	123,50
1970-01-01	125,00	163,50	153,67	142,00
1970-02-01	134,00	152,50	150,67	146,50
1970-03-01	110,00	129,50	146,33	146,50
1970-04-01	102,00	122,00	123,00	137,25

Fonte: Autor (2021).

Um ponto muito relevante a ser observado é que as médias são feitas da seguinte forma: a coluna “rollmean_02” é a média dos dois valores anteriores da coluna original “DriversKilled”, a coluna “rollmean_03” é a média dos três últimos valores e assim por diante.

Foram criadas onze novas colunas, com a primeira sendo a “rollmean_02”, que seria a média dos dois valores anteriores, até a coluna “rollmean_12”, contendo o valor médio dos doze períodos anteriores.

4.3.2.8.3 Criação das colunas defasadas no tempo com valores de desvio padrão

Seguindo com a criação das *features*, foi feita a mesma lógica de criação de novas colunas defasadas no tempo, porém neste caso os valores inseridos foram os valores do desvio padrão.

Na Tabela 8 abaixo é ilustrado como foi feito a criação dessas 12 novas colunas.

Tabela 8 – Conjunto de dados com a criação das colunas rollsd

Date	DriversKilled	rollsd_02	rollsd_03	rollsd_04	rollsd_05
1969-01-01	107,00				
1969-02-01	97,00				
1969-03-01	102,00	7,07			
1969-04-01	87,00	3,54	5,00		
1969-05-01	119,00	10,61	7,64	8,54	
1969-06-01	106,00	22,63	16,01	13,38	11,87
1969-07-01	110,00	9,19	16,09	13,18	11,78
1969-08-01	106,00	2,83	6,66	13,48	11,78
1969-09-01	107,00	2,83	2,31	6,13	11,67
1969-10-01	134,00	0,71	2,08	1,89	5,50
1969-11-01	147,00	19,09	15,89	13,28	12,07
1969-12-01	180,00	9,19	20,40	20,34	18,62
1970-01-01	125,00	23,33	23,71	30,32	30,80

Fonte: Autor (2021).

Como observado, a coluna “rollsd_02” foi preenchida com os valores dos desvios padrão dos dois períodos anteriores da coluna contendo os valores originais, a coluna “rollsd_03”, seguindo a mesma linha de raciocínio, foi preenchida com o valor do desvio padrão dos três períodos anteriores da coluna original e assim por diante.

Foram criadas onze novas colunas, ou seja, a última coluna “rollsd_12” foi preenchida com o desvio padrão dos doze valores anteriores. Outro exemplo, a coluna “rollsd_09” foi preenchida levando em consideração o cálculo do desvio padrão dos nove períodos anteriores e assim sucessivamente.

4.3.2.8.4 Criação das colunas defasadas no tempo com valores de desvio padrão, valores máximos, a diferença e a divisão

Por fim, foram criados conjunto de quatro tipos de *features,* para as combinações de X e Y onde X>1, Y>X e Y<13:

- lagsd_XtoY: desvio padrão considerando valores da LagX e da LagY;
- lag_max_XtoY: valor máximo entre todos os valores entre LagX e LagY;
- lagdiff_XtoY: valor da diferença entre a LagY e a LagX;
- lagdiv_XtoY: valor da divisão de LagY pela LagX.

A Tabela 9 apresenta um exemplo dos cálculos.

Tabela 9 – Conjunto de dados com a criação das colunas lagsd, lagmax, lagdiff e lagdiv.

Date	DriversKilled	lagsd_1to3	lagmax_1to3	lagdiff_1to3	lagdiv_1to3
1969-01-01	107				
1969-02-01	97				
1969-03-01	102				
1969-04-01	87	3,54	107,00	-5,00	0,95
1969-05-01	119	7,07	102,00	-10,00	0,90
1969-06-01	106	12,02	119,00	17,00	1,17
1969-07-01	110	13,44	119,00	19,00	1,22
1969-08-01	106	6,36	119,00	-9,00	0,92
1969-09-01	107	0,00	110,00	0,00	1,00

Fonte: Autor (2021).

Todas as técnicas acima aplicadas foram feitas visando criar o máximo de variáveis que trazem uma perspectiva de temporalidade para o modelo de *machine learning* na tentativa de melhorar a eficácia do modelo e fazer com que o modelo observe as características da série temporal analisada. Essa abordagem pode ser replicada para qualquer base de dados de séries temporais.

4.3.2.8.5 Exclusão das linhas com valores faltantes e ajuste no conjunto das *features*

Juntamente com a criação das novas *features*, criou-se valores faltantes nas linhas por conta da defasagem dos valores. Dessa forma, para que esses dados faltantes não interferissem negativamente na previsão, removeu-se as linhas que continham esses valores faltantes. O resultado é visto na Tabela 10 a seguir, que mostra parte da porção superior da base de dados para elucidar a supressão das linhas com os valores faltantes (12 primeiros períodos).

Tabela 10 – Parte superior do conjunto de dados com a remoção das linhas contendo dados faltantes.

Date	DriversKilled	kms	PetrolPrice	law	Ano	Mês	Trimestre	lag_01	lag_02	lag_03	lag_04	lag_05
1970-01-01	125	9130	0,1010	0	70	1	1	180,00	147,00	134,00	107,00	106,00
1970-02-01	134	8933	0,1010	0	70	2	1	125,00	180,00	147,00	134,00	107,00
1970-03-01	110	11000	0,1000	0	70	3	1	134,00	125,00	180,00	147,00	134,00
1970-04-01	102	10733	0,0990	0	70	4	2	110,00	134,00	125,00	180,00	147,00
1970-05-01	103	12912	0,0980	0	70	5	2	102,00	110,00	134,00	125,00	180,00
1970-06-01	111	12926	0,0980	0	70	6	2	103,00	102,00	110,00	134,00	125,00
1970-07-01	120	13990	0,0970	0	70	7	3	111,00	103,00	102,00	110,00	134,00
1970-08-01	129	14926	0,0970	0	70	8	3	120,00	111,00	103,00	102,00	110,00
1970-09-01	122	12900	0,0970	0	70	9	3	129,00	120,00	111,00	103,00	102,00
1970-10-01	183	12034	0,0960	0	70	10	4	122,00	129,00	120,00	111,00	103,00
1970-11-01	169	10643	0,0960	0	70	11	4	183,00	122,00	129,00	120,00	111,00
1970-12-01	190	10742	0,0950	0	70	12	4	169,00	183,00	122,00	129,00	120,00
1971-01-01	134	10266	0,0970	0	71	1	1	190,00	169,00	183,00	122,00	129,00

Fonte: Autor (2021).

Foram excluídas as primeiras doze linhas da base de dados, pois a *feature* com a maior defasagem foi de doze períodos, consequentemente as linhas que continham valores faltantes foram as doze primeiras.

Feito isso, a primeira observação do dataset passou a ser janeiro de 1970 e não mais janeiro de 1969.

4.3.2.9 Separação dos dados em conjunto de treino e conjunto de teste

Foi mantida a mesma proporção da separação feita anteriormente. Esse passo está aqui somente para fins didáticos da aplicação do guia.

4.3.2.10 Aplicação do modelo XGBoost com features adicionais

Foram mantidos os mesmos parâmetros anteriores, como mostrado abaixo:

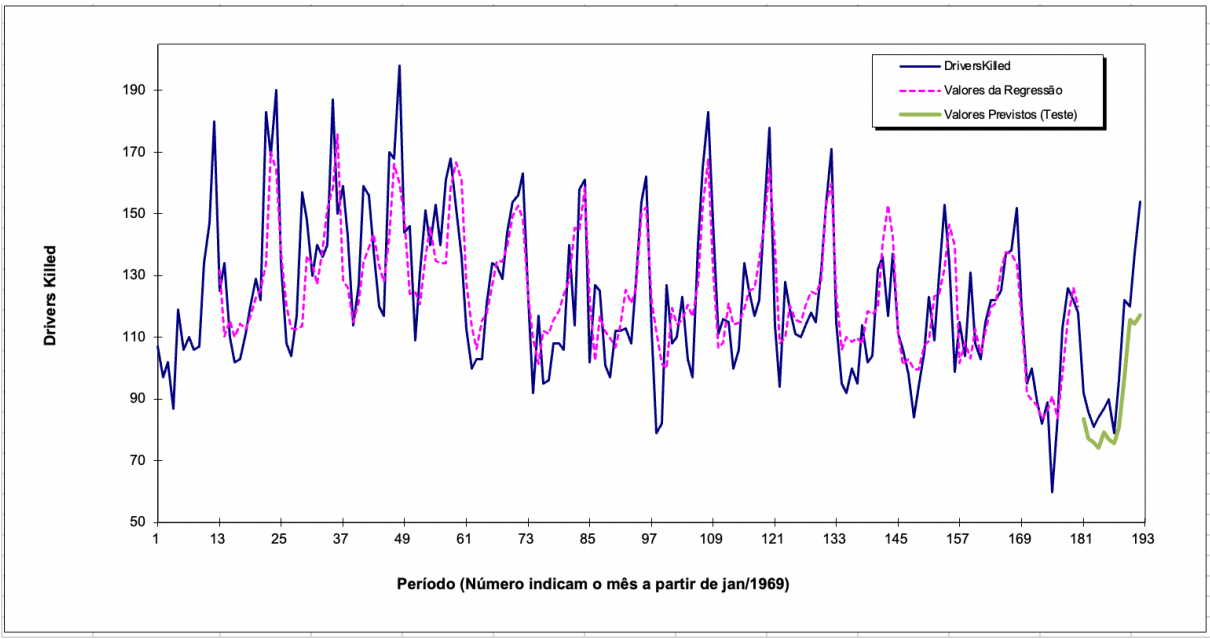
- booster = "gblinear"
- eta = 0.01
- max_depth = 10
- subsample = 0.7
- colsample_bytree = 0.7
- gamma = 0.2

4.3.2.11 Resultados XGBoost usando as features adicionais

Neste item serão apresentados os resultados obtidos do modelo, incluindo o RMSE, o gráfico de importância das variáveis, gráfico do valor previsto em comparação com os valores do conjunto teste e o gráfico de resíduos.

O resultado da previsão comparado com os valores reais do conjunto de dados treino e teste está exposto na Figura 11 abaixo.

Figura 11 – Previsão do modelo XGBoost.



Fonte: Autor (2021).

Com o resultado obtido, fica claro visualizar que a previsão seguiu a tendência de queda e também acompanhou a sazonalidade, comprovando graficamente os resultados calculados de RMSE.

O resultado do RMSE da aplicação do XGBoost com os ajustes feitos pode ser verificado no Quadro 7 na página seguinte.

Quadro 7 - Valores RMSE da aplicação do modelo XGBoost.

RMSE Conjunto Treino	**14,8**
RMSE Conjunto Teste 3 Meses	**7,7**
RMSE Conjunto Teste 6 Meses	**9,3**
RMSE Conjunto Teste 12 Meses	**16,5**

Fonte: Autor (2021).

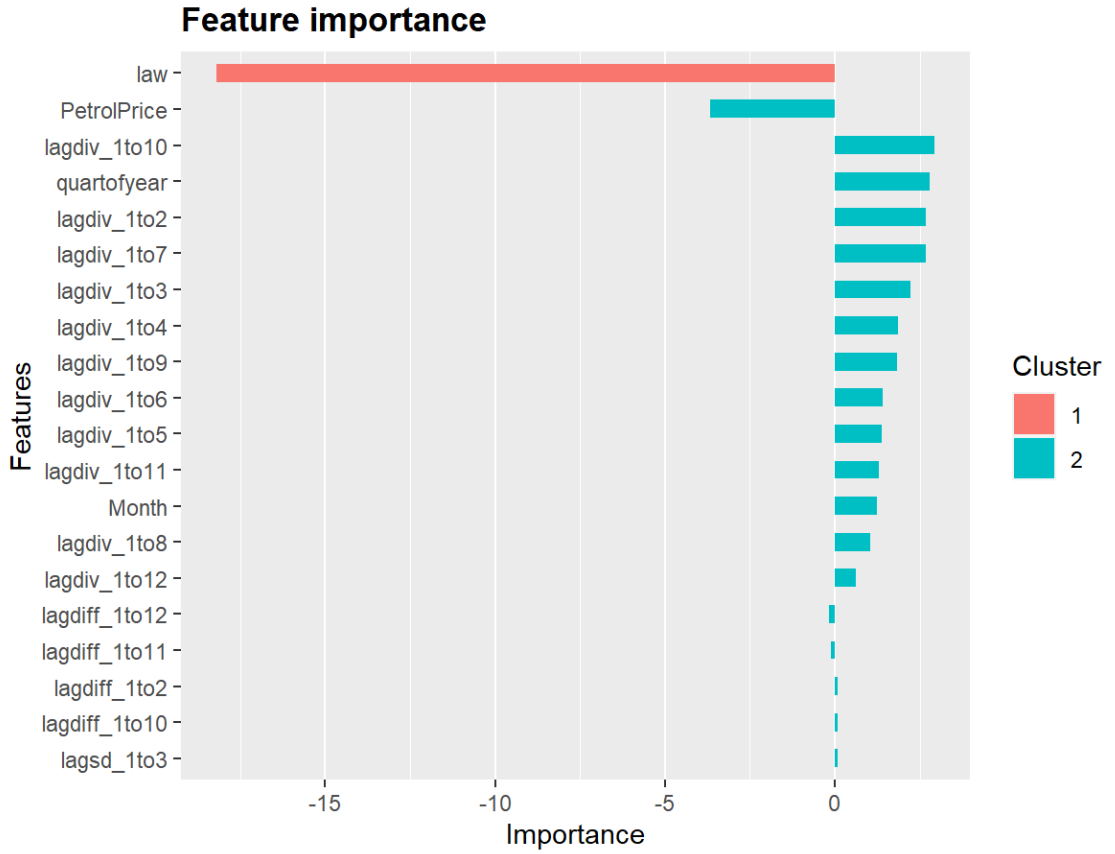
4.3.2.11.1 Gráfico de importância das variáveis

Uma análise interessante para saber quais variáveis têm maior influência sobre o resultado modelo é chamada de análise de importância das variáveis. No caso do algoritmo XGBoost, um gráfico da importância poderá ser gerado, permitindo visualizar quais foram as variáveis que mais impactaram na previsão de forma quantitativa.

Dessa forma, as variáveis com o maior nível de importância devem ser analisadas e tratadas com maior atenção, pois elas são as variáveis que mais afetaram os resultados do modelo e consequentemente a previsão.

A Figura 12 a seguir apresenta o gráfico de importância.

Figura 12 - Importância das variáveis.



Fonte: Autor (2021).

Percebeu-se que de fato a variável mais impactante ao modelo foi a variável *"law"*, ou seja, a coluna com os valores referentes a eficácia da lei que obriga motoristas a fazerem uso de cinto de segurança, confirmando que se a lei estava ou não em vigor no momento do acidente é um fator que tem grande influência para o modelo. Vale a pena salientar que o valor da influência é negativo, ou seja, quanto maior o valor da coluna "law", menor o valor de "DriversKilled", porém, como visto anteriormente, essa coluna pode ter o valor 0 ou 1. O valor 0 representando que a lei do uso de cinto de segurança não estava em vigor e 1 representando que a lei estava em vigor no momento do acidente. Dito isso, o resultado mostra que a lei foi efetiva e reduziu o número de mortes.

O preço do petróleo junto com as variáveis "lagdiv" tem grande relevância para este modelo. Contudo, vale lembrar que esse modelo usa bastante aleatoriedade, e a ordem de importância entre variáveis com valores próximos pode mudar a cada vez que é executado.

A variável "quartofyear", que representa o trimestre e a variável "Month" que se refere a coluna mês, também foram apontadas como relevantes indicando uma forte relação da sazonalidade com os meses do ano, o que foi também observado anteriormente na Figura 3.

Seguindo com a análise das variáveis que mais impactaram o modelo, nota-se a presença das *features* criadas como uma das maiores parcelas na representatividade da previsão, provando que essa estratégia da criação das novas *features* realmente foi efetiva.

Também se notou que a variável “PetrolPrice” e “law” apareceram entre as mais relevantes, trazendo uma atenção para a importância de realizar uma análise multivariada. Análise essa que leva em consideração as variáveis independentes presentes no *dataset* original, constatando que não somente as *features* criadas, mas as variáveis independentes têm grande importância para previsão, que se deixadas de lado, podem acarretar em uma grande perda na precisão do resultado.

Esta etapa final é de validação, pois se os resultados obtidos estiverem próximos ao objetivo do analista, o modelo estará preparado para realizar as previsões futuras. Nos casos em que a comparação gráfica dos valores do conjunto de validação com os valores resultantes do modelo não apresentar resultados satisfatórios, o leitor deverá retornar às etapas anteriores para realizar as alterações sugeridas, até atingir o resultado desejado.

O grau de exigência quanto aos resultados, varia de acordo com as necessidades do analista. Cabe ao analista avaliar se deve seguir com o modelo e os resultados obtidos ou se deve retomar e aperfeiçoar o modelo.

5 RESULTADOS

Este trabalho reforça o fato de que um modelo não será o melhor sempre. A cada conjunto de dados o analista deve sempre usar os modelos que julgar necessário. Algumas vezes o valor da previsão não vem de um único modelo, mas sim da combinação do resultado de vários modelos.

Vimos a grande vantagem de serem incluídas outras variáveis independentes aos modelos de previsão, quando comparamos os modelos de *machine learning* com os modelos tradicionais.

No caso do presente trabalho, como mostra a Tabela 11 abaixo, o modelo de regressão obteve a melhor performance entre todos os modelos em todos os horizontes de previsão. Mas também é importante verificar o impacto que as *features* adicionais causaram na performance do modelo XGBoost. Vale frisar que a previsão está sendo realizado para os próximos 12 meses, ou seja, um horizonte bem desafiador para qualquer modelo de previsão de séries temporais.

Os modelos tradicionais obtiveram os piores resultados em todos os horizontes de previsão, comprovando a superioridade dos modelos de *machine learning*.

Tabela 11 - Comparação dos resultados dos modelos.

Modelo	RMSE Conjunto Treinamento	RMSE Conjunto de Teste 3 meses	RMSE Conjunto de Teste 6 meses	RMSE Conjunto de Teste 12 meses
Holt	22,1	32,7	32,2	28,8
Holt-Winters	20,6	19,2	33,1	49,2
Regressão	16,9	4,5	3,7	11,9
XGBoost	21,3	17,2	12,2	17,4
XGBoost com Feat. Adicionais	14,8	7,7	9,3	16,5

Fonte: Autor (2021).

5.1 CONSIDERAÇÕES SOBRE OS MODELOS TRADICIONAIS

A principal vantagem dos modelos tradicionais é de serem de simples implementação e fácil entendimento. Não só o modelo é de fácil entendimento, mas também a análise dos resultados e coeficientes. E por esta razão ainda é bastante adotado no mercado.

Contudo, vale a pena salientar os seguintes pontos de fraqueza:

- Os coeficientes são encontrados por meio da otimização do erro de previsão, sem premissas estatísticas, como a normalidade dos resíduos em torno de zero. Ou seja, o resultado normalmente incorpora um "viés". Isso ocorre pois os parâmetros do modelo são otimizados levando-se em conta apenas o erro de previsão, seja qual for a medida escolhida;
- Ainda com relação a otimização realizada nos modelos tradicionais, à medida que a quantidade dos dados cresce, devido a não linearidade é comum existirem várias soluções que sejam ótimos locais, com o mesmo valor de erro. Contudo essas soluções podem gerar previsões futuras diferentes;
- Grande dependência das informações associadas a data da informação. O analista geralmente toma por base a frequência com que os dados são obtidos para definir um único período a ser utilizado nos modelos. Esse é o caso da referência para o cálculo do coeficiente de sazonalidade no modelo Holt-Winters. Contudo, além da imposição pelo modelo de se testar os períodos isoladamente, também podem existir outras "sazonalidades" nos dados não sincronizadas ao período principal escolhido pelo analista;
- Finalmente, o maior problema é o impedimento de incluir novas variáveis que podem estar correlacionadas com a variável de interesse.

5.2 CONSIDERAÇÕES SOBRE OS MODELOS DE *MACHINE LEARNING*

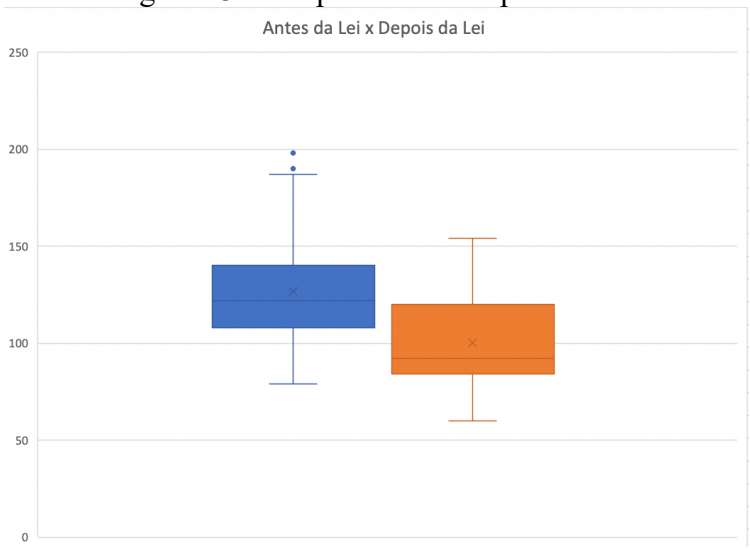
Inicialmente, observou-se que o modelo de regressão teve uma performance superior, indicando que os dados possuíam uma certa regularidade na sazonalidade, que foi corretamente capturada pelo modelo.

No caso do algoritmo XGBoost, este se beneficiou das *features* adicionais para alcançar um melhor resultado, ainda abaixo dos valores da regressão, mas com precisão razoável para um período menor, ou seja, modelos mais complexos de ML aplicados a séries temporais precisariam de mais *features* para obter melhores resultados.

Outro ponto que vale destaque é que o modelo XGBoost conseguiu identificar a importância das variáveis da "lei de uso cinto de segurança" (LAW) e do "Preço do combustível" (PetrolPrice) no resultado da previsão. No caso da variável "law", mesmo essa variável tendo ainda uma presença reduzida no *dataset*.

Levando em consideração a proporção de dados com e sem a variável LAW, pode-se verificar na Figura 13 abaixo a diferença entre o número de mortes dos dois grupos.

Figura 13 - Boxplot antes e depois da lei.



Fonte: Autor (2021).

Comprova-se que após a lei entrar em vigor o número de mortes causadas por acidentes de trânsito diminuiu. Antes da lei entrar em vigor a média anual era de aproximadamente 120 e após a lei a média caiu para menos de 100. Os valores de máximos e mínimos também tiveram uma queda após a lei entrar em vigor. Outro fator que confirma que a variável “law” tem muita influência no resultado da previsão e não deveria ser ignorada para ter uma previsão mais assertiva.

6 CONCLUSÃO

O analista não sabe de antemão qual será o melhor modelo, e faz parte do processo do analista testar várias opções. Adicionalmente, cada modelo poderá explorar as *features* de modo diferente, e disponibilizar novas *features* ao modelo torna-se necessário. Por essa razão a área de estudo chamada de "*features engineering*" tem ganhado cada vez mais importância. No caso específico de séries temporais, foi mostrado no presente trabalho como a criação de novas *features* permite que modelos antes restritos a aplicações genéricas de ML podem ser empregados em séries temporais. Contudo, o analista terá sempre que explorar, em um processo iterativo e incremental, como que novas *features* contribuem para cada modelo, e finalmente decidir qual é o melhor conjunto entre modelo e *features* criadas.

Com um mesmo algoritimo de ML existem outras formas de realizar a busca pelos melhores parâmetros para configurar esse algoritimo, porém como o foco do presente trabalho foi de apresentar um guia detalhado com as etapas necessárias para viabilizar a utilização de tais modelos em séries temporais, não foi possível dedicar mais tempo para esta atividade. Também vale ressaltar que mesmo com poucas alterações para conseguir melhores resultados, os modelos de ML obtiveram resultados muito superiores aos modelos tradicionais apresentados.

A criação de novas *features* e a preparação dos dados como apresentado neste trabalho pode munir o analista de um conjunto maior de possíveis modelos a serem usados na previsão de séries temporais.

6.1 ALCANCE DOS OBJETIVOS

O presente trabalho ocorreu conforme as etapas descritas no Capítulo 3. Dessa forma, todos os objetivos definidos no Capítulo 1 foram alcançados.

O primeiro objetivo específico que foi escolher o conjunto de dados para ser utilizado no estudo de caso foi concluído com sucesso no item 3.1.

A apresentação dos modelos tradicionais de previsão de séries temporais foi realizada no item 2.3, concluindo assim o segundo objetivo específico.

O terceiro objetivo específico que era descrever os passos de como os dados deveriam ser preparados detalhadamente para serem submetidos a um modelo genérico de *machine learning* foi concluído no item 4.3, onde foram expostas as etapas do guia.

A seleção dos modelos de ML pode ser verificada no Capítulo 4 onde foram escolhidos o modelo de regressão múltipla e o XGBoost, concluindo assim o quarto objetivo específico.

Já o quinto objetivo específico foi comtemplado no Capítulo 5 como um todo, onde foram analisados e comparados os resultados dos diferentes modelos de previsão.

Por fim, foram destacadas as vantagens de cada abordagem apresentada, finalizando e concluindo assim todos os objetivos propostos para este presente trabalho.

6.2 SUGESTÕES DE MELHORIAS PARA TRABALHOS FUTUROS

Neste trabalho foi explorado apenas dois modelos de complexidade bem diferente com o objetivo de enfatizar os pontos apresentados anteriormente neste capitulo, além de destacar as vantagens de poder contar com os modelos de ML na previsão de Séries Temporais.

Para trabalhos futuros, sugere-se explorar diferentes algoritmos de ML, dedicando um pouco mais de tempo na configuração dos seus parâmetros, e no ajuste das *features*. Nem todas as *features* criadas inicialmente precisam estar no modelo final. Porém, a ideia deste presente trabalho foi mostrar ao leitor como preparar os dados de séries temporais para que fosse possível usar modelos consagrados de ML e obter bons resultados.

REFERÊNCIAS

ABEPRO. ABEPRO - **Associação Brasileira de Engenharia de Produção** | A Profissão. Disponível em: < http://portal.abepro.org.br/a-profissao/>. Acesso em: 20 de outubro de 2021.

ALPAYDIN, E. **Introduction to Machine Learning.** Massachusetts, USA: MIT Press, 2004.

AL-SABA, T., EL-AMIN, I. **Artificial neural networks as applied to long-term demand forecasting**. Artificial Intelligence in Engineering. [S.l], v. 13, n. 2, p. 189-197, abril 1999.

ARMSTRONG, J. **Principles of Forecasting**: a Handbook for Researchers and Practitioners. Boston: Kluwer Academic Publishers, 2001.

BAKAR, N. A.; ROSBI, S. **Data Clustering Using Autoregressive Integrated Moving Average (ARIMA) Model for Islamic Country Currency**: An Econometrics Method for Islamic Financial Engineering. The International Journal of Engineering and Science, v. 06, n. 06, p. 22-31, 2017.

BITTENCOURT, G. **Inteligência artificial**: Ferramentas e teorias. 3. ed. Florianópolis: UFSC, 2006.

BOX, P.; JENKINS, G. M. **Time series analysis**: forecasting and control. 1. ed. San Francisco: Holden-day Inc, 1976.

BREIMAN L, FRIEDMAN JH, STONE CJ, OLSHEN RA, **Classification and regression trees**. Chapman and Hall, New York, 1993.

C. CHATFIELD, "**Model uncertainty and forecast accuracy**", J. Forecasting 15 (1996).

CARVALHO, A. C. P. DE L. F. DE. **Inteligência Artificial**: riscos, benefícios e uso responsável. Estudos Avançados, v. 35, p. 21–36, 19 abr. 2021.

CARVALHO, H. **Análise multivariada de dados qualitativos**: utilização do SPSS. Lisboa: Edições Sílabo, 2004.

CECATTO, C.; BELFIORE, P. **O uso de métodos de previsão de demanda nas indústrias alimentícias brasileiras.** Gestão & Produção, v. 22, n. 2, p. 404–418, jun. 2015.

CHEN, T.; GUESTRIN, C. XGBoost: **A Scalable Tree Boosting System**. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 785-794, 2016.

CLIFF, T.; RAGSDALE. **Spreadsheet Modeling & Decision Analysis**: A Practical Introduction to Management Science, 6 ed. 2012.

DAWOOD, E. G. **Geo-locating UEs Using Multi-output Decision Tree Regressor**. Florida Institute of Technology, 2019.

DOBRA, A. **Classification and Regression Tree Construction**. Cornell University, 2002.

Ertel, W. (2017). **Introduction to Artificial Intelligence**. 2nd ed. London: Springer.

F. GIROSI, M. JONES, AND T. POGGIO, "**Priors, stabilizers and basics functions**: From regularization to radial, tensor and additive splines." AI Memo No: 1430, MIT AI Lab, 1993.

Facelli, K., Lorena, A. C., Gama, J., & de Carvalho, A. C, P. L. F. **Inteligência Artificial**: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, (2019).

G. ZHANG, B.E. PATUWO, M.Y. HU, "**Forecasting with artificial neural networks**: The state of the art", International Journal of Forecasting 14 (1998).

GHAHRAMANI, Z. **Probabilistic machine learning and artificial intelligence**. Nature, v. 521, n. 7553, p. 452–459, maio 2015.

GUO, Y. et al. **Machine learning-based thermal response time ahead energy demand prediction for building heating systems**. Applied Energy, v. 221, p. 16–27, 1 jul. 2018.

HAIR jr., J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise Multivariada de dados.** 5. Ed. Porto Alegre: Bookman, 2009.

Harvey, A. C. and Durbin, J. **The effects of seat belt legislation on British road casualties:** A case study in structural time series modelling. Journal of the Royal Statistical Society Series A, 149, 187–227. doi: 10.2307/2981553, (1986).

HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed. New Jersey: Prentice Hall, 2009.

HYNDMAN, R. J.; MAKRIDAKIS, S.; WHEELWRIGHT, S. C. **Forecasting: Methods and Applications.** 3ª. ed. New York: John Wiley& Sons, 1998.

HOCHREITER S, SCHMIDHUBER J. **Long Short-Term Memory**. Neural Computation, 1997.

JEAN, N. et al. **Combining satellite imagery and machine learning to predict poverty.** Science, v. 353, n. 6301, p. 790–794, 19 ago. 2016.

JOHNSON, R. A.; Wichern, D. W. **Applied multivariate statistical analysis.** New Jersey: Prentice – Hall, 1999.

L. BREIMAN, J. H. FRIEDMAN, R. A. OLSHEN, AND C. J. STONE, **Classification and Regression Trees.** Chapman and Hall, 1984.

LUDERMIR, T. B. **Inteligência Artificial e Aprendizado de Máquina**: estado atual e tendências. Estudos Avançados, v. 35, n. 101, p. 85–94, abr. 2021.

MAHESH, B. **Machine Learning Algorithms** - A Review. v. 9, n. 1, p. 7, 2018.

MAKRIDAKIS, S.; WHEELWRIGHT, S.; HYNDMAN, R. **Forecasting:** Methods and Applications. 3. ed., New York: John Wiley & Sons, 1998.

MANLY, B. F. J. **Multivariate statistical methods.** A primer. 2. Ed. London: Chapman & Hall, 1997.

MAURYA, P. K. Crop **Value Forecasting using Decision Tree Regressor and Models.** European Journal of Molecular & Clinical Medicine, v. 7, 2020.

MONTGOMERY, D.; JOHNSON, L.; GARDINER, J. **Forecasting and Time Series Analysis.** New York: McGraw-Hill, 1990.

MOON, M; MENTZER, J.; SMITH, C.; GARVER, M. **Seven Keys to Better Forecasting.** Business Horizons. v. 41, n. 5, p. 44-52, 1998.

MORETTIN, Pedro Alberto. TOLOI, Clécia Maria de Castro. **Modelos para previsão de Séries Temporais.** 2. ed. Rio de Janeiro: Edgard Blucher (2006).

MURDICK, R. G.; GEORGOFF, D. M. **Forecasting:** a Systems Approach. Technological Forecasting and Social Change. v. 44, n. 1, p. 1-16, 1993.

P.J. BROCKWELL AND R.A. DAVIS. **Introduction to Time Series and Forecasting.** Springer (2002).

P.J. BROCKWELL AND R.A. DAVIS. **Time Series:** Theory and methods, 1991.

R.H. SHUMWAY AND D.S. STOFFER. **Time Series Analysis and Its Applications.** With R Examples. 2nd edition (2006).

SATRIO, C. B. A. et al. **Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET.** 5th International Conference on Computer Science and Computational Intelligence, 2020.

SEBER, G.A.F., Lee, A.L.: **Linear Regression Analysis**, vol. 329. Wiley, Hoboken (2012) series forecasting", Information Sciences 178 (2008).

SOUZA, R. C.; CAMARGO, M. E. **Análise e previsão de séries temporais:** os modelos ARIMA. 2 ed. Rio de Janeiro: Regional, 2004.

T. ESCOVEDO, A. KOSHIYAMA, **Introdução a Data Science**: Algoritmos de Machine Learning e métodos de análise. São Paulo, Ed. Casa do Código (2020).

TUBINO, Dálvio Ferrari. (2009) – **Manual de Planejamento e Controle da Produção**. 2a ed. São Paulo: Atlas

WEIGEND, A. GERSHENFELD, N. **Time Series Prediction Forecasting the Future and Understanding the Past.** Addison, Wesley, USA, 1994.

APÊNDICE A – BASE DE DADOS UTILIZADA

Date	DriversKilled	kms	PetrolPrice	law
1969-01-01	107	9059	0,103	0
1969-02-01	97	7685	0,102	0
1969-03-01	102	9963	0,102	0
1969-04-01	87	10955	0,101	0
1969-05-01	119	11823	0,101	0
1969-06-01	106	12391	0,101	0
1969-07-01	110	13460	0,104	0
1969-08-01	106	14055	0,104	0
1969-09-01	107	12106	0,104	0
1969-10-01	134	11372	0,103	0
1969-11-01	147	9834	0,103	0
1969-12-01	180	9267	0,102	0
1970-01-01	125	9130	0,101	0
1970-02-01	134	8933	0,101	0
1970-03-01	110	11000	0,100	0
1970-04-01	102	10733	0,099	0
1970-05-01	103	12912	0,098	0
1970-06-01	111	12926	0,098	0
1970-07-01	120	13990	0,097	0
1970-08-01	129	14926	0,097	0
1970-09-01	122	12900	0,097	0
1970-10-01	183	12034	0,096	0
1970-11-01	169	10643	0,096	0
1970-12-01	190	10742	0,095	0
1971-01-01	134	10266	0,097	0
1971-02-01	108	10281	0,096	0
1971-03-01	104	11527	0,095	0
1971-04-01	117	12281	0,095	0
1971-05-01	157	13587	0,094	0
1971-06-01	148	13049	0,094	0
1971-07-01	130	16055	0,093	0
1971-08-01	140	15220	0,093	0
1971-09-01	136	13824	0,093	0
1971-10-01	140	12729	0,092	0
1971-11-01	187	11467	0,092	0
1971-12-01	150	11351	0,091	0
1972-01-01	159	10803	0,091	0
1972-02-01	143	10548	0,090	0
1972-03-01	114	12368	0,090	0
1972-04-01	127	13311	0,089	0
1972-05-01	159	13885	0,089	0

1972-06-01	156	14088	0,088	0
1972-07-01	138	16932	0,089	0
1972-08-01	120	16164	0,088	0
1972-09-01	117	14883	0,089	0
1972-10-01	170	13532	0,088	0
1972-11-01	168	12220	0,087	0
1972-12-01	198	12025	0,087	0
1973-01-01	144	11692	0,086	0
1973-02-01	146	11081	0,086	0
1973-03-01	109	13745	0,085	0
1973-04-01	131	14382	0,084	0
1973-05-01	151	14391	0,085	0
1973-06-01	140	15597	0,084	0
1973-07-01	153	16834	0,084	0
1973-08-01	140	17282	0,084	0
1973-09-01	161	15779	0,083	0
1973-10-01	168	13946	0,081	0
1973-11-01	152	12701	0,083	0
1973-12-01	136	10431	0,094	0
1974-01-01	113	11616	0,092	0
1974-02-01	100	10808	0,108	0
1974-03-01	103	12421	0,107	0
1974-04-01	103	13605	0,114	0
1974-05-01	121	14455	0,112	0
1974-06-01	134	15019	0,111	0
1974-07-01	133	15662	0,110	0
1974-08-01	129	16745	0,108	0
1974-09-01	144	14717	0,107	0
1974-10-01	154	13756	0,105	0
1974-11-01	156	12531	0,119	0
1974-12-01	163	12568	0,118	0
1975-01-01	122	11249	0,133	0
1975-02-01	92	11096	0,131	0
1975-03-01	117	12637	0,128	0
1975-04-01	95	13018	0,124	0
1975-05-01	96	15005	0,119	0
1975-06-01	108	15235	0,116	0
1975-07-01	108	15552	0,115	0
1975-08-01	106	16905	0,115	0
1975-09-01	140	14776	0,114	0
1975-10-01	114	14104	0,112	0
1975-11-01	158	12854	0,111	0
1975-12-01	161	12956	0,115	0

1976-01-01	102	12177	0,114	0
1976-02-01	127	11918	0,112	0
1976-03-01	125	13517	0,112	0
1976-04-01	101	14417	0,110	0
1976-05-01	97	15911	0,108	0
1976-06-01	112	15589	0,108	0
1976-07-01	112	16543	0,109	0
1976-08-01	113	17925	0,108	0
1976-09-01	108	15406	0,106	0
1976-10-01	128	14601	0,106	0
1976-11-01	154	13107	0,105	0
1976-12-01	162	12268	0,103	0
1977-01-01	112	11972	0,101	0
1977-02-01	79	12028	0,100	0
1977-03-01	82	14033	0,099	0
1977-04-01	127	14244	0,102	0
1977-05-01	108	15287	0,103	0
1977-06-01	110	16954	0,102	0
1977-07-01	123	17361	0,100	0
1977-08-01	103	17694	0,093	0
1977-09-01	97	16222	0,092	0
1977-10-01	140	14969	0,091	0
1977-11-01	165	13624	0,090	0
1977-12-01	183	13842	0,089	0
1978-01-01	148	12387	0,088	0
1978-02-01	111	11608	0,088	0
1978-03-01	116	15021	0,087	0
1978-04-01	115	14834	0,085	0
1978-05-01	100	16565	0,085	0
1978-06-01	106	16882	0,084	0
1978-07-01	134	18012	0,084	0
1978-08-01	125	18855	0,084	0
1978-09-01	117	17243	0,083	0
1978-10-01	122	16045	0,083	0
1978-11-01	153	14745	0,085	0
1978-12-01	178	13726	0,085	0
1979-01-01	114	11196	0,084	0
1979-02-01	94	12105	0,085	0
1979-03-01	128	14723	0,088	0
1979-04-01	119	15582	0,090	0
1979-05-01	111	16863	0,091	0
1979-06-01	110	16758	0,109	0
1979-07-01	114	17434	0,114	0

1979-08-01	118	18359	0,113	0
1979-09-01	115	17189	0,111	0
1979-10-01	132	16909	0,109	0
1979-11-01	153	15380	0,108	0
1979-12-01	171	15161	0,108	0
1980-01-01	115	14027	0,104	0
1980-02-01	95	14478	0,107	0
1980-03-01	92	16155	0,107	0
1980-04-01	100	16585	0,112	0
1980-05-01	95	18117	0,111	0
1980-06-01	114	17552	0,112	0
1980-07-01	102	18299	0,110	0
1980-08-01	104	19361	0,108	0
1980-09-01	132	17924	0,106	0
1980-10-01	136	17872	0,104	0
1980-11-01	117	16058	0,102	0
1980-12-01	137	15746	0,103	0
1981-01-01	111	15226	0,105	0
1981-02-01	106	14932	0,104	0
1981-03-01	98	16846	0,117	0
1981-04-01	84	16854	0,115	0
1981-05-01	94	18146	0,113	0
1981-06-01	105	17559	0,114	0
1981-07-01	123	18655	0,119	0
1981-08-01	109	19453	0,124	0
1981-09-01	130	17923	0,123	0
1981-10-01	153	17915	0,121	0
1981-11-01	134	16496	0,121	0
1981-12-01	99	13544	0,117	0
1982-01-01	115	13601	0,113	0
1982-02-01	104	15667	0,108	0
1982-03-01	131	17358	0,109	0
1982-04-01	108	18112	0,111	0
1982-05-01	103	18581	0,111	0
1982-06-01	115	18759	0,115	0
1982-07-01	122	20668	0,115	0
1982-08-01	122	21040	0,117	0
1982-09-01	125	18993	0,119	0
1982-10-01	137	18668	0,118	0
1982-11-01	138	16768	0,117	0
1982-12-01	152	16551	0,117	0
1983-01-01	120	16231	0,113	0
1983-02-01	95	15511	0,114	1

1983-03-01	100	18308	0,113	1
1983-04-01	89	17793	0,118	1
1983-05-01	82	19205	0,118	1
1983-06-01	89	19162	0,118	1
1983-07-01	60	20997	0,120	1
1983-08-01	84	20705	0,119	1
1983-09-01	113	18759	0,119	1
1983-10-01	126	19240	0,118	1
1983-11-01	122	17504	0,118	1
1983-12-01	118	16591	0,118	1
1984-01-01	92	16224	0,118	1
1984-02-01	86	16670	0,115	1
1984-03-01	81	18539	0,116	1
1984-04-01	84	19759	0,115	1
1984-05-01	87	19584	0,115	1
1984-06-01	90	19976	0,115	1
1984-07-01	79	21486	0,115	1
1984-08-01	96	21626	0,115	1
1984-09-01	122	20195	0,114	1
1984-10-01	120	19928	0,116	1
1984-11-01	137	18564	0,116	1
1984-12-01	154	18149	0,116	1

APÊNDICE B – APLICAÇÃO MODELO HOLT

Ano	Mês	Período	DrivesKilled	Nível Base	Tendência	Previsão
1969	1	1	107	107,00	0	--
	2	2	97	98,72	0	107,00
	3	3	102	101,44	0	98,72
	4	4	87	89,49	0	101,44
	5	5	119	113,92	0	89,49
	6	6	106	107,36	0	113,92
	7	7	110	109,55	0	107,36
	8	8	106	106,61	0	109,55
	9	9	107	106,93	0	106,61
	10	10	134	129,34	0	106,93
	11	11	147	143,96	0	129,34
	12	12	180	173,79	0	143,96
1970	1	13	125	133,40	0	173,79
	2	14	134	133,90	0	133,40
	3	15	110	114,12	0	133,90
	4	16	102	104,09	0	114,12
	5	17	103	103,19	0	104,09
	6	18	111	109,65	0	103,19
	7	19	120	118,22	0	109,65
	8	20	129	127,14	0	118,22
	9	21	122	122,89	0	127,14
	10	22	183	172,64	0	122,89
	11	23	169	169,63	0	172,64
	12	24	190	186,49	0	169,63
1971	1	25	134	143,04	0	186,49
	2	26	108	114,04	0	143,04
	3	27	104	105,73	0	114,04
	4	28	117	115,06	0	105,73
	5	29	157	149,78	0	115,06
	6	30	148	148,31	0	149,78
	7	31	130	133,15	0	148,31
	8	32	140	138,82	0	133,15
	9	33	136	136,49	0	138,82
	10	34	140	139,39	0	136,49
	11	35	187	178,80	0	139,39
	12	36	150	154,96	0	178,80
1972	1	37	159	158,30	0	154,96
	2	38	143	145,64	0	158,30
	3	39	114	119,45	0	145,64
	4	40	127	125,70	0	119,45
	5	41	159	153,26	0	125,70
	6	42	156	155,53	0	153,26
	7	43	138	141,02	0	155,53
	8	44	120	123,62	0	141,02
	9	45	117	118,14	0	123,62
	10	46	170	161,07	0	118,14

	11	47	168	166,81	0	161,07
	12	48	198	192,63	0	166,81
1973	1	49	144	152,38	0	192,63
	2	50	146	147,10	0	152,38
	3	51	109	115,56	0	147,10
	4	52	131	128,34	0	115,56
	5	53	151	147,10	0	128,34
	6	54	140	141,22	0	147,10
	7	55	153	150,97	0	141,22
	8	56	140	141,89	0	150,97
	9	57	161	157,71	0	141,89
	10	58	168	166,23	0	157,71
	11	59	152	154,45	0	166,23
	12	60	136	139,18	0	154,45
1974	1	61	113	117,51	0	139,18
	2	62	100	103,02	0	117,51
	3	63	103	103,00	0	103,02
	4	64	103	103,00	0	103,00
	5	65	121	117,90	0	103,00
	6	66	134	131,23	0	117,90
	7	67	133	132,69	0	131,23
	8	68	129	129,64	0	132,69
	9	69	144	141,53	0	129,64
	10	70	154	151,85	0	141,53
	11	71	156	155,29	0	151,85
	12	72	163	161,67	0	155,29
1975	1	73	122	128,83	0	161,67
	2	74	92	98,35	0	128,83
	3	75	117	113,79	0	98,35
	4	76	95	98,24	0	113,79
	5	77	96	96,39	0	98,24
	6	78	108	106,00	0	96,39
	7	79	108	107,66	0	106,00
	8	80	106	106,29	0	107,66
	9	81	140	134,19	0	106,29
	10	82	114	117,48	0	134,19
	11	83	158	151,02	0	117,48
	12	84	161	159,28	0	151,02
1976	1	85	102	111,87	0	159,28
	2	86	127	124,39	0	111,87
	3	87	125	124,90	0	124,39
	4	88	101	105,12	0	124,90
	5	89	97	98,40	0	105,12
	6	90	112	109,66	0	98,40
	7	91	112	111,60	0	109,66
	8	92	113	112,76	0	111,60
	9	93	108	108,82	0	112,76
	10	94	128	124,70	0	108,82
	11	95	154	148,95	0	124,70
	12	96	162	159,75	0	148,95

1977	1	97	112	120,23	0	159,75
	2	98	79	86,10	0	120,23
	3	99	82	82,71	0	86,10
	4	100	127	119,37	0	82,71
	5	101	108	109,96	0	119,37
	6	102	110	109,99	0	109,96
	7	103	123	120,76	0	109,99
	8	104	103	106,06	0	120,76
	9	105	97	98,56	0	106,06
	10	106	140	132,86	0	98,56
	11	107	165	159,46	0	132,86
	12	108	183	178,95	0	159,46
1978	1	109	148	153,33	0	178,95
	2	110	111	118,29	0	153,33
	3	111	116	116,39	0	118,29
	4	112	115	115,24	0	116,39
	5	113	100	102,63	0	115,24
	6	114	106	105,42	0	102,63
	7	115	134	129,08	0	105,42
	8	116	125	125,70	0	129,08
	9	117	117	118,50	0	125,70
	10	118	122	121,40	0	118,50
	11	119	153	147,56	0	121,40
	12	120	178	172,76	0	147,56
1979	1	121	114	124,12	0	172,76
	2	122	94	99,19	0	124,12
	3	123	128	123,04	0	99,19
	4	124	119	119,70	0	123,04
	5	125	111	112,50	0	119,70
	6	126	110	110,43	0	112,50
	7	127	114	113,39	0	110,43
	8	128	118	117,21	0	113,39
	9	129	115	115,38	0	117,21
	10	130	132	129,14	0	115,38
	11	131	153	148,89	0	129,14
	12	132	171	167,19	0	148,89
1980	1	133	115	123,99	0	167,19
	2	134	95	99,99	0	123,99
	3	135	92	93,38	0	99,99
	4	136	100	98,86	0	93,38
	5	137	95	95,66	0	98,86
	6	138	114	110,84	0	95,66
	7	139	102	103,52	0	110,84
	8	140	104	103,92	0	103,52
	9	141	132	127,16	0	103,92
	10	142	136	134,48	0	127,16
	11	143	117	120,01	0	134,48
	12	144	137	134,07	0	120,01
1981	1	145	111	114,97	0	134,07
	2	146	106	107,55	0	114,97

	3	147	98	99,64	0	107,55
	4	148	84	86,69	0	99,64
	5	149	94	92,74	0	86,69
	6	150	105	102,89	0	92,74
	7	151	123	119,54	0	102,89
	8	152	109	110,81	0	119,54
	9	153	130	126,70	0	110,81
	10	154	153	148,47	0	126,70
	11	155	134	136,49	0	148,47
	12	156	99	105,46	0	136,49
1982	1	157	115	113,36	0	105,46
	2	158	104	105,61	0	113,36
	3	159	131	126,63	0	105,61
	4	160	108	111,21	0	126,63
	5	161	103	104,41	0	111,21
	6	162	115	113,18	0	104,41
	7	163	122	120,48	0	113,18
	8	164	122	121,74	0	120,48
	9	165	125	124,44	0	121,74
	10	166	137	134,84	0	124,44
	11	167	138	137,45	0	134,84
	12	168	152	149,49	0	137,45
1983	1	169	120	125,08	0	149,49
	2	170	95	100,18	0	125,08
	3	171	100	100,03	0	100,18
	4	172	89	90,90	0	100,03
	5	173	82	83,53	0	90,90
	6	174	89	88,06	0	83,53
	7	175	60	64,83	0	88,06
	8	176	84	80,70	0	64,83
	9	177	113	107,44	0	80,70
	10	178	126	122,80	0	107,44
	11	179	122	122,14	0	122,80
	12	180	118	118,71	0	122,14
1984	1	181	92			118,71
	2	182	86			118,71
	3	183	81			118,71
	4	184	84			118,71
	5	185	87			118,71
	6	186	90			118,71
	7	187	79			118,71
	8	188	96			118,71
	9	189	122			118,71
	10	190	120			118,71
	11	191	137			118,71
	12	192	154			118,71

APÊNDICE C – APLICAÇÃO MODELO HOLT-WINTERS

Ano	Mês	Período	DriversKilled	Nível Base	Tendência	Sazonal	Previsão
1969	1	1	107	--	--	-9,833	--
	2	2	97	--	--	-19,833	--
	3	3	102	--	--	-14,833	--
	4	4	87	--	--	-29,833	--
	5	5	119	--	--	2,167	--
	6	6	106	--	--	-10,833	--
	7	7	110	--	--	-6,833	--
	8	8	106	--	--	-10,833	--
	9	9	107	--	--	-9,833	--
	10	10	134	--	--	17,167	--
	11	11	147	--	--	30,167	--
	12	12	180	116,8	0,0	63,167	--
1970	1	13	125	117,58	0,50	1,59	107,00
	2	14	134	119,56	1,49	2,86	98,25
	3	15	110	121,21	1,60	-12,43	106,22
	4	16	102	123,18	1,85	-24,10	92,97
	5	17	103	124,02	1,17	-13,19	127,19
	6	18	111	125,06	1,08	-12,97	114,36
	7	19	120	126,17	1,10	-6,39	119,31
	8	20	129	127,79	1,45	-2,86	116,43
	9	21	122	129,35	1,52	-8,19	119,41
	10	22	183	132,32	2,49	39,36	148,03
	11	23	169	134,98	2,60	32,72	164,98
	12	24	190	137,13	2,30	56,35	200,75
1971	1	25	134	139,14	2,11	-2,87	141,03
	2	26	108	139,75	1,11	-20,06	144,12
	3	27	104	139,85	0,43	-27,94	128,43
	4	28	117	140,31	0,45	-23,58	116,17
	5	29	157	141,99	1,27	5,49	127,58
	6	30	148	143,99	1,76	-1,73	130,29
	7	31	130	145,36	1,50	-12,33	139,36
	8	32	140	146,70	1,39	-5,40	144,01
	9	33	136	147,93	1,28	-10,66	139,90
	10	34	140	147,19	-0,07	8,53	188,57
	11	35	187	147,42	0,13	37,26	179,85
	12	36	150	145,32	-1,36	22,13	203,90
1972	1	37	159	144,70	-0,86	8,50	141,09
	2	38	143	144,63	-0,33	-7,86	123,77
	3	39	114	144,20	-0,40	-29,44	116,36
	4	40	127	144,09	-0,21	-19,28	120,23
	5	41	159	144,28	0,06	11,60	149,37
	6	42	156	144,89	0,43	6,77	142,61
	7	43	138	145,53	0,57	-9,15	132,99
	8	44	120	145,24	-0,01	-18,54	140,70
	9	45	117	144,50	-0,49	-21,82	134,57
	10	46	170	144,74	-0,01	19,61	152,54
	11	47	168	144,15	-0,40	28,38	181,99

	12	48	198	145,08	0,49	42,52	165,88
1973	1	49	144	145,16	0,21	2,10	154,08
	2	50	146	145,72	0,45	-2,47	137,51
	3	51	109	145,85	0,24	-34,35	116,73
	4	52	131	146,26	0,35	-16,62	126,81
	5	53	151	146,31	0,15	7,02	158,22
	6	54	140	145,92	-0,22	-1,63	153,24
	7	55	153	146,38	0,24	1,29	136,55
	8	56	140	147,12	0,57	-10,98	128,08
	9	57	161	149,15	1,54	0,48	125,87
	10	58	168	150,60	1,48	18,15	170,31
	11	59	152	150,90	0,69	10,32	180,46
	12	60	136	149,17	-0,92	5,63	194,11
1974	1	61	113	146,70	-1,96	-21,61	150,36
	2	62	100	142,99	-3,13	-29,31	142,28
	3	63	103	139,76	-3,20	-35,95	105,52
	4	64	103	135,86	-3,67	-27,38	119,95
	5	65	121	131,44	-4,17	-4,54	139,22
	6	66	134	127,61	-3,94	3,68	125,64
	7	67	133	124,01	-3,72	6,39	124,96
	8	68	129	121,11	-3,17	1,52	109,31
	9	69	144	119,00	-2,46	16,72	118,42
	10	70	154	117,34	-1,93	30,41	134,69
	11	71	156	116,67	-1,09	29,53	125,73
	12	72	163	117,32	0,07	32,16	121,22
1975	1	73	122	118,48	0,80	-4,96	95,78
	2	74	92	119,36	0,85	-28,02	89,97
	3	75	117	121,57	1,76	-15,17	84,27
	4	76	95	123,29	1,74	-27,98	95,96
	5	77	96	124,01	1,06	-20,09	120,49
	6	78	108	124,21	0,48	-9,49	128,75
	7	79	108	123,73	-0,16	-8,26	131,08
	8	80	106	122,78	-0,69	-10,60	125,09
	9	81	140	122,14	-0,66	17,47	138,81
	10	82	114	119,91	-1,71	6,35	151,90
	11	83	158	118,63	-1,42	36,05	147,74
	12	84	161	117,69	-1,10	39,54	149,37
1976	1	85	102	116,20	-1,37	-11,08	111,63
	2	86	127	116,50	-0,25	-2,51	86,81
	3	87	125	117,24	0,41	0,02	101,08
	4	88	101	118,12	0,73	-20,79	89,67
	5	89	97	118,78	0,68	-21,20	98,76
	6	90	112	119,54	0,73	-8,20	109,96
	7	91	112	120,27	0,73	-8,27	112,01
	8	92	113	121,11	0,81	-8,95	110,40
	9	93	108	120,61	-0,07	-2,45	139,39
	10	94	128	120,59	-0,03	7,05	126,90
	11	95	154	120,45	-0,11	34,39	156,60
	12	96	162	120,43	-0,05	40,88	159,88
1977	1	97	112	120,49	0,03	-9,37	109,30

	2	98	79	118,90	-1,06	-27,27	118,01
	3	99	82	116,36	-2,05	-22,75	117,86
	4	100	127	115,70	-1,12	0,46	93,52
	5	101	108	115,18	-0,72	-11,92	93,37
	6	102	110	114,62	-0,61	-5,83	106,27
	7	103	123	114,73	-0,13	2,69	105,75
	8	104	103	114,48	-0,21	-10,63	105,64
	9	105	97	113,66	-0,62	-11,86	111,83
	10	106	140	113,87	-0,07	19,69	120,10
	11	107	165	114,50	0,40	45,06	148,20
	12	108	183	116,03	1,15	58,16	155,78
1978	1	109	148	118,85	2,27	16,14	107,82
	2	110	111	121,83	2,74	-16,39	93,85
	3	111	116	125,17	3,14	-13,75	101,83
	4	112	115	127,73	2,75	-8,28	128,77
	5	113	100	129,72	2,24	-23,71	118,57
	6	114	106	131,12	1,68	-18,61	126,12
	7	115	134	132,74	1,64	1,74	135,49
	8	116	125	134,43	1,68	-9,84	123,75
	9	117	117	135,81	1,47	-16,46	124,25
	10	118	122	135,83	0,51	-2,51	156,97
	11	119	153	135,16	-0,28	27,04	181,39
	12	120	178	134,25	-0,70	48,62	193,03
1979	1	121	114	132,07	-1,69	-6,52	149,69
	2	122	94	129,55	-2,24	-29,08	113,99
	3	123	128	127,91	-1,84	-4,58	113,55
	4	124	119	126,12	-1,81	-7,51	117,79
	5	125	111	124,74	-1,52	-17,10	100,60
	6	126	110	123,44	-1,37	-15,19	104,61
	7	127	114	121,66	-1,64	-4,49	123,81
	8	128	118	120,34	-1,43	-4,88	110,18
	9	129	115	119,44	-1,08	-8,50	102,46
	10	130	132	119,03	-0,63	7,74	115,85
	11	131	153	118,71	-0,42	31,84	145,44
	12	132	171	118,46	-0,31	51,21	166,91
1980	1	133	115	118,29	-0,21	-4,38	111,64
	2	134	95	118,33	-0,05	-25,27	89,00
	3	135	92	117,38	-0,65	-18,36	113,70
	4	136	100	116,35	-0,91	-13,36	109,22
	5	137	95	115,30	-1,00	-19,22	98,34
	6	138	114	114,92	-0,59	-5,74	99,12
	7	139	102	114,01	-0,80	-9,47	109,85
	8	140	104	113,03	-0,92	-7,63	108,33
	9	141	132	113,28	-0,14	9,52	103,61
	10	142	136	113,78	0,28	17,33	120,89
	11	143	117	112,86	-0,52	13,49	145,90
	12	144	137	111,24	-1,25	34,36	163,55
1981	1	145	111	110,21	-1,10	-0,96	105,60
	2	146	106	110,02	-0,49	-11,20	83,83
	3	147	98	109,82	-0,30	-14,03	91,18

	4	148	84	109,01	-0,64	-21,08	96,15
	5	149	94	108,57	-0,50	-16,14	89,15
	6	150	105	108,18	-0,43	-4,05	102,33
	7	151	123	108,78	0,26	6,22	98,28
	8	152	109	109,35	0,47	-2,81	101,41
	9	153	130	110,26	0,76	16,29	119,34
	10	154	153	112,04	1,45	32,98	128,35
	11	155	134	113,78	1,64	17,95	126,98
	12	156	99	113,31	0,23	2,13	149,78
1982	1	157	115	113,64	0,30	0,58	112,59
	2	158	104	113,99	0,33	-10,40	102,74
	3	159	131	115,60	1,18	5,46	100,30
	4	160	108	117,30	1,53	-13,28	95,71
	5	161	103	118,84	1,53	-15,94	102,68
	6	162	115	120,32	1,50	-4,89	116,33
	7	163	122	121,56	1,33	2,39	128,03
	8	164	122	122,97	1,38	-1,59	120,09
	9	165	125	123,71	0,95	6,36	140,65
	10	166	137	123,80	0,38	19,88	157,63
	11	167	138	124,01	0,26	15,33	142,12
	12	168	152	125,33	0,97	18,38	126,39
1983	1	169	120	126,02	0,78	-3,79	126,88
	2	170	95	125,91	0,19	-23,99	116,40
	3	171	100	124,79	-0,69	-14,57	131,56
	4	172	89	123,20	-1,29	-27,13	110,83
	5	173	82	120,91	-1,96	-31,15	105,97
	6	174	89	117,92	-2,65	-20,80	114,07
	7	175	60	112,87	-4,25	-34,21	117,65
	8	176	84	107,67	-4,89	-16,21	107,03
	9	177	113	102,94	-4,78	8,81	109,14
	10	178	126	98,49	-4,56	24,93	118,04
	11	179	122	94,46	-4,21	23,42	109,26
	12	180	118	90,64	-3,95	24,33	108,63
1984	1	181	92	--	--	--	82,90
	2	182	86	--	--	--	58,76
	3	183	81	--	--	--	64,22
	4	184	84	--	--	--	47,72
	5	185	87	--	--	--	39,75
	6	186	90	--	--	--	46,16
	7	187	79	--	--	--	28,80
	8	188	96	--	--	--	42,85
	9	189	122	--	--	--	63,93
	10	190	120	--	--	--	76,11
	11	191	137	--	--	--	70,64
	12	192	154	--	--	--	67,61

APÊNDICE D – DECOMPOSIÇÃO DA DATA COMPLETA

Date	DriversKilled	kms	PetrolPrice	law	Ano	Mês	Trimestre
1969-01-01	107	9059	0,103	0	69	1	1
1969-02-01	97	7685	0,102	0	69	2	1
1969-03-01	102	9963	0,102	0	69	3	1
1969-04-01	87	10955	0,101	0	69	4	2
1969-05-01	119	11823	0,101	0	69	5	2
1969-06-01	106	12391	0,101	0	69	6	2
1969-07-01	110	13460	0,104	0	69	7	3
1969-08-01	106	14055	0,104	0	69	8	3
1969-09-01	107	12106	0,104	0	69	9	3
1969-10-01	134	11372	0,103	0	69	10	4
1969-11-01	147	9834	0,103	0	69	11	4
1969-12-01	180	9267	0,102	0	69	12	4
1970-01-01	125	9130	0,101	0	70	1	1
1970-02-01	134	8933	0,101	0	70	2	1
1970-03-01	110	11000	0,100	0	70	3	1
1970-04-01	102	10733	0,099	0	70	4	2
1970-05-01	103	12912	0,098	0	70	5	2
1970-06-01	111	12926	0,098	0	70	6	2
1970-07-01	120	13990	0,097	0	70	7	3
1970-08-01	129	14926	0,097	0	70	8	3
1970-09-01	122	12900	0,097	0	70	9	3
1970-10-01	183	12034	0,096	0	70	10	4
1970-11-01	169	10643	0,096	0	70	11	4
1970-12-01	190	10742	0,095	0	70	12	4
1971-01-01	134	10266	0,097	0	71	1	1
1971-02-01	108	10281	0,096	0	71	2	1
1971-03-01	104	11527	0,095	0	71	3	1
1971-04-01	117	12281	0,095	0	71	4	2
1971-05-01	157	13587	0,094	0	71	5	2
1971-06-01	148	13049	0,094	0	71	6	2
1971-07-01	130	16055	0,093	0	71	7	3
1971-08-01	140	15220	0,093	0	71	8	3
1971-09-01	136	13824	0,093	0	71	9	3
1971-10-01	140	12729	0,092	0	71	10	4
1971-11-01	187	11467	0,092	0	71	11	4
1971-12-01	150	11351	0,091	0	71	12	4
1972-01-01	159	10803	0,091	0	72	1	1
1972-02-01	143	10548	0,090	0	72	2	1
1972-03-01	114	12368	0,090	0	72	3	1
1972-04-01	127	13311	0,089	0	72	4	2
1972-05-01	159	13885	0,089	0	72	5	2

1972-06-01	156	14088	0,088	0	72	6	2
1972-07-01	138	16932	0,089	0	72	7	3
1972-08-01	120	16164	0,088	0	72	8	3
1972-09-01	117	14883	0,089	0	72	9	3
1972-10-01	170	13532	0,088	0	72	10	4
1972-11-01	168	12220	0,087	0	72	11	4
1972-12-01	198	12025	0,087	0	72	12	4
1973-01-01	144	11692	0,086	0	73	1	1
1973-02-01	146	11081	0,086	0	73	2	1
1973-03-01	109	13745	0,085	0	73	3	1
1973-04-01	131	14382	0,084	0	73	4	2
1973-05-01	151	14391	0,085	0	73	5	2
1973-06-01	140	15597	0,084	0	73	6	2
1973-07-01	153	16834	0,084	0	73	7	3
1973-08-01	140	17282	0,084	0	73	8	3
1973-09-01	161	15779	0,083	0	73	9	3
1973-10-01	168	13946	0,081	0	73	10	4
1973-11-01	152	12701	0,083	0	73	11	4
1973-12-01	136	10431	0,094	0	73	12	4
1974-01-01	113	11616	0,092	0	74	1	1
1974-02-01	100	10808	0,108	0	74	2	1
1974-03-01	103	12421	0,107	0	74	3	1
1974-04-01	103	13605	0,114	0	74	4	2
1974-05-01	121	14455	0,112	0	74	5	2
1974-06-01	134	15019	0,111	0	74	6	2
1974-07-01	133	15662	0,110	0	74	7	3
1974-08-01	129	16745	0,108	0	74	8	3
1974-09-01	144	14717	0,107	0	74	9	3
1974-10-01	154	13756	0,105	0	74	10	4
1974-11-01	156	12531	0,119	0	74	11	4
1974-12-01	163	12568	0,118	0	74	12	4
1975-01-01	122	11249	0,133	0	75	1	1
1975-02-01	92	11096	0,131	0	75	2	1
1975-03-01	117	12637	0,128	0	75	3	1
1975-04-01	95	13018	0,124	0	75	4	2
1975-05-01	96	15005	0,119	0	75	5	2
1975-06-01	108	15235	0,116	0	75	6	2
1975-07-01	108	15552	0,115	0	75	7	3
1975-08-01	106	16905	0,115	0	75	8	3
1975-09-01	140	14776	0,114	0	75	9	3
1975-10-01	114	14104	0,112	0	75	10	4
1975-11-01	158	12854	0,111	0	75	11	4
1975-12-01	161	12956	0,115	0	75	12	4

1976-01-01	102	12177	0,114	0	76	1	1
1976-02-01	127	11918	0,112	0	76	2	1
1976-03-01	125	13517	0,112	0	76	3	1
1976-04-01	101	14417	0,110	0	76	4	2
1976-05-01	97	15911	0,108	0	76	5	2
1976-06-01	112	15589	0,108	0	76	6	2
1976-07-01	112	16543	0,109	0	76	7	3
1976-08-01	113	17925	0,108	0	76	8	3
1976-09-01	108	15406	0,106	0	76	9	3
1976-10-01	128	14601	0,106	0	76	10	4
1976-11-01	154	13107	0,105	0	76	11	4
1976-12-01	162	12268	0,103	0	76	12	4
1977-01-01	112	11972	0,101	0	77	1	1
1977-02-01	79	12028	0,100	0	77	2	1
1977-03-01	82	14033	0,099	0	77	3	1
1977-04-01	127	14244	0,102	0	77	4	2
1977-05-01	108	15287	0,103	0	77	5	2
1977-06-01	110	16954	0,102	0	77	6	2
1977-07-01	123	17361	0,100	0	77	7	3
1977-08-01	103	17694	0,093	0	77	8	3
1977-09-01	97	16222	0,092	0	77	9	3
1977-10-01	140	14969	0,091	0	77	10	4
1977-11-01	165	13624	0,090	0	77	11	4
1977-12-01	183	13842	0,089	0	77	12	4
1978-01-01	148	12387	0,088	0	78	1	1
1978-02-01	111	11608	0,088	0	78	2	1
1978-03-01	116	15021	0,087	0	78	3	1
1978-04-01	115	14834	0,085	0	78	4	2
1978-05-01	100	16565	0,085	0	78	5	2
1978-06-01	106	16882	0,084	0	78	6	2
1978-07-01	134	18012	0,084	0	78	7	3
1978-08-01	125	18855	0,084	0	78	8	3
1978-09-01	117	17243	0,083	0	78	9	3
1978-10-01	122	16045	0,083	0	78	10	4
1978-11-01	153	14745	0,085	0	78	11	4
1978-12-01	178	13726	0,085	0	78	12	4
1979-01-01	114	11196	0,084	0	79	1	1
1979-02-01	94	12105	0,085	0	79	2	1
1979-03-01	128	14723	0,088	0	79	3	1
1979-04-01	119	15582	0,090	0	79	4	2
1979-05-01	111	16863	0,091	0	79	5	2
1979-06-01	110	16758	0,109	0	79	6	2
1979-07-01	114	17434	0,114	0	79	7	3

1979-08-01	118	18359	0,113	0	79	8	3
1979-09-01	115	17189	0,111	0	79	9	3
1979-10-01	132	16909	0,109	0	79	10	4
1979-11-01	153	15380	0,108	0	79	11	4
1979-12-01	171	15161	0,108	0	79	12	4
1980-01-01	115	14027	0,104	0	80	1	1
1980-02-01	95	14478	0,107	0	80	2	1
1980-03-01	92	16155	0,107	0	80	3	1
1980-04-01	100	16585	0,112	0	80	4	2
1980-05-01	95	18117	0,111	0	80	5	2
1980-06-01	114	17552	0,112	0	80	6	2
1980-07-01	102	18299	0,110	0	80	7	3
1980-08-01	104	19361	0,108	0	80	8	3
1980-09-01	132	17924	0,106	0	80	9	3
1980-10-01	136	17872	0,104	0	80	10	4
1980-11-01	117	16058	0,102	0	80	11	4
1980-12-01	137	15746	0,103	0	80	12	4
1981-01-01	111	15226	0,105	0	81	1	1
1981-02-01	106	14932	0,104	0	81	2	1
1981-03-01	98	16846	0,117	0	81	3	1
1981-04-01	84	16854	0,115	0	81	4	2
1981-05-01	94	18146	0,113	0	81	5	2
1981-06-01	105	17559	0,114	0	81	6	2
1981-07-01	123	18655	0,119	0	81	7	3
1981-08-01	109	19453	0,124	0	81	8	3
1981-09-01	130	17923	0,123	0	81	9	3
1981-10-01	153	17915	0,121	0	81	10	4
1981-11-01	134	16496	0,121	0	81	11	4
1981-12-01	99	13544	0,117	0	81	12	4
1982-01-01	115	13601	0,113	0	82	1	1
1982-02-01	104	15667	0,108	0	82	2	1
1982-03-01	131	17358	0,109	0	82	3	1
1982-04-01	108	18112	0,111	0	82	4	2
1982-05-01	103	18581	0,111	0	82	5	2
1982-06-01	115	18759	0,115	0	82	6	2
1982-07-01	122	20668	0,115	0	82	7	3
1982-08-01	122	21040	0,117	0	82	8	3
1982-09-01	125	18993	0,119	0	82	9	3
1982-10-01	137	18668	0,118	0	82	10	4
1982-11-01	138	16768	0,117	0	82	11	4
1982-12-01	152	16551	0,117	0	82	12	4
1983-01-01	120	16231	0,113	0	83	1	1
1983-02-01	95	15511	0,114	1	83	2	1

1983-03-01	100	18308	0,113	1	83	3	1
1983-04-01	89	17793	0,118	1	83	4	2
1983-05-01	82	19205	0,118	1	83	5	2
1983-06-01	89	19162	0,118	1	83	6	2
1983-07-01	60	20997	0,120	1	83	7	3
1983-08-01	84	20705	0,119	1	83	8	3
1983-09-01	113	18759	0,119	1	83	9	3
1983-10-01	126	19240	0,118	1	83	10	4
1983-11-01	122	17504	0,118	1	83	11	4
1983-12-01	118	16591	0,118	1	83	12	4
1984-01-01	92	16224	0,118	1	84	1	1
1984-02-01	86	16670	0,115	1	84	2	1
1984-03-01	81	18539	0,116	1	84	3	1
1984-04-01	84	19759	0,115	1	84	4	2
1984-05-01	87	19584	0,115	1	84	5	2
1984-06-01	90	19976	0,115	1	84	6	2
1984-07-01	79	21486	0,115	1	84	7	3
1984-08-01	96	21626	0,115	1	84	8	3
1984-09-01	122	20195	0,114	1	84	9	3
1984-10-01	120	19928	0,116	1	84	10	4
1984-11-01	137	18564	0,116	1	84	11	4
1984-12-01	154	18149	0,116	1	84	12	4

APÊNDICE E – APLICAÇÃO MODELO DE REGRESSÃO

Ano	Período	Mês	kms	PetrolPrice	Ano	Trimestre	DriversKilled	Valor Previsto
1969	1	1	9059	0,103	69	1	107	116,64
	2	2	7685	0,102	69	1	97	124,65
	3	3	9963	0,102	69	1	102	118,69
	4	4	10955	0,101	69	2	87	124,98
	5	5	11823	0,101	69	2	119	124,19
	6	6	12391	0,101	69	2	106	124,50
	7	7	13460	0,104	69	3	110	128,15
	8	8	14055	0,104	69	3	106	128,36
	9	9	12106	0,104	69	3	107	137,89
	10	10	11372	0,103	69	4	134	150,51
	11	11	9834	0,103	69	4	147	158,53
	12	12	9267	0,103	69	4	180	163,00
1970	13	1	9130	0,101	70	1	125	118,21
	14	2	8933	0,101	70	1	134	121,32
	15	3	11000	0,1	70	1	110	116,72
	16	4	10733	0,099	70	2	102	127,63
	17	5	12912	0,098	70	2	103	122,62
	18	6	12926	0,098	70	2	111	124,96
	19	7	13990	0,097	70	3	120	130,99
	20	8	14926	0,097	70	3	129	129,94
	21	9	12900	0,097	70	3	122	139,76
	22	10	12034	0,096	70	4	183	152,86
	23	11	10643	0,096	70	4	169	160,35
	24	12	10742	0,095	70	4	190	162,96
1971	25	1	10266	0,097	71	1	134	117,06
	26	2	10281	0,096	71	1	108	119,98
	27	3	11527	0,095	71	1	104	118,39
	28	4	12281	0,095	71	2	117	124,97
	29	5	13587	0,094	71	2	157	123,16
	30	6	13049	0,094	71	2	148	127,52
	31	7	16055	0,093	71	3	130	126,43
	32	8	15220	0,093	71	3	140	131,88
	33	9	13824	0,093	71	3	136	139,38
	34	10	12729	0,092	71	4	140	153,33
	35	11	11467	0,092	71	4	187	160,34
	36	12	11351	0,091	71	4	150	163,74
1972	37	1	10803	0,091	72	1	159	119,28
	38	2	10548	0,09	72	1	143	123,19
	39	3	12368	0,09	72	1	114	118,91
	40	4	13311	0,089	72	2	127	125,38
	41	5	13885	0,089	72	2	159	125,67
	42	6	14088	0,088	72	2	156	127,90
	43	7	16932	0,089	72	3	138	126,23
	44	8	16164	0,088	72	3	120	132,02
	45	9	14883	0,089	72	3	117	138,52
	46	10	13532	0,088	72	4	170	153,40
	47	11	12220	0,087	72	4	168	161,18

	48	12	12025	0,087	72	4	198	164,29
1973	49	1	11692	0,086	73	1	144	119,63
	50	2	11081	0,086	73	1	146	124,25
	51	3	13745	0,085	73	1	109	117,47
	52	4	14382	0,084	73	2	131	125,06
	53	5	14391	0,085	73	2	151	126,83
	54	6	15597	0,084	73	2	140	125,38
	55	7	16834	0,084	73	3	153	130,19
	56	8	17282	0,084	73	3	140	130,94
	57	9	15779	0,083	73	3	161	139,42
	58	10	13946	0,081	73	4	168	156,66
	59	11	12701	0,083	73	4	152	162,43
	60	12	10431	0,094	73	4	136	166,66
1974	61	1	11616	0,092	74	1	113	117,02
	62	2	10808	0,108	74	1	100	112,94
	63	3	12421	0,107	74	1	103	110,00
	64	4	13605	0,114	74	2	103	110,88
	65	5	14455	0,112	74	2	121	111,33
	66	6	15019	0,111	74	2	134	112,24
	67	7	15662	0,11	74	3	133	119,81
	68	8	16745	0,108	74	3	129	119,41
	69	9	14717	0,107	74	3	144	129,82
	70	10	13756	0,105	74	4	154	143,86
	71	11	12531	0,119	74	4	156	142,49
	72	12	12568	0,118	74	4	163	145,33
1975	73	1	11249	0,133	75	1	122	94,85
	74	2	11096	0,131	75	1	92	98,98
	75	3	12637	0,128	75	1	117	97,49
	76	4	13018	0,124	75	2	95	107,79
	77	5	15005	0,119	75	2	96	105,84
	78	6	15235	0,116	75	2	108	109,15
	79	7	15552	0,115	75	3	108	117,92
	80	8	16905	0,115	75	3	106	115,35
	81	9	14776	0,114	75	3	140	126,13
	82	10	14104	0,112	75	4	114	139,12
	83	11	12854	0,111	75	4	158	146,67
	84	12	12956	0,115	75	4	161	146,33
1976	85	1	12177	0,114	76	1	102	103,30
	86	2	11918	0,112	76	1	127	107,82
	87	3	13517	0,112	76	1	125	104,35
	88	4	14417	0,11	76	2	101	111,57
	89	5	15911	0,108	76	2	97	109,66
	90	6	15589	0,108	76	2	112	113,23
	91	7	16543	0,109	76	3	112	118,48
	92	8	17925	0,108	76	3	113	116,39
	93	9	15406	0,106	76	3	108	129,19
	94	10	14601	0,106	76	4	128	141,49
	95	11	13107	0,105	76	4	154	149,94
	96	12	12268	0,103	76	4	162	156,58
1977	97	1	11972	0,101	77	1	112	112,37

	98	2	12028	0,1	77	1	79	115,15
	99	3	14033	0,099	77	1	82	110,77
	100	4	14244	0,102	77	2	127	117,57
	101	5	15287	0,103	77	2	108	115,55
	102	6	16954	0,102	77	2	110	112,42
	103	7	17361	0,1	77	3	123	121,44
	104	8	17694	0,093	77	3	103	126,74
	105	9	16222	0,092	77	3	97	135,11
	106	10	14969	0,091	77	4	140	149,63
	107	11	13624	0,09	77	4	165	157,54
	108	12	13842	0,089	77	4	183	159,72
1978	109	1	12387	0,088	78	1	148	119,17
	110	2	11608	0,088	78	1	111	124,41
	111	3	15021	0,087	78	1	116	114,88
	112	4	14834	0,085	78	2	115	126,09
	113	5	16565	0,085	78	2	100	122,13
	114	6	16882	0,084	78	2	106	123,95
	115	7	18012	0,084	78	3	134	129,14
	116	8	18855	0,084	78	3	125	128,44
	117	9	17243	0,083	78	3	117	137,33
	118	10	16045	0,083	78	4	122	151,06
	119	11	14745	0,085	78	4	153	157,03
	120	12	13726	0,085	78	4	178	163,16
1979	121	1	11196	0,084	79	1	114	126,55
	122	2	12105	0,085	79	1	94	125,01
	123	3	14723	0,088	79	1	128	116,04
	124	4	15582	0,09	79	2	119	121,05
	125	5	16863	0,091	79	2	111	118,15
	126	6	16758	0,109	79	2	110	110,32
	127	7	17434	0,114	79	3	114	114,23
	128	8	18359	0,113	79	3	118	113,82
	129	9	17189	0,111	79	3	115	121,67
	130	10	16909	0,109	79	4	132	133,22
	131	11	15380	0,108	79	4	153	141,80
	132	12	15161	0,108	79	4	171	144,99
1980	133	1	14027	0,104	80	1	115	105,04
	134	2	14478	0,107	80	1	95	104,00
	135	3	16155	0,107	80	1	92	100,24
	136	4	16585	0,112	80	2	100	105,06
	137	5	18117	0,111	80	2	95	102,42
	138	6	17552	0,112	80	2	114	106,29
	139	7	18299	0,11	80	3	102	114,07
	140	8	19361	0,108	80	3	104	113,75
	141	9	17924	0,106	80	3	132	122,58
	142	10	17872	0,104	80	4	136	133,29
	143	11	16058	0,102	80	4	117	143,51
	144	12	15746	0,103	80	4	137	146,45
1981	145	1	15226	0,105	81	1	111	100,71
	146	2	14932	0,104	81	1	106	104,76
	147	3	16846	0,117	81	1	98	92,47

	148	4	16854	0,115	81	2	84	102,96
	149	5	18146	0,113	81	2	94	101,79
	150	6	17559	0,114	81	2	105	105,74
	151	7	18655	0,119	81	3	123	108,12
	152	8	19453	0,124	81	3	109	104,63
	153	9	17923	0,123	81	3	130	113,22
	154	10	17915	0,121	81	4	153	123,77
	155	11	16496	0,121	81	4	134	131,36
	156	12	13544	0,117	81	4	99	146,92
1982	157	1	13601	0,113	82	1	115	102,60
	158	2	15667	0,108	82	1	104	100,36
	159	3	17358	0,109	82	1	131	95,96
	160	4	18112	0,111	82	2	108	101,36
	161	5	18581	0,111	82	2	103	102,03
	162	6	18759	0,115	82	2	115	101,41
	163	7	20668	0,115	82	3	122	103,75
	164	8	21040	0,117	82	3	122	103,60
	165	9	18993	0,119	82	3	125	112,31
	166	10	18668	0,118	82	4	137	123,43
	167	11	16768	0,117	82	4	138	133,37
	168	12	16551	0,117	82	4	152	136,56
1983	169	1	16231	0,113	83	1	120	93,62
	170	2	15511	0,114	83	1	95	98,05
	171	3	18308	0,113	83	1	100	90,78
	172	4	17793	0,118	83	2	89	99,06
	173	5	19205	0,118	83	2	82	96,27
	174	6	19162	0,118	83	2	89	98,82
	175	7	20997	0,12	83	3	60	100,25
	176	8	20705	0,119	83	3	84	104,30
	177	9	18759	0,119	83	3	113	113,82
	178	10	19240	0,118	83	4	126	121,99
	179	11	17504	0,118	83	4	122	130,74
	180	12	16591	0,118	83	4	118	136,47
1984	181	1	16224	0,118	84	1	92	91,35
	182	2	16670	0,115	84	1	86	93,87
	183	3	18539	0,116	84	1	81	88,82
	184	4	19759	0,115	84	2	84	94,28
	185	5	19584	0,115	84	2	87	97,31
	186	6	19976	0,115	84	2	90	98,26
	187	7	21486	0,115	84	3	79	102,06
	188	8	21626	0,115	84	3	96	103,94
	189	9	20195	0,114	84	3	122	112,16
	190	10	19928	0,116	84	4	120	121,30
	191	11	18564	0,116	84	4	137	128,69
	192	12	18149	0,116	84	4	154	132,60