UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

William Jones Beckhauser

Churn Prediction in Enterprises with High Customer Turnover

Florianópolis
2024

William Jones Beckhauser

Churn Prediction in Enterprises with High Customer Turnover

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do título de Mestre em Ciência da Computação.
Supervisor:: Prof. Renato Fileto, Dr.

Florianópolis
2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Beckhauser, William Jones
Churn Prediction in Enterprises with High Customer Turnover / William Jones Beckhauser ; orientador, Renato Fileto, 2024.
60 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2024.

Inclui referências.

1. Ciência da Computação. 2. Churn Prediction. 3. High Customer Turnover. 4. Machine Learning. 5. Classification Model Comparison. I. Fileto, Renato. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. III. Título.

William Jones Beckhauser

Churn Prediction in Enterprises with High Customer Turnover

O presente trabalho em nível de Mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Jonata Tyska, Dr.
Universidade Federal de Santa Catarina

Prof. Mateus Grellert, Dr.
Universidade Federal do Rio Grande do Sul

Prof. Ricardo Marcondes Marcacini, Dr.
Universidade de São Paulo

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Ciência da Computação.

Documento assinado digitalmente
Marcio Bastos Castro
Data: 07/03/2024 19:31:11-0300
CPF: ***.707.720-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

Documento assinado digitalmente
Renato Fileto
Data: 07/03/2024 17:03:15-0300
CPF: ***.315.998-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Renato Fileto, Dr.
Supervisor

Florianópolis, 2024.

Este trabalho é dedicado aos meus amigos, família e à ciência.

ACKNOWLEDGEMENTS

Quero expressar meu sincero agradecimento a todos que tornaram possível a realização deste trabalho. Ao meu orientador, colegas, amigos e minha família, meu profundo reconhecimento pelo apoio, orientação e recursos fornecidos ao longo desta jornada. Nada se constrói sozinho, e este trabalho é mais um resultado do esforço conjunto de muitos.

“We can only see a short distance ahead,
but we can see plenty there that needs to be done.”
(Alan Turing)

RESUMO

A maioria das pesquisas sobre modelos de aprendizado de máquina para predizer perda de clientes se concentra em setores como telecomunicações. No entanto, esse problema pode ser particularmente desafiador em setores com alta rotatividade de clientes (HCT, do inglês High Customer Turnover), tais como entrega de alimentos, comércio eletrônico e jogos. O presente estudo visa delinear HCT e determinar as abordagens mais eficazes para prever a perda de clientes em organizações com HCT. Para tanto, foi primeiramente realizada uma análise do estado da arte. Ela permitiu selecionar três conjuntos de dados representativos de diferentes setores com HCT para experimentos. Além disso, foram identificadas as abordagens mais promissoras na literatura. Com base nisso, neste trabalho, realizamos dois experimentos. Em ambos empregamos diversos modelos de aprendizado de máquina tradicionais (SVM, Decision Tree e Random Forest) e redes neurais (Multilayer Perceptron, CNN). Entretanto, o primeiro experimento usa dados convencionais de perfil e de transações dos clientes (e.g., faixa etária e de renda, quantidade de compras, despesas financeiras e produtos adquiridos), enquanto o segundo explora textos em que os clientes avaliam suas compras mais recentes. Esses últimos experimentos também usam modelos para a geração de embeddings de texto, tais como o Word2Vec, além de modelos de linguagem como BERT e RoBERTa. Todos os modelos foram submetidos a testes de treinamento, nos quais foram exploradas diferentes combinações de hiperparâmetros. A avaliação dos modelos foi realizada com métricas como acurácia, cobertura, pre-cisão e F1-score. A metodologia aplicada é uma adaptação do processo de referência do CRISP-DM, para agilizar a seleção dos melhores modelos. Ela inclui segmentação RFM (Recência, Frequência e Valor Monetário) como parte da preparação dos da-dos, segmentando os clientes com base em suas compras, e classificando o nivel de lealdade com a empresa. Os resultados mostram que modelos de aprendizado de máquina, notadamente Random Forest e SVM, alcançam resultados superiores (F1-Score em torno de 93%) quando a segmentação RFM é empregada. Em um contexto alternativo que exclui segmentação RFM, os modelos não ultrapassam 75% de F1-Score. Nos experimentos usando somente textos de avaliações que clientes fizeram de suas compras como característica para classificação, considerando somente um subconjunto dos registros de compras com tais avaliações, o BERT alcançou F1-score de 91%. Esses resultados sugerem o potencial, ainda subexplorado, da aplicação de processamento de linguagem natural na previsão de perda de clientes.

Palavras-chave: Predição de perda de clientes. Aprendizado de máquina. Comparação de modelos de classificação. Alta rotatividade de clientes.

RESUMO EXPANDIDO

Introdução

A previsão de potenciais perdas de clientes (em inglês churn), permite a adoção de medidas preventivas para restaurar a lealdade, sendo crucial para a sobrevivência e crescimento das empresas em um ambiente de intensa concorrência corporativa. Estudos demonstram que um modesto aumento de 5% na taxa de retenção de clientes pode resultar em um incremento nos lucros de 25% a 95%. A previsão de churn constitui um aspecto crítico do Gerenciamento de Relacionamento com o Cliente (CRM). Melhorar a retenção de clientes ajuda a equilibrar a balança de consumidores, reduzindo a necessidade de adquirir novos clientes e, consequentemente, cortando custos de marketing, através da obtenção de ganhos financeiros por meio de estratégias eficazes de retenção. A demanda crescente por previsão de churn motiva a utilização de tecnologias de mineração de dados e Aprendizado de Máquina (ML, do ingles Machine Learning) para automatizar essa tarefa. Uma variedade de modelos de ML tem sido investigada recentemente, especialmente em setores como telecomunicações. No entanto, algumas empresas apresentam características que tornam a previsão de churn particularmente desafiadora e relevante: setores altamente competitivos, transações não contratuais, operações entre entidade de negócio e consumidor, além de altas taxas de churn. Essas características somadas exacerbam a dificuldade dessas empresas em reter clientes, dando origem ao conceito de Alta Rotatividade de Clientes (HCT, do inglês High Customer Turnover). A previsão automática e precisa de churn é particularmente importante para setores com HCT, como os de entrega de comida e o de jogos online, onde os clientes não têm obrigações contratuais que os vinculem. Algumas pesquisas mostram que 70% a 90% dos jogadores deixam de jogar 10 dias após a primeira partida. Apesar do interesse crescente na previsão de churn para empresas com HCT, há uma escassez de pesquisas nessa área, com os setores de telecomunicações e financeiro dominando a pesquisa em previsão de churn, enquanto o setor de jogos conta com apenas 2% dos estudos. Empresas operando sem contratos enfrentam dificuldades significativas para detectar churn de clientes em comparação com empresas contratuais, que utilizam um modelo baseado em assinatura, evidenciando a importância da previsão de churn e a escassez de pesquisas em empresas com HCT.

Objetivos

O objetivo geral deste trabalho é determinar as características (por exemplo, dados de transações, textos de avaliação de compras) e os modelos de ML e de linguagem mais eficazes para prever a perda de clientes em empresas com HCT. Os objetivos específicos desta pesquisa são: (i) analisar o estado da arte na previsão de churn em empresas com HCT; (ii) projeto e implementação de um arcabouço (framework) para suportar o treinamento eficiente e a avaliação de modelos para previsão de churn em empreendimentos com HCT; (iii) avaliação comparativa de modelos de aprendizado de máquina por meio de experimentos conduzidos em conjuntos de dados de perfil e transacionais dos clientes de empresas de três setores diversos; (iv) aplicação de Modelos de Linguagem (LMs, do inglês Language Models) para previsão de churn usando dados textuais de avaliações de compras dos clientes de uma dessas empre-

sas; (v) investigação do desempenho de vários modelos de aprendizado de máquina e LMs pré-treinados por meio de experimentos para identificar combinações ótimas de características e modelos para prever churn em indústrias diversas com HCT.

Metodologia

A metodologia deste estudo é baseada em uma abordagem empírica e quantitativa, utilizando dados de empresas reais e fontes públicas. Ela desdobrou-se em várias etapas, incluindo: (i) estudos preliminares para definir o escopo da pesquisa, formular hipóteses e objetivos, e avaliar conjuntos de dados iniciais; (ii) uma revisão sistemática da literatura sobre previsão de churn para selecionar e analisar trabalhos relevantes; (iii) análise comparativa do desempenho de modelos de previsão de churn, incluindo a seleção de modelos e conjuntos de dados para este estudo; (iv) desenvolvimento de um framework para a previsão de churn; (v) realização de experimentos para treinar, testar e avaliar os modelos selecionados; (vi) análise dos resultados experimentais usando estatísticas descritivas e métricas de desempenho e (vii) disseminação dos resultados da pesquisa para a comunidade acadêmica e praticantes, com alguns artigos resultantes ainda em fase de preparação.

Resultados e Discussão

Neste estudo, foi introduzido e aplicado o framework CP-HCT para a comparação de modelos de ML e modelos de linguagem na previsão de churn em datasets de empresas com HCT. O CP-HCT se destacou por permitir o treinamento e a seleção eficientes de modelos de ML, apoiando uma extensão do processo CRISP-DM com um teste rápido para avaliar modelos com configurações padrão de parâmetros antes das fases críticas de definição e otimização de hiperparâmetros, modelagem e avaliação de desempenho. O CP-HCT permite primeirmente determinar os modelos mais promissores para os datasets específicos. Adicionalmente, a fase de preparação de dados é otimizada pelo uso da segmentação RFM para transformar dados transacionais em informações categóricas para o treinamento. Os resultados dos experimentos realizados no âmbito deste trabalho revelaram que o modelo Random Forest (RF) obteve a melhor acurácia na maioria dos conjuntos de dados analisados, beneficiando-se significativamente do framework CP-HCT e da segmentação RFM, enquanto modelos como Máquina de Vetores de Suporte (SVM) e Multilayer Perceptron (MLP) também mostraram resultados positivos. Experimentos focados em dados de avaliações de compras, postadas por clintes de um aplicativo de entrega de alimentos, revelaram que os modelos MLP e SVM, usando embeddings do BERT, destacaram-se ao alcançar mais de 96% de F1-score. Isso demonstra o potencial do uso de dados textuais como as avaliações postadas por clientes após suas compras, mesmo quando limitados a análise de uma única característica (revisões postadas clientes, neste caso). Além disso, combinações de modelos de ML e LMs, além de outras tecnologias de processamento de linguagem natural, como os atuais grandes modelos de linguagem (LLMs) podem contribuir na melhoria do desempenho da previsão de churn.

Palavras-chave: Predição de perda de clientes. Aprendizado de máquina. Comparação de modelos de classificação. Alta rotatividade de clientes.

ABSTRACT

Most research on machine learning models to predict customer churn focuses on sectors such as telecommunications. However, this problem can be particularly challenging in industries with high customer turnover (HCT), such as food delivery, e-commerce, and gaming. This study aims to outline HCT and determine the most effective approaches for predicting customer churn in organizations with HCT. To this end, a state-of-the-art analysis was conducted. It allowed for the selection of three representative datasets from different HCT sectors for experiments. Additionally, the most promising approaches from the literature were identified. Based on this, in this work, we conducted two experiments. In both, we employed various traditional machine learning models (SVM, Decision Tree, and Random Forest) and neural networks (Multilayer Perceptron, CNN). However, the first experiment uses conventional customer profile and transaction data (e.g., age and income range, number of purchases, financial expenses, and products purchased), while the second one exploits texts in which customers review their most recent purchases. These latter experiments also use models for generating text embeddings, such as Word2Vec, as well as language models like BERT and RoBERTa. All models were subjected to training tests, where different combinations of hyperparameters were explored. The models were evaluated using metrics such as accuracy, coverage, precision, and F1-score. The applied methodology is an adaptation of the CRISP-DM reference process, to speed up the selection of the best models. It includes RFM segmentation (Recency, Frequency, and Monetary Value) as part of data preparation, segmenting customers based on their purchases, and classifying their loyalty level to the company. The results show that machine learning models, notably the Random Forest and SVM, achieve superior results (around 93% F1 Score) when RFM segmentation is employed. In an alternative context that excludes RFM segmentation, the models do not surpass the F1-score threshold of 75%. In experiments using only texts of customer purchase reviews as a classification feature, considering only a subset of purchase records with those reviews, BERT achieved an F1-score of 91%. These results suggest the still unexploited potential of applying natural language processing in predicting customer churn.

Keywords: Churn Prediction. Machine Learning. Classification Model Comparison. High Customer Turnover. High Churn Rate.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

The prediction of potential customer churn (lost) enables taking measures in advance to restore loyalty. It is essential for the survival and growth of businesses amidst intense corporate competition (GAO, L. et al., 2023). A modest 5% customer retention rate allows companies to increase their profits between 25% to 95% (GALLO, 2014). Churn prediction is a critical aspect of Customer Relationship Management (CRM). Effective churn prediction helps identifying a subset of potential consumers with a likelihood of churning, allowing preventive measures to strengthen relationships with them (TRAN; LE; NGUYEN, 2023). Improving customer retention helps balance the consumer scale, reducing the imperative for acquiring new customers and consequently cutting marketing costs. This is achieved by realizing financial gains through effective retention strategies (HOCHSTEIN et al., 2023).

The growing demand for churn prediction and prevention motivates the use of data mining and Machine Learning (ML) technologies (SUH, 2023) to automate this task. A variety of ML models for doing so have been investigated in recent years, specially in sectors like telecommunication (JAIN; KHUNTETA; SRIVASTAVA, 2021), (AHN et al., 2020b), (GEILER; AFFELDT; NADIF, 2022), (SOBREIRO et al., 2022). However, some other enterprises have four key traits that make churn prediction partic-ularly challenging and relevant: highly competitive sectors, non-contractual transactions, business-to-consumer operations, and high churn rates. Together, these traits exacer-bate the difficulty of these companies to retain customers. High churn rates usually results from the first three characteristics, proposed in this work. Thus, in this work we refer to this situation as High Customer Turnover (HCT). Accurate automatic prediction of churn is particularly important for sectors with HCT, such as online gaming, as their costumers have no contractual obligations binding them. Studies have shown that 70% to 90% of players stop playing 10 days after the first match (KIM, D.; LEE, E.; ŔHEE, ́2017), (MILOŠEV́IC; ŽIVIC; ANDJELKOVIC, 2017).

High churn rate makes it more crucial for companies to identify customers at risk of churn and take proactive measures to retain them. However, despite the growing interest in churn prediction for enterprises with HCT, there is a shortage of research in this area. Until 2020, the telecommunications and the finance sectors, which do not usually have HCT, accounted for 44% and 18% of the research in churn prediction, respectively, while the gaming sector accounted for just 2% (SOBREIRO et al., 2022).

Companies operating without contracts, such as those categorized under HCT, struggle to easily detect customer churn compared to contractual companies utilizing a subscription-based model. Contractual companies can identify when a customer is likely to churn as they have to cancel their subscription or renew it, usually prior to effectively discontinue the relationship. Meanwhile, contractless companies face

numerous obstacles in trying to detect customer churn, since customer behaviors are not fixed and cannot be predetermined, but instead, constantly vary (KANDIL, 2023). Considering the relevance of churn prediction and the dearth of research in High Customer Turnover (HCT) companies, the central inquiry guiding this study is:

> What features and models are more effective for churn prediction in industries with High Customer Turnover?

1.1 OBJECTIVES

The general objective of this work is to determine effective features (e.g. transaction data, purchase review texts), ML and language models for predicting customer churn in HCT companies.

Specific Objectives

Considering the general objective and the major challenges for developing this work, the following specific objectives stand out:

1. Analysis of the state of the art in churn prediction in enterprises with HCT, aiming to identify and select the most effective or promising ML models, features and useful datasets from distinct sectors for our experiments.
2. Design and implementation of a framework to support efficient training and evaluation of models for churn prediction;
3. Comparative evaluation of machine learning models through experiments conducted on transactional datasets spanning three diverse sectors.
4. Application of current Large Language Models (LLMs) for churn prediction using textual data from customer purchase reviews, which have been unexploited to leverage churn prediction.
5. Investigation of the performance of various machine learning models and pre-trained LLMs through experiments to identify optimal combinations of features and models for predicting churn in diverse industries.

1.2 METHODOLOGY

The research characterization for this study is empirical in nature, employing an exploratory approach with a quantitative focus. The research's aim is applied in its essence, utilizing experimental procedures and real-world company data provided by actual businesses, along with publicly available data to answer the research question.

Throughout the course of the study, the objectives have been pursued through well-defined stages. Preliminary studies played a pivotal role in framing the scope of

the work and subsequent phases. In the following, the major stages of our research are outline with concise descriptions:

1. **Preliminary Studies:** Exploratory bibliographic research, delineating the problem, formulating research questions and hypotheses, and defining objectives and methods. Initial assessment of available datasets.
2. **Systematic Review of the Literature on Churn Prediction:** Comprehensive review of literature related to churn prediction followed by work selection, critical analysis, and comparative assessment.
3. **Model Performance Comparison:** Comparative analysis of the performance (accuracy, precision, etc.) of models employed for churn prediction within HCT environments. Study of key techniques for enhancing input data quality. Selection of the models and datasets to be employed in this research.
4. **Development of the framework for Churn Prediction:** Conception and implementation of a framework for efficient data processing and model training.
5. **Conducting Churn Prediction Experiments:** Using the developed framework for efficiently training, testing and evaluating selected models with distinct datasets, features and combinations of hyperparameters.
6. **Analysis of Experimental Results:** Use of descriptive statistical techniques along with performance measures like accuracy and F1-score to compare models' performance.
7. **Dissemination of the Research Results:** Code, datasets, manuals, presentations and articles are available for the research and practitioners community, being some of them under preparation yet.

1.3 CONTRIBUTIONS

The major contributions of this work are: (i) extensive training and evaluation of ML models for churn prediction in industries with HCT; (ii) experimental results showing the superiority of the Random Forest (RF) and the SVM models in the HCT contexts considered; (iii) demonstrating the effectiveness of textual data over conventional data for predicting churn in the food delivery sector; (iv) the preparation and publication for community reuse of an annotated dataset from a real European company of the food delivery sector; (v) a framework for training, testing, selecting, evaluating, and optimizing models for churn prediction, which can be useful in situations requiring extensive experimentation.

1.4 TEXT ORGANIZATION

This document is organized as follows. Chapter 2 lays the foundations of this work by elucidating concepts like customer churn and HCT, besides discussing technologies and datasets employed in our experiments. Chapter 3 depicts two bibliographical reviews, carried out in the scope of this work, about the state of the art in the area of churn prediction, with focus on HCT enterprises. They reveal the most promising approaches found in the literature according with exploratory facets that we consider useful for their comparison. The framework developed for our experiments is described in Chapter 4. Chapter 5 describes the analysis and preparation of transactional and textual data. Chapter 6 details the plan of experiments conducted as part of this work, results and discusses. Finally, Chapter 7 sums up insights derived from our work, with a brief description of derived articles, etither already accepted or being prepared yet, and enumerates some future research directions.

2 FOUNDATIONS

This chapter provides the theoretical basis for this work, including the following topics: customer churn, High Customer Turnover (HCT), datasets for experiments and suitable ML models for churn prediction in HCT enterprises, such as Random Forest, Decision Tree, SVM and Multilayer Perceptron. It also discusses embeddings and Large Language Models such as BERT and RoBERTa, that we employ in some experiments.

2.1 CUSTOMER CHURN

Churn occurs when a customer discontinue using a product or service, becoming inactive. The term "churn" is commonly used to refer to any type of customer attrition, voluntary or involuntary (BERRY; LINOFF, 2004). Dissatisfaction with the quality of service, high costs, unattractive plans, lack of understanding of the service plan, poor support, and other factors may contribute to customer churn. Additionally, customers may terminate their contract without the intention of switching to a competitor due to changes in their situation, such as financial difficulties that prevent them from continuing to use the service or relocation to an area where the company does not operate or the service is not available (MUSTAFA; SOOK LING; ABDUL RAZAK, 2021).

Customer churn is measured by the churn rate (Z) during a specific period of time, such as a month or a year; whichever is relevant to the business. It can be calculated by the following formula, where Y refers to the number of customers who were inactive during the time period; B is the number of active customers during that period; and K is the number of new customers that arrived during that period.

$$Z = Y/(B + K)$$

2.2 HIGH CUSTOMER TURNOVER

High Customer Turnover (HCT) is a term proposed in this work to refer to the following characteristics of enterprises in which churn is more common and its prediction is often more challenging and crucial than in most business.

High Churn Rate, for this research, is a churn rate above the average of industries that we are aware of in churn publications, i.e. an annual turnover rate between 20% and 40% (HASHMI; BUTT; IQBAL, 2013). Very differently, the gaming industry, for example, can have an astounding churn rate of up to 70% (KIM, D.; LEE, E.; RHEE, 2017). Companies with high churn rates usually have the other following characteristics that contribute to this.

Non-contractual Negotiations means that HCT organizations rely on informal and flexible negotiations that do not result in a legally binding relationship between the institution and the client, as in telecom companies. While non-contractual negotiations

can be more collaborative and involve discussion, it also creates challenges in differentiating customers who have ended their relationship with the organization from those who are simply on a temporary pause between transactions (MARTÍNEZ et al., 2020). Additionally, there are no constraints on discontinuing the usage of services or products since there is no formal link with the entity.

Business to Customer (B2C) interactions, i.e. between a company and an individual customer, is another characteristic of HCT. These interactions are often facilitated through retail channels such as brick-and-mortar stores, e-commerce websites, and mobile apps (IACOBUCCI; HIBBARD, 1999). Companies that rely on B2C transactions need to be concerned with campaigns, targeting, behavioral research and retention. As a result, management differs from organizations in business to business or business to government systems.

Highly Competitive Sectors is a defining feature of HCT. In intensely competitive markets, there are many companies offering similar goods or services, creating fierce rivalry. This type of market structure creates pressure on HCT companies to increase efficiency and reduce costs to remain profitable, making it challenging for new companies to break through and exercise market creativity (MAKOWSKI; OSTROY, 2001). Thus, they have a greater need of accurate churn prediction.

2.3 DATASETS FOR EXPERIMENTS

The quality and the quantity of data used for machine learning models are crucial factors that significantly impact their efficiency (SOUZA et al., 2022). To ensure the best performance of these models, it is essential to use relevant and useful data. One way to achieve this is by using previously tested and validated data. Thus, in our experiments described in section 4. We use the well-known datasets gambling (COUSSEMENT; DE BOCK, 2013) and e-commerce (SANTOS, 2021). Other valuable dataset especially obtained for our experiments come from a company of the food delivery domain. We standardized feature identifiers of these three datasets, and made them available together for convenience [1]. They are described in the following.

2.3.1 Food Delivery

Food Delivery data obtained from a prominent European food group whose identity is kept confidential at their request. It has data about 76446 deliveries to 7249 customers, mainly for lunch and dinner on weekdays (Monday to Friday), during the 12-month period between April 20, 2022, and April 21, 2023. Table 1 lists the features available in this dataset grouped by their data types. They include the number of purchases, delivery region, type of orders (lunch or dinner), quantities of desserts, drinks,

[1] Available at https://github.com/WilliamBeckhauser/ChurnPredictionHCT

dishes, and other items, discounts offered, expenditure amounts, average spending, discounts availed, marketing communication, participation in the loyalty plan, types of discounts, and gender. In addition, about 40% of customers reviewed their last purchase, resulting in a subset of data, with the text feature Review. For this dataset, we consider 30 days without purchasing as churn, because it is the period of time commonly waited in loyalty programs to intensify promotions (WENGER, 2021), and we did not find references about churn in food delivery.

Table 1 – Food Delivery dataset summary.

# of Variables	Data Type	Labels
13	Numeric	# of Purchases; Avg. Spending; Discount type; Review score; Days from last Order
3	String	Delivery; Language; Gender; RFM
8	Boolean	Lunch Purchase; email Marketing; First Order; Churn (target)
1	Text	Reviews

2.3.2 Gambling

Gambling data generously provided by bwin Interactive Entertainment AG. This dataset is freely accessible from Cambridge Health Alliance, a teaching hospital affiliated with the Harvard Medical School. It consists of data from 2445 users spanning from February 2005 to February 2007, covering more than 99 variables, and the target “churn", grouped by data type in Table 2. It variables include customer profile details such as birthday, age, gender, language, and country. Additionally, transaction data met-rics like average and frequency of purchases and purchases by type are also included. We use 120 days to consider customer churn in this dataset as in (COUSSEMENT; DE BOCK, 2013).

Table 2 – Gambling dataset summary.

# of Variables	Data Type	Label
95	Numeric	Frequencies; Spending; Sum of Errors and Gains; Age
3	String	Country of Residence; Language; Gender; RFM
1	Boolean	Churn (target)

2.3.3 E-commerce

E-commerce data sourced from Kaggle’s public data repository. It is called “Brazilian E-Commerce Public Dataset" by Olist (KAGGLE, 2018), and encompasses 24886 purchases of 23903 customers made between the years 2016 and 2018. The dataset provides extensive information regarding customer transactions, including the number of purchases, Amounts Spent, customer ratings, purchase integrity, purchase

status, payment method, product weight, purchase category, as well as the geographical locations of both the seller and the buyer (refer to Table 3 for a summary). A period of 90 days is employed to ascertain customer churn in the e-commerce dataset as in (OLIVEIRA, 2012).

Table 3 – E-commerce dataset summary.

# of Variables	Data Type	Label
8	Numeric	Number of Purchases; Amounts Spent; Customer Ratings
10	String	Payment Method; Product Description; Address; RFM
1	Boolean	Churn (target)

Table 4 shows the distribution of churn and non-churn customers in the Food Delivery, Gambling, and E-commerce datasets. Notice that in the Food Delivery and E-commerce datasets, churners account for around 50% of the total customers, while in the gambling dataset, this percentage is slightly lower at 45%. Therefore, the data samples are quite balanced. For a more comprehensive description of the datasets, please refer to Appendix A, which includes detailed information such as feature types, descriptions, and distributions.

Table 4 – Churn and non-churn distribution.

	Food Delivery		Gambling		E-commerce	
	Customers	%	Customers	%	Customers	%
Churners	3620	50%	811	45%	12360	52%
Non-churners	3629	50%	990	55%	11542	48%
Total	7249	100%	1801	100%	23902	100%
Churn Time	30 days		120 days		90 days	

2.3.4 RFM Segmentation of Customer Data

RFM segmentation (CHENG; CHEN, Y.-S., 2009) is a data preparation technique that distinguishes customers from extensive datasets in a given number of classes. It is done based on measures of Recency, Frequency and Monetary Value, as described in the following:

Recency (R) refers to the time interval since the last purchase. A shorter interval indicates higher R;

Frequency (F) refers to the number of transactions within a specific period (e.g., annually, quarterly, or monthly). A higher frequency is denoted by a larger F;

Monetary value (M) represents the amount spent on purchases within a specific period, with a higher monetary value corresponding to a larger M.

2.4 MACHINE LEARNING FOR CHURN PREDICTION

Machine learning aims to enhance computational performance in specific tasks using observed data (GHAHRAMANI, 2015). ML algorithms are commonly used for doing Classification, Regression or Clustering. These algorithms fall into tree types of learning: supervised, unsupervised and reinforcement learning. In the case of this work, we are exclusively dealing with classifiers based on supervised learning.

Decision Tree (DT) is a tree-like model of decisions and their possible consequences, including chance events, resource costs, and utility (BURSTEINAS; LONG, 2000). Decision trees are constructed by recursively partitioning the data into subsets based on the values of the input features. At each node of the tree, a decision is made based on the value of a feature, and the data is split into two or more subsets. This process is repeated until a stopping criterion is met, such as a maximum tree depth or a minimum number of samples per leaf (SIERS; ISLAM, 2020), (KIM, S.; LEE, H., 2022).

Random Forests (RF) is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model (NEW. . . , 2022). In a Random Forest, each decision tree is built using a random subset of the features in the dataset, which helps to reduce overfitting and increase the diversity of the trees. During the training process, the algorithm creates a large number of decision trees, and each tree is trained on a random subset of the training data. When making a prediction, the Random Forest combines the predictions of all the trees to arrive at a final classification (ABDULLAH; PRASETYO, 2020), (COUSSEMENT; DE BOCK, 2013), (PERIŠIĆ; PAHOR, 2022).

Support Vector Machine (SVM) models are supervised and can be used for classification or regression. The input to an SVM classifier can be a set of features (feature-based methods) or rich representations structured like trees or graphs (kernel methods) (PENG, 2013). SVM works by finding the hyperplane that best separates the positive and negative instances in the feature space. The hyperplane is chosen to maximize the margin between the two classes, which is the distance between the hyperplane and the closest points from each class (LIANG et al., 2021), (XIAHOU; HARADA, 2022), (PERIÁÑEZ et al., 2017).

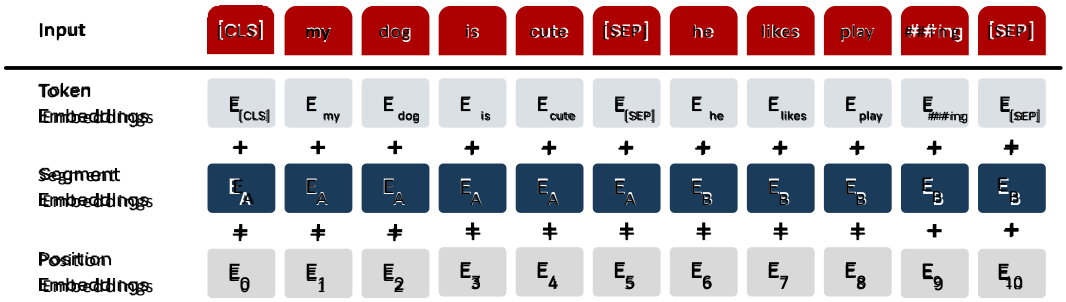
Multilayer Perceptron (MLP) is a type of artificial neural network that is capable of modeling many systems and is inclusive of other kinds of artificial neural networks. They are particularly useful for tasks that require a great deal of computation within a short time interval and have a high degree of flexibility to deal with numerous possible inputs (HUANG, 1992). MLPs consist of multiple layers of nodes, with each node in a layer connected to every node in the previous and next layers (ALMEIDA et al., 2019), (WU, X. et al., 2022).

2.5 NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a field of artificial intelligence that employs a variety of computational methods to make human language accessible to computers. It aims to give computers the ability to understand and generate human language (TORREGROSA et al., 2023). Text Embedding is a key concept in NLP. It allows representing tokens or words from a vocabulary, or sequences of them, as vectors which capture properties like meaning similarities and correspondences (INCITTI; URLI; SNIDARO, 2023), a longstanding challenge in NLP (PAWAR, 2018).

Word Embeddings project words into a vector space (YIN et al., 2023). Contextualized word embedding models, like BERT, also encode meaning variations determined by contexts. Current word embedding techniques originate from neural network models that predict the use of words in a text. They have become indispensable in NLP applications (RHEAULT; COCHRANE, 2018). Figure 1 illustrates the BERT contextualized embedding representation (DEVLIN et al., 2019). The embedded representation of each input is the sum of the respective token embedding, segmentation embedding and the position embedding.

Figure 1 – BERT input representation (DEVLIN et al., 2019).



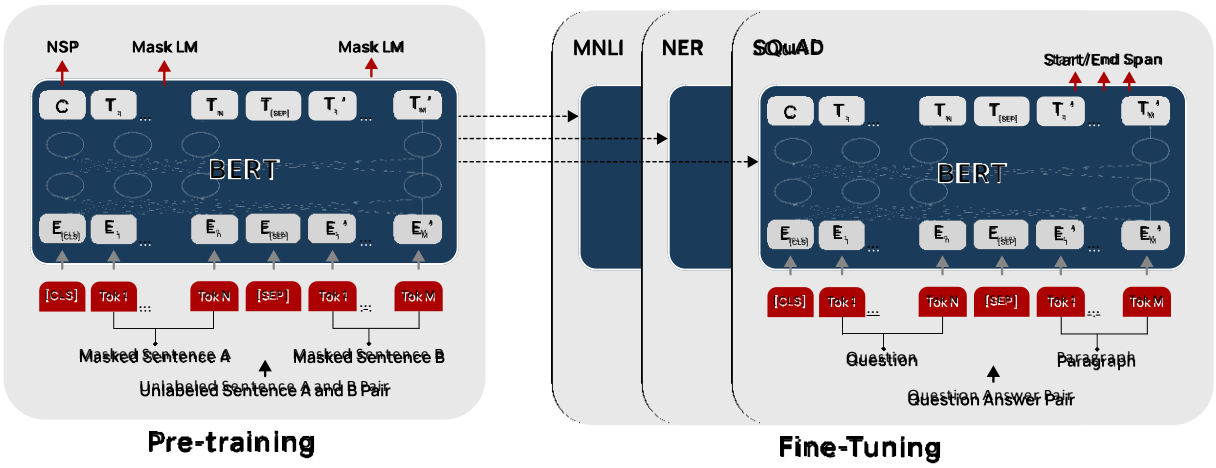
2.5.1 Large Language Models

Large Language Models (LLMs) are computational models that have the capability to understand and generate human language (CHANG et al., 2023). LLMs such as GPT, Bard, BERT and RoBERTa are pre-trained on large amounts of text data and can be fine-tuned for specific tasks, making them highly versatile (DU et al., 2023). They can be used to encode both text and rich layout information in visually rich documents (WEI, M.; HE, Y.; ZHANG, Q., 2020).

BERT uses the masked language modelling (MLM) objective, where some of the tokens of a input sequence are randomly masked, and the objective is to predict them. BERT applies a Transformer encoder to attend to bi-directional contexts during pre-training, jointly conditioning on both left and right context in all layers (LIU, Q.; KUSNER; BLUNSOM, 2020). In addition, BERT uses a next-sentence-prediction (NSP) objective. Given two input sentences, NSP predicts whether the second sentence is the

actual next sentence of the first sentence. The NSP objective aims to improve the tasks, such as question answering and natural language inference, which require reasoning over sentence pairs (CHOI et al., 2021).

Figure 2 – Overall pre-training and fine-tuning procedures for BERT (DEVLIN et al., 2019).



Apart from output layers, the same architectures are used in both pre-training and fine-tuning, see figure 2. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers) (DEVLIN et al., 2019).

RoBERTa is a post-BERT method, with simple modifications that include: (1) training the model longer, with bigger batches, over more data; (2) removing the next sentence prediction objective; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data (LIU, Y. et al., 2019).

LLama (GAO, P. et al., 2023), **Bard** (ALI et al., 2022) and **GPT** (WU, S. et al., 2023) are autoregressive LLMs that utilize deep learning to generate text resembling that produced by humans. In simple terms, it is a computational system designed to create sequences of words, code, or other data from an initial input called a prompt. These models are employed, for instance, in machine translation to statistically predict sequences of words. The language model is trained on an unlabeled dataset consisting of texts from sources such as Wikipedia and many other predominantly English-language websites, but also in other languages. To generate relevant results, it is necessary to train these LLMs with billions of parameters, instead of the the million required by previous models like BERT (FLORIDI; CHIRIATTI, 2020).

3 LITERATURE REVIEW

We conducted two bibliographical reviews, whose processes and results are described in this chapter. The review described in Section 3.1 aimed to identify the latest strides in ML models tailored for predicting customer churn on datasets sourced from HCT companies. The review described in Section 3.2 investigated advances centered around Natural Language Processing (NLP) or particularly Large Language Models (LLM), intended to predict churn using text features in both contexts, HCT and non-HCT.

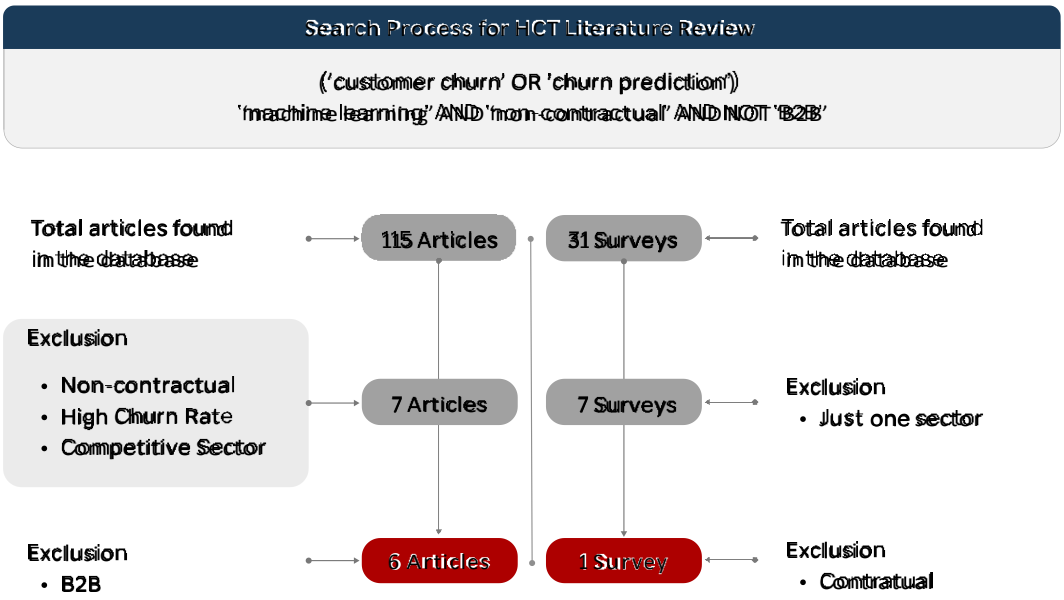
3.1 CHURN PREDICTION IN HCT ENTERPRISES

Related work often used the keywords 'machine learning', 'customer churn', and 'non-contractual'. In addition, we consider that churn prediction is more relevant in B2C environments. Thus, we employed the search expression:

('customer churn' OR 'churn prediction') AND
('machine learning') AND ('non-contractual') AND NOT ('B2B')

The search was conducted across multiple databases, including IEEE Xplore, Google Scholar, and Periódicos Capes. The search process resulted in the identification of 146 articles. Then, we manually filtered these articles by title, abstract, and keywords, considering only works published between January 2002 and February 2023, and applied the inclusion and exclusion criteria illustrated in Figure 3.

Figure 3 – Search process for literature review, with exclusion and inclusion criteria.



The articles found were classified in surveys or experimental. The 31 surveys were cataloged separately. Of these, only 6 ones had data from companies with non-

contractual negotiation, and only 1 made a distinction between companies with contractual and non-contractual data (AHN et al., 2020a). Upon excluding the surveys, 115 articles remained. However, only 7 of them met the criteria of non-contractual sales, high churn rate, and belonging to a highly competitive sector. However, out of the remaining 7, 1 article was focused on B2B and was excluded.

Table 5 compares the selected works and our proposal according with the sector of the data used to train and evaluate the ML models, the models that provided the best accuracy (ACC) in the respective works and their accuracy measures.

Table 5 – Churn Prediction in HCT Enterprises: Comparison of selected articles.

Work	Sector	Best Model	Acc.
(WU, X. et al., 2022)	Web Browser	MLP	94%
(PERIÁÑEZ et al., 2017)	Gaming	SVM	96%
(PERIŠIĆ; PAHOR, 2022)	Gaming	RF	71%
(COUSSEMENT; DE BOCK, 2013)	Gambling	RF	78%
(XIAHOU; HARADA, 2022)	E-Commerce	SVM	91%
(KIM, S.; LEE, H., 2022)	E-Commerce	DT	90%
This Work	Food Delivery, E-Commerce and Gambling	MLP, SVM, DT and RF	88.41% - 92.52%

Our bibliographical review and the survey of Sobreiro, Martinho, Alonso and Berrocal 2022 (SOBREIRO et al., 2022) revealed that 3 of the models chosen for our research are among the most used in published articles about churn prediction in general, namely: Random Forests in 30 articles, SVM in 29 articles, and Decision Trees in 52 articles. On the other hand, the Multilayer Perceptron was used in just 4 articles published until 2020, according to (SOBREIRO et al., 2022).

Four of the six articles listed in table 5 do not publish the datasets used in their experiments (PERIÁÑEZ et al., 2017; PERIŠIĆ; PAHOR, 2022; XIAHOU; HARADA, 2022; KIM, S.; LEE, H., 2022), while one required prior request (WU, X. et al., 2022). We were able to get just the “Gambling" dataset (COUSSEMENT; DE BOCK, 2013), on which the Random Forest algorithm achieved an accuracy of 78%. In our preliminary experiments the Random Forest achieved an accuracy of 92.52% on this dataset, due to different filtering methods and the application of the RFM segmentation technique.

Our research also employs other datasets which were not used in previous studies, such as the “Food Delivery" and the “E-commerce" datasets. To the best of our knowledge, our work is the first to employ the former in churn prediction in the “Food Delivery” industry.

3.2 NLP IN CHURN PREDICTION

We also realized a literature review to identify the use NLP advances to predict customer churn. Related work often used the keywords 'machine learning' and 'customer churn'. Thus, we employed the search expression:

('customer churn' OR 'churn prediction') AND ('machine learning') AND ('NLP' or 'LLM')

The search was conducted across multiple databases, including IEEE Xplore, Google Scholar, and Periódicos Capes. Then, we manually filter the retrieved articles by title, abstract, and keywords, considering only works published between January 2010 and February 2023 Our exclusion criteria were the lack of implementation of at least one classification model and the absence of English or Portuguese versions in articles. The search process led to the identification of 12 articles, and out of these, 5 articles met the criteria, illustrated in Figure 4.

Figure 4 – Search process for literature review, with exclusion and inclusion criteria.

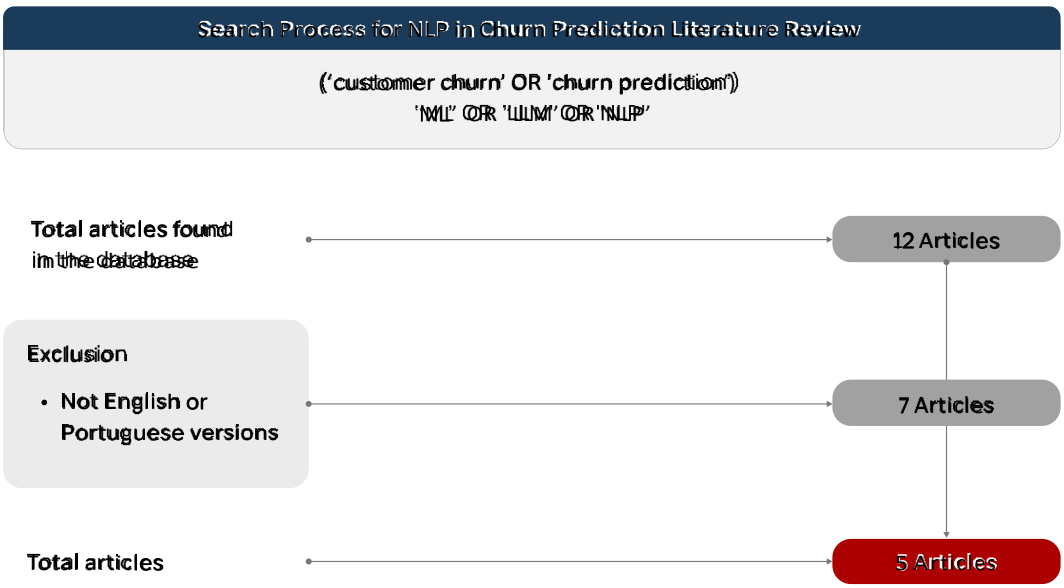


Table 6 lists the selected articles, with the kind of data employed in their respective experiments, ML techiques used and highest performanceobtained. Four of these studies use data from social media, specifically focusing on microblogging data from Twitter. Among these studies, two employed SVM and RNN models, yielding similar out-comes: around 75-78% F1-Score. Both of these studies were authored by Hadi Amiri and Hal Daume (AMIRI; DAUME III, 2016), (AMIRI; DAUME III, 2015). The third article employed a CNN (GRIDACH; HADDAD; MULKI, 2017) approach, achieving superior results compared to the first two studies (86% F1 score).

Table 6 – NLP in Churn Prediction: Comparison of selected articles.

Work	Data	Techniques	Performance
(AMIRI; DAUME III, 2016)	Microblog (tweets)	SVM	F1 Score: 75%
(AMIRI; DAUME III, 2015)	Microblog (tweets)	RNN	F1 Score: 78%
(GRIDACH; HADDAD; MULKI, 2017)	Microblog (tweets)	CNN	F1 Score: 84%
(IBITOYE; ONIFADE, 2022)	Microblog (tweets)	SVM	Accucary: 94%
(ZHONG; LI, 2019)	Transcript Data of Phone Calls (telecom)	CNN	F1 Score: 91%
This Work (2023)	Reviews (Food Delivery)	BERT, MLP, RoBERTa, RF, CNN, SVM	F1 Score, 50% - 97%

The fourth study, which also used data from microblogs, aimed to deepen the concept of "clustered social influence" within the realm of Twitter. Although the SVM exhibited favorable results, its assessment was constrained due to the limited dataset (IBITOYE; ONIFADE, 2022).

The fifth study also utilized the CNN model as its core framework, differentiating itself by employing proprietary data from a telecommunications company, specifically Phone Call Transcripts. This distinction sets it apart from previously published works. However, the use of quite old techniques like Word2vec requires a reconsideration of the choices made (ZHONG; LI, 2019).

The application of NLP for predicting churn remains relatively unexplored. In addition, the articles found present limitations in their applications or unsatisfactory results. According to Zhong and Li, two factors make social media less relevant for predicting churn: (1) customers prefer to contact the company directly instead of posting on social networks and (2) there is a shortage of training data (ZHONG; LI, 2019).

After reviewing the articles, the opportunity to implement these approaches with greater robustness and evaluate them on real datasets from companies that record customer reviews or interactions arose. For doing this, we utilize the food delivery dataset, focusing on the most recent purchase review of each customer (available for around 40% of the costumers). Furthermore, we exploit contemporary models like BERT and RoBERTa, which, to the best our knowledge, have not being applied for churn prediction yet (KOROTEEV, 2021), neither for just generating embeddings for further processing nor for being fine-tuned to solve the churn classification task. For comparative purposes, we complement the experiments by using Word2Vec with CBOW and Skip-gram for generating token embeddings to be feed as features to CNN (ZHONG; LI, 2019), SVM (IBITOYE; ONIFADE, 2022) and Random Forest (SOBREIRO et al., 2022) models.

4 THE CP-HCT FRAMEWORK

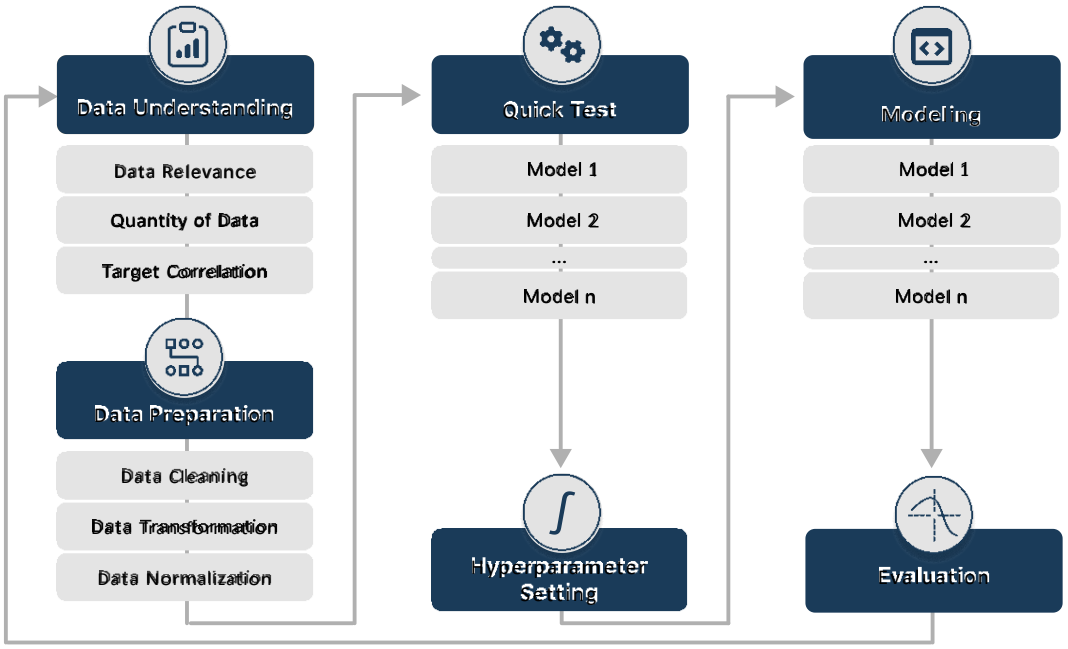
This chapter presents the framework designed to carry out the experiments carried out for this work. It supports an inspiration from CRISP-DM, encompassing Data Understanding, Data Preparation, Quick Test, Hyperparameter Setting and Evaluation.

4.1 THE CP-HCT FRAMEWORK

We developed a framework called CP-HCT (Churn Prediction in enterprises with High Customer Turnover)[1] to help face the challenge of predicting churn effectively in HCT enterprises. It helps preliminary training and testing of a number of ML models with a variety of parameter configurations. The goal is achieving the best results while reducing the computational costs for training of alternative models, with distinct hyperparemeter configurations for choosing the best ones.

Figure 5 illustrates the process supported by CP-HCT for training, testing, filtering, adjusting and evaluating ML models for churn prediction in HCT enterprises. It is an inspiration from the well-known CRISP-DM process (CHAPMAN et al., 2000), which has six phases in its traditional methodology. The main innovations of the CP-HCT framework are the modules to support the steps Quick Test and Hyperparameter Configuration. The other phases (Data Understanding Data, Peparation, Modeling and Evaluation) are those of CRISP-DM, but with a focus on churn in HCT enterprises. In the following subsections we explain each phase of the CP-HCT process that we have applied and the the modules that we have developed to support some of them.

Figure 5 – The CP-HCT process for building and selecting churn prediction models



[1] Available at https://github.com/WilliamBeckhauser/ChurnPredictionHCT

4.1.1 Data Understanding

During this phase, we assess the relevance and volume of the data by identifying missing values, outliers, and duplicate records. This step involves calculating the percentage of missing values per column, spotting potential discrepancies using statistical measures such as the Z-score, and exploring the possibility of generating new data points, like average values, for instance, the average amount spent per order.

We examine the dispersion of feature values and analyze the correlation between features and the target variable (churn). Moreover, we make adjustments like renaming and describing columns to streamline the analysis and address issues such as data imbalance. It is important to ensure the balance between customers who churned out and those that did not, for properly train the models.

4.1.2 Data Preparation

In this phase, we utilize insights from data understanding to structure, balance, and clean the data effectively. We handle null values by excluding or substituting them with "0" where applicable. Data type conversions, like turning integers into floats, ensure consistency. Categorical data, such as gender and country, are managed using one-hot encoding. This method transforms categorical variables into binary vectors, aiding analysis and processing.

The primary features are transactional data related to customer purchases. While profile data is limited, it allows the creation of new features like average purchase expenditure and percentage of goal achievement in games. Additionally, these features support dataset segmentation strategies such as RFM (Recency, Frequency, Monetary value) (WEI, J.-T.; LIN, S.-Y.; WU, H.-H., 2010). RFM segmentation categorizes customers effectively using shared features (e.g., purchase date, amount paid) divided into quartiles. This approach gauges segment importance and assigns a comprehensive score to customers, labeling them as Not Fans, Switchers, Loyal, or Champions.

For textual data, we use the Googletrans library to standardize to English. This involves a process that removes emojis, special characters, and converts text to lowercase.

4.1.3 Quick Test

In this phase, we perform a quick test using the most promising models from literature review, section 3, identified based on their performance on similar tasks or their suitability for the specific problem at hand. The quick Test allows us to discern the accuracy potential of our models, enabling the identification of options to be excluded. For instance, in a scenario with an extensive list of models to be tested. The models are

then trained after the data understanding and preparation phases, with hyperparameter setings recommended by the ML software, (PEDREGOSA et al., 2011).

4.1.4 Hyperparameter Setting

After the Quick Test, we conducted an extensive search for the best configuration of hyperparameters of the selected models, i.e., the configuration that yields optimal performance. In our first experiments, the filtering phase selected 4 models: Multilayer Perceptron, SVM, Decision Trees and Random Forest. In the second experiment we used the BERT and RoBERTa models. The list of hyperparameters for each of these models with the respective values that we tried for the respective parameter (second column) are listed in the appendix 13. The hyperparameter tuning tried all the combinations of parameter configurations to identify the optimal settings for each model.

By fine-tuning the hyperparameters of each model, we aimed to extract the most valuable characteristics of each model to achieve the best possible performance. The resulting hyperparameter configuratios are used in subsequent phases to evaluate and compare the performance of each model.

4.1.5 Evaluation

During the Evaluation phase, we measure the performance of our models using various evaluation metrics such as accuracy, recall, F1 score, precision, specificity, true positive rate, false positive rate, ROC-AUC, and cross-validation ROC-AUC. In this article, we focus on three key evaluation metrics: accuracy, F1 score, and precision. These metrics provide insights into the overall performance of the model and its ability to correctly predict positive instances (precision) and avoid false positives (F1 score).

5 DATA ANALYSIS AND PREPARATION

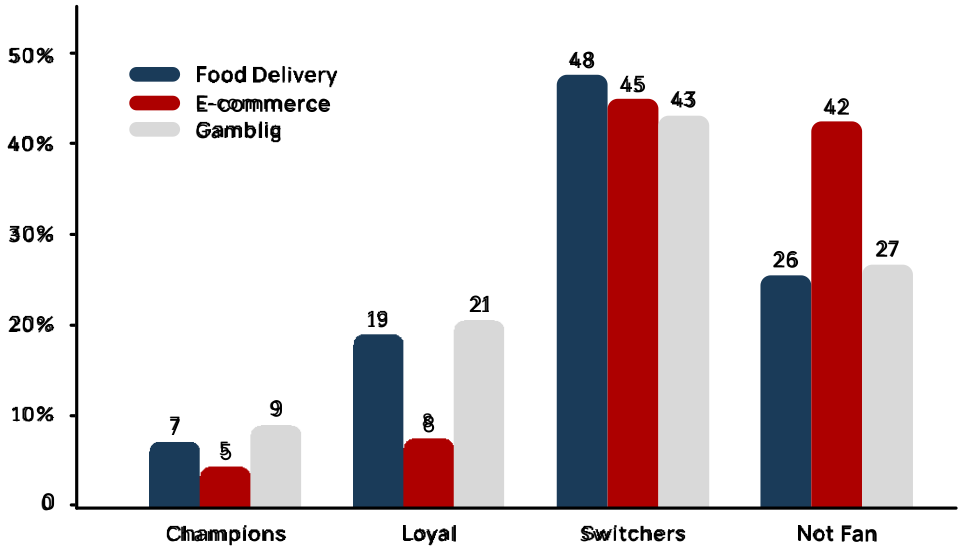
The chapter is structured into two sections: Transactional Data (5.1) and Textual Data (5.2). In the first section, we detail the process of data cleaning, including the removal of leading spaces and null value imputation, followed by customer segmentation using RFM analysis. This section also outlines the use of Pearson's correlation coefficient to filter features in datasets based on their correlation with churn. Techniques such as MinMaxScaler and OneHotEncoder are applied for data normalization and encoding. The second section examines customer reviews from the Food Delivery dataset, employing NLP techniques for multilingual data analysis, including Googletrans for translation and BERT.

5.1 TRANSACTIONAL DATA

Data preprocessing and analysis begin with data cleaning. This step involves removing leading spaces from column names, replacing null values with zeros, and performing type conversions for consistency across the dataset.

Following data cleaning, RFM (Recency, Frequency, Monetary Value) analysis is conducted for customer segmentation. Relevant columns are selected and renamed to align with the RFM framework. Quartiles for each RFM metric are calculated, assigning scores to customers. These scores enable segmentation into labels such as 'Not_Fan', 'Switchers', 'Loyal', and 'Champions', using regular expressions for specific RFM score combinations. The result of this segmentation is visually represented in Figure 6.

Figure 6 – Percentage Distribution of Customer Segments in RFM Analysis



In the data understanding and preparation phase, the dispersion of feature values and correlations between features and the target variable, churn, are examined using Pearson's linear correlation coefficient. Features with less than a 5% correlation in e-

commerce and gambling datasets and less than a 10% correlation in the food delivery dataset are excluded from further analysis. Table 7 provides a snapshot of descriptive statistics for some selected features. Appendix A.1 provides a comprehensive overview of the features, descriptions, and values for the three datasets.

The preprocessing involve steps supported by Python libraries: pandas is used for data frame manipulation, numpy for numerical computations, matplotlib and seaborn for visualization, and Google Colab's files module for file operations. Steps include normalizing data with MinMaxScaler and encoding categorical variables with OneHotEncoder from sklearn.preprocessing.

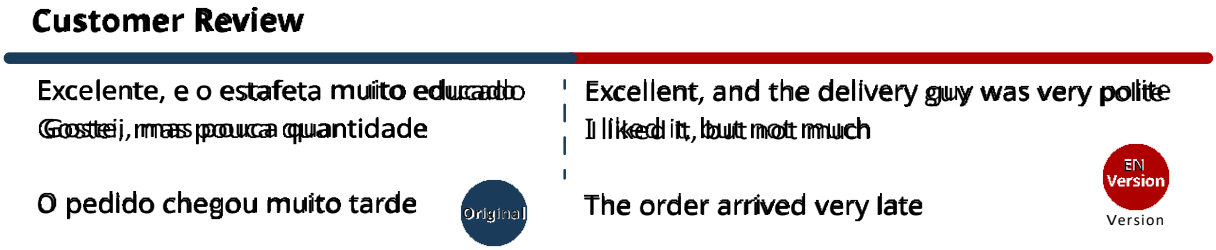
Table 7 – Statistics for some selected features from the Food Delivery dataset

index	orders	plates	totalValue	discountTotal	DaysLastOrder
count	7249	7249	7249	7249	7249
mean	10.55	13.75	88.78	27.51	51.48
std	16.68	44.96	132.57	97.08	53.16
min	1.0	0.0	0.0	0.0	1.0
25%	2.0	2.0	15.0	0.0	7.0
50%	4.0	5.0	37.0	0.0	30.0
75%	12.0	14.0	102.0	7.0	86.0
max	237.0	2672.0	985.0	978.0	182.0

5.2 TEXTUAL DATA

Among the scrutinized datasets, only Food Delivery contains textual data, in the form of customer reviews. Around 40% of the customers filled out a review for their respective last purchase. Figure 7 shows three examples of customer reviews, one per line, with the original version of the reviews on the left column, and their respective English version, obtained by using Googletrans, on the right column. Most reviews are short as these examples.

Figure 7 – Examples of reviews from the Food Delivery dataset.

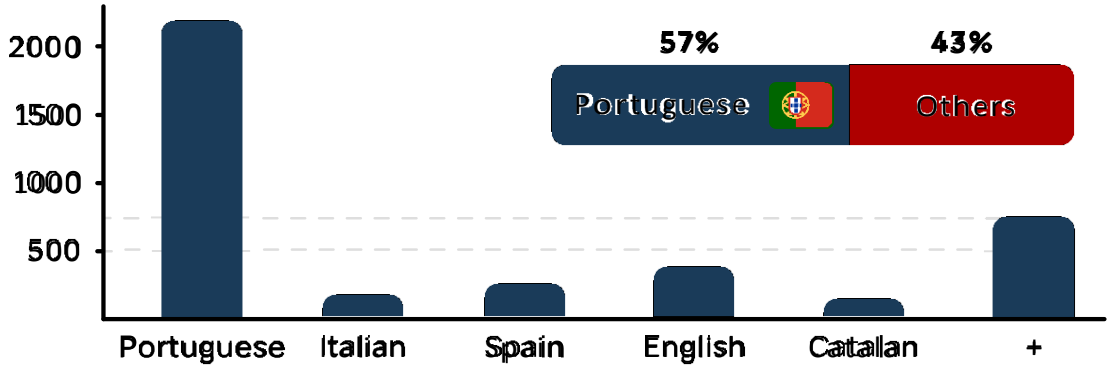


These reviews consist of concise texts describing customer experiences, encompassing a range of topics from complaints about undelivered items to discussions about meal prices, portion sizes and compliments.

This dataset contains reviews written in more than ten languages. Our analysis identified Portuguese (specifically from Portugal) as the most common language, ap-

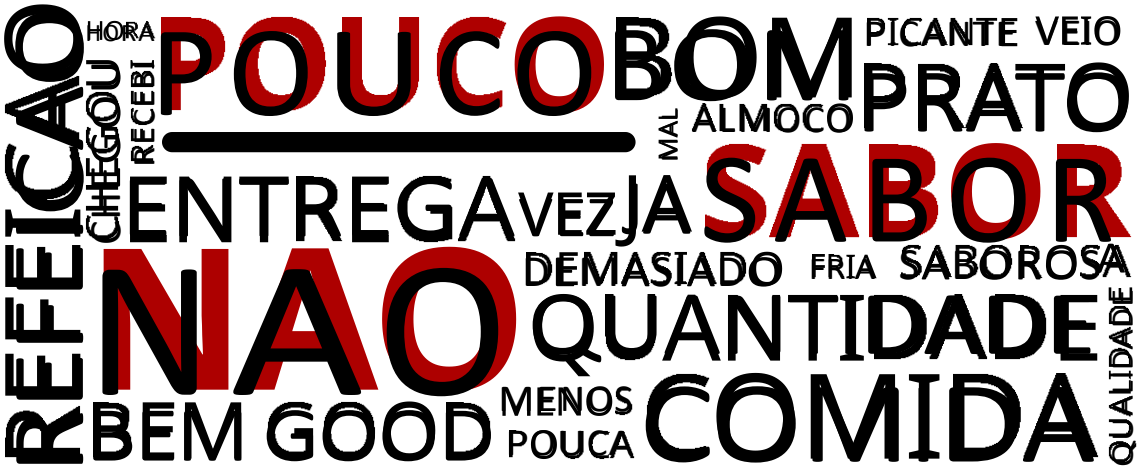
pearing in 57% of the reviews. English, Spanish, Italian, and Catalan also make up a significant portion, accounting for the remaining 43%, as shown in Figure 8.

Figure 8 – Distribution of Review Languages in the Dataset



The prevalence of Portuguese within the dataset results in the frequent use of certain keywords, including "sabor" (flavor), "quantidade" (quantity), "bom" (good), "comida" (meal), and "pouco" (little). These terms are prominently featured in the Word Cloud, Figure 9 highlighting their frequency and importance within the reviews.

Figure 9 – Visualizing Linguistic Trends: The Predominance of Key Portuguese Terms in Multilingual Reviews



To improve our analysis and draw comparisons between models trained on multilingual reviews versus those trained exclusively on English reviews, we utilized the Googletrans library for English translations. This initiative aimed at unifying the dataset's linguistic diversity to overcome the challenges of multi-language analysis. Despite the overall precision of the translations, our manual review process uncovered about 30 instances where the translations deviated from the original meanings, mainly in reviews that were succinct, often one or two words long. We limited our manual verification of translation fidelity to languages including Portuguese, Spanish, Catalan, and Italian. After completing the translations, we separated the dataset into two for

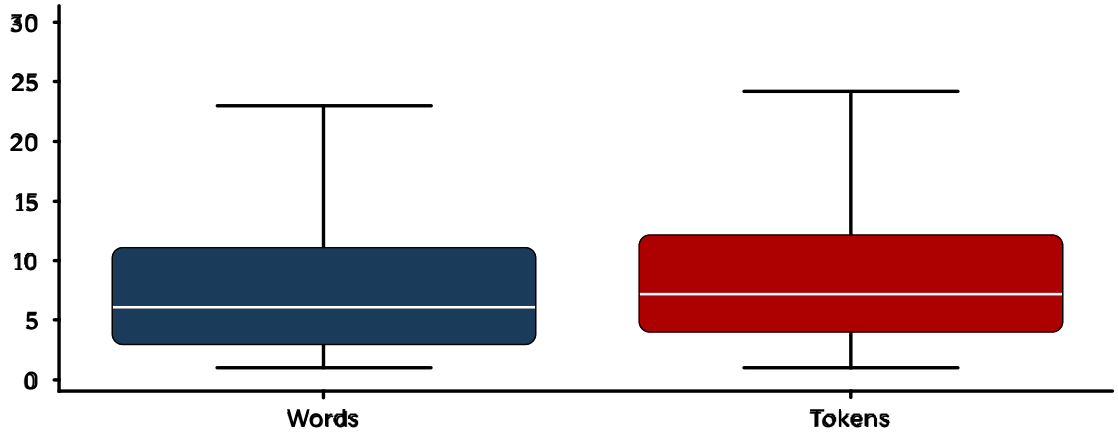
further analysis: one with reviews in multiple languages and the other with only English-language reviews.

Prior to initiating model training, the dataset underwent a preprocessing phase. We employed a variety of tokenization and embedding techniques for model training, such as BERT, RoBERTa, and Word2Vec, leveraging both skip-gram and CBOW models. The tokenization and embedding processes were greatly facilitated by the transformers library from Hugging Face, such as the use of BertTokenizer and BertModel for BERT embeddings. Furthermore, the PyTorch library, with its DataLoader and Dataset classes, played a crucial role in managing and batching data efficiently, enabling the model to process the data in segments during both training and evaluation.

The dataset was initially loaded from a CSV file into a DataFrame, providing a structured format with columns for 'id', 'comment', and 'churn'. It was then divided into training and test sets, with the test set comprising 20% of the data, using the train_test_split function from sklearn.model_selection to ensure the model's ability to generalize to new, unseen data.

A custom ChurnDataset class was created to manage the dataset, including methods for accessing individual entries and their attributes, such as comments and labels. This setup enabled precise encoding of comments via BertTokenizer, thus preparing the input for the BERT model by adding special tokens, attention masks, and padding to standardize sequence lengths. DataLoaders were set up for both training and testing datasets with defined batch sizes. The distribution of words and tokens is illustrated in Figure 10.

Figure 10 – Distribution of Words and Tokens in Preprocessed Dataset



6 EXPERIMENTS AND RESULTS

This chapter describes the experiments carried out within the scope of this work to determine the best features and models for churn prediction in enterprises with HCT. Section 6.1 reports the experiments using customer profile and transaction data to build and evaluate the machine learning models MLP, SVM, DT, and RF. Section 6.2 discusses additional investigations into multilingual customer feedback, and the impact of translating these reviews into English, using large language models such as BERT and RoBERTa for initial analysis. These studies also evaluate the performance of RF, SVM, MLP, and CNN models by employing embeddings generated through Word2Vec, utilizing both skip-gram and CBOW.

6.1 CHURN PREDICTION USING PROFILE AND TRANSACTION FEATURES

For the execution of the experiments, the Python programming environment on Google Colaboration, with the Pro plan, was used, providing robust computational resources such as 12.7 GB of RAM and 78.2 GB of disk space. The implementation of the machine learning models—Multilayer Perceptron, SVM, Decision Trees, and Random Forest—and performance evaluation were conducted using a set of scikit-learn libraries: MLPClassifier, SVC, DecisionTreeClassifier, and RandomForestClassifier for constructing the models. Additionally, model evaluation and selection processes employed libraries such as train_test_split for data partitioning, cross_val_score and GridSearchCV for model evaluation and hyperparameter tuning, and various metrics from sklearn.metrics for performance assessment, such as: accuracy_score, f1_score, precision_score, recall_score, roc_auc_score, roc_curve, classification_report, confusion_matrix.

The datasets were analyzed and prepared as described in section 5.1 to train the four models: Multilayer Perceptron, SVM, Decision Trees and Random Forests. The default scikit-learn hyperparameters (PEDREGOSA et al., 2011) were used to train the models for the quick test. Table 8 presents the preliminary results of this test. The Random Forest achieved the best accuracies on the Food Delivery and E-commerce datasets, namely 89.59% and 89.42%. However, SVM achieved the best accuracy of 92.52% on the Gambling dataset. Diff AVG refers to the difference between the accuracy of each model and the average accuracy (AVG) of the four models for the respective dataset.

During the Hyperparameter Setting phase, we searched for the best combination of hyperparameters for each model and dataset, by combining the selected values presented in Appendix 13. The combinations were generated using code created in this research. The resulting best combinations are shown in Appendix 14. The Random Forest best parameter configuration for the Gambling and E-commerce dasets

Table 8 – Preliminary results obtained by using default parameter values.

	Food Delivery		**Gambling**		**E-commerce**	
	ACC	Diff.AVG	ACC	Diff.AVG	ACC	Diff.AVG
Multilayer Perceptron	89.0%	1.7%	91.2%	-0.3%	87.7%	0.2%
Support Vector Machine	84.1%	-3.2%	**92.5%**	0.9%	88.4%	0.8%
Decision Trees	86.6%	-0.8%	91.1%	-0.4%	84.7%	-2.8%
Random Forest	**89.6%**	2.3%	91.4%	-0.2%	**89.4%**	1.9%
AVG	87.3%		91.6%		87.6%	

combines the 'entropy' criterion, 100 estimators, None for the maximum depth. For the Food Delivery dataset, the best configuration is 10 for the maximum depth set and 50 for the number of estimators. The SVM model uses 'rbf' kernel, C=1, 'scale' gamma for Gambling and E-commerce.

The DT model vest configuration uses the 'entropy' criterion, 10 for maximum depth, 2 for minimum samples split, and 4 for minimum samples leaf for the Gambling. For Food Delivery 10 is better for minimum samples split. The hyperparameters are the same for MLP on the three datasets, with hidden layer sizes 5, alpha=0.001, solver='adam', and activation='relu'. A test set size of 0.2 was employed throughout the training process, i.e. taking 20% of the data for testing purposes.

Table 9 presents the final results. The Random Forest model provided the highest accuracy (90.69%) for the Food Delivery dataset, followed by the Decision Tree model (90.21%), and the Multilayer Perceptron model (88.41%). The precision scores for all models were close, with the Random Forest model having the highest precision score (91.09%).

Table 9 – Final results after parameter optimization.

		Accuracy	**F1-score**	**Precision**
Food Delivery	Multilayer Perceptron	88.4%	88.4%	89.7%
	Support Vector Machine	87.9%	88.0%	78.6%
	Decision Tree	90.2%	89.9%	90.2%
	Random Forest	**90.7%**	**90.9%**	**91.1%**
Gambling	Multilayer Perceptron	90.6%	90.2%	91.4%
	Support Vector Machine	**92.5%**	**92.4%**	**93.2%**
	Decision Tree	90.9%	90.7%	90.9%
	Random Forest	**92.5%**	91.3%	92.6%
E-commerce	Multilayer Perceptron	89.2%	89.2%	90.1%
	Support Vector Machine	89.1%	89.0%	**90.9%**
	Decision Tree	88.4%	88.1%	81.7%
	Random Forest	**89.6%**	**89.3%**	**90.9%**

For the Gambling datase, the Support Vector Machine and Random Forest models presented the highest accuracy scores (both at 92.52%), followed by the Decision Tree (90.86%). The precision scores for all models are relatively close, with the Support Vector Machine having the highest precision score (93.17%).

Finally, for the E-commerce dataset, the Random Forest model yielded the highest accuracy (89.56%), followed by the SVM (89.06%), and the MLP (89.17%). The precision scores for all models are close, with the SVM algorithm having the highest precision score (90.91%).

Additional tests were conducted without employing RFM segmentation for training and testing the models.This led to a consistent decrease of at least 15% in accuracy across all models. The RF model achieved the highest accuracy, reaching 74%.

Overall, the results show that the RF algorithm performs the best across the three domains, with the SVM algorithm also performing well. The Multilayer Perceptron algorithm has the lowest accuracy scores across the datasets of the three domains, although the difference in performance between then is quite small.

The random forest achieved superior results due to its ability to adapt to non-linear data (RADHIKA et al., 2023). This is because our features range from customer marketing permissions to whether it's a lunch or dinner customer.

6.2 CHURN PREDICTION USING MULTILINGUAL AND TRANSLATED REVIEWS

The computational experiments were conducted on a MacBook Pro with an Apple M2 Pro chip, featuring a 12-core CPU and a 19-core GPU. System specifications comprised 16GB of unified memory and a 512GB SSD. Jupyter Notebook served as the development and execution environment.

Python libraries essential for the experiments were selected for ML and NLP tasks. The transformers library was used for loading and fine-tuning BERT and RoBERTa models. The gensim library facilitated the training and application of Word2Vec models for text vectorization. For ML models such as SVM and RF, and for data preprocessing tasks like dataset splitting, the scikit-learn library was employed. PyTorch was chosen for training CNN models, leveraging GPU acceleration to enhance training efficiency.

In this experiment, we utilized a portion of the Food Delivery dataset that consists of textual reviews of the most recent purchase, which are available for around 40% of the customers. These reviews went through a preprocessing procedure involving the removal of emojis and punctuation marks, as well as being converted to lowercase. The dataset was divided into training and test sets, using a holdout method, with each set representing 80% and 20% of the data, respectively.

Text tokenization and encoding into embeddings were accomplished using BERT, RoBERTa, as well as Word2Vec with both Skip-gram and CBOW. BERT and RoBERTa models were loaded and fine-tuned for the task of binary sequence classification. The outcomes obtained from Word2Vec were utilized as inputs for the SVM, CNN, MLP, and RF.

In the initial testing phase, we applied hyperparameters that had been explored in prior research pertaining to the BERT and RoBERTa models. This involved utilizing

12 epochs and a batch size of 16. Both models produced meaningful results, achieving an F1 Score of approximately 85%. The models trained using Word2Vec, specifically the CNN model, exhibited inferior performance and were excluded from the subsequent stage, as their results were below 60%.

During the model performance and hyperparameter optimization phase, we divide it into two subsections.

6.2.1 Word2Vec Embeddings

The models trained with Word2Vec embeddings, both CBOW and Skip-gram, exhibited lower performance compared to those trained with BERT and RoBERTa embeddings. However, an improvement was still observed when the reviews were translated into English. This trend was consistent across RF, SVM, and MLP models. The performance metrics and hyperparameters for models trained with Word2Vec embeddings are presented in Table 10.

Table 10 – Performance Metrics and Hyperparameters using Word2Vec Embeddings

Model	Language	Accuracy	F1-Score	Model Specific Hyperparameters
				Word2Vec CBOW
RF	Multilingual	72.95%	72.26%	Max depth: 10, Estimators: 100
RF	Translated	75.74%	75.64%	Max depth: 20, Estimators: 200
SVM	Multilingual	62.05%	52.07%	C: 10, Kernel: linear
SVM	Translated	72.52%	71.41%	C: 10, Kernel: linear
MLP	Multilingual	67.56%	67.60%	Activation: relu, Layers: (100,), Iter: 300
MLP	Translated	71.85%	71.69%	Activation: tanh, Layers: (100,), Iter: 200
				Word2Vec Skip-gram
RF	Multilingual	74.02%	73.61%	Max depth: 10, Estimators: 200
RF	Translated	78.41%	78.37%	Max depth: 10, Estimators: 200
SVM	Multilingual	73.76%	73.09%	C: 10, Kernel: linear
SVM	Translated	78.55%	78.65%	C: 10, Kernel: rbf
MLP	Multilingual	75.23%	74.65%	Activation: tanh, Layers: (50,), Iter: 300
MLP	Translated	79.76%	79.95%	Activation: relu, Layers: (100,), Iter: 300

6.2.2 BERT

For models utilizing BERT embeddings, a noticeable improvement in both accuracy and F1-score was observed when reviews were translated into English before training. Specifically, the English translations led to higher performance metrics across all models tested, including RF, MLP, SVM, and MLP. The optimal hyperparameters varied between the original and English datasets, reflecting the distinct characteristics of the multilingual and translated texts. The detailed performance metrics and hyperparameters are summarized in Table 11. Beyond the contents of the table, experiments incorporating embeddings and fine-tuning with the RoBERTa model were conducted.

However, these trials did not surpass BERT's performance in any scenario, showing an average decline of 3-5%.

Table 11 – Performance and Hyperparameters of Models Trained with BERT Embeddings

Model	Language	Accuracy	F1-Score	Hyperparameters
RF	Multilingual	93.40%	92.20%	bootstrap: False, max_depth: None, min_samples_leaf: 4, min_samples_split: 2, n_estimators: 100
RF	Translated	**96.11%**	95.15%	bootstrap: True, max_depth: 30, min_samples_leaf: 4, min_samples_split: 2, n_estimators: 200
SVM	Multilingual	93.46%	93.64%	C: 10, degree: 2, gamma: 'scale', kernel: 'rbf'
SVM	Translated	**96.19%**	**96.34%**	C: 0.1, degree: 2, gamma: 'scale', kernel: 'linear'
MLP	Multilingual	93.03%	93.23%	activation: 'tanh', alpha: 0.05, hidden_layer_sizes: (50, 50), learning_rate: 'adaptive', solver: 'sgd'
MLP	Translated	**96.61%**	**96.65%**	activation: 'tanh', alpha: 0.001, hidden_layer_sizes: (50, 100, 50), learning_rate: 'constant', solver: 'sgd'
Fine-tune	Original	86,94%	84,48%	Ír': 2e-05, 'batch_size': 32
Fine-tune	Translated	89,14%	86,58%	Ír': 2e-05, 'batch_size': 16

6.2.3 Comparison of Word2Vec and BERT Results

The experimental results show a clear advantage in translating multilingual reviews into English before model training, particularly for models leveraging BERT embeddings. This improvement can be attributed to the extensive pre-training of these models on diverse English corpora, which likely enhances their understanding and representation of the text when it is in English.

Furthermore, the variation in optimal hyperparameters between the original and translated texts suggests that the characteristics of the language significantly influence model tuning and performance. Specifically, models trained on English texts tended to perform better with different configurations of hyperparameters, indicating the necessity of a tailored approach to model optimization based on the language of the text.

It is also noteworthy that despite the lower performance of Word2Vec models compared to BERT and RoBERTa, the translation of reviews into English still resulted in performance gains.

6.3 DISCUSSION

The RFM segmentation of customers, in the experiments with conventional data of their profiles and transactions, led to an average increase of 15% in prediction accuracy. This result may raise doubts, mainly because one of the RFM segmentation criteria considers the date of the last purchase. It could have introduced a bias in the training data, since the churn in each dataset was defined as at least a certain number of days without new purchases after the last one. We carried out further experiments to investigate this possible bias. Table 12 shows the models accuracy with segmentation using the three RFM measures or just pairs of them. Excluding F or M led to an accuracy increase of up to 5% over RFM. However, the omission of the R measure resulted in an accuracy decline between 13% and 20%. It suggests that, in fact, the use of the R measure for customer segmentation, as we initially did, allows dubious gains in models' performance. We found out that it is a common practice in the literature, without mentioning this bias. One possibility to alleviate it would be to increase the number of segmentation labels (in addition to the four labels used in this work ('Not_Fan', 'Switchers', 'Loyal', and 'Champions'), for diluting the bias.

Table 12 – Models' accuracy with costumer segmentaion using different combinations of RFM measures

	Model	**RFM**	**RF**	**RM**	**FM**
Food Delivery	MLP	88.4%	90.21%	90.28%	**75.59%**
	SVM	87.9%	90.62%	90.83%	74.21%
	DT	90.2%	**90.83%**	90.62%	74.69%
	RF	**90.7%**	90.55%	**91.17%**	75.31%
Gambling	MLP	90.6%	**95.84%**	94.46%	74.52%
	SVM	**92.5%**	95.01%	**95.29%**	**74.79%**
	DT	90.9%	91.69%	93.91%	70.64%
	RF	92.5%	90.86%	93.07%	72.85%
E-commerce	MLP	89.2%	89.40%	89.67%	72.91%
	SVM	89.1%	89.29%	89.92%	73.25%
	DT	88.4%	89.01%	90.42%	70.26%
	RF	**89.6%**	**89.44%**	**91.17%**	**76.45%**

Still in the first experiment, the Random Forest model showed superiority in most datasets mainly due to its ability to deal with non-linear data. This is particularly relevant in contexts where customer characteristics are diverse and complex, as can be seen in the datasets from the food delivery, e-commerce and gambling sectors. Customer data in these areas does not adhere to traditional sale models, such as contractual ties.

The experiments with the texts of the customer reviews highlighted gains in model efficacy when reviews were translated into English. It is probably due to the superiority of language models like BERT trained with large volumes of English corpora, instead of the Portuguese language of the original texts. The increase in performance

can be linked to the fusion of synonymous or similar terms into unified expressions in English, thus simplifying connections. For instance, the terms 'estafeta' and 'entregador' are consolidated under 'delivery guy'.

The superior performance of models working solely on textual data from customer reviews over those working with conventional data of customer profiles and transactions can be attributed to the rich variety of information present in the review texts. Though usually short, many reviews present customer complaints and other information that can be crucial for predicting churn and understanding the reasons.

In addition, the methodology used to collect review data, including the review texts and numeric scores may also be a relevant factor. The methodology employed to collect the data used in this work allows feedback for each item purchased, including a numeric rating and, for items with negative ratings (between 1 and 3), a text explaining the reasons. In contrast, textual feedback on positive experiences (ratings from 4 to 5) is limited to the end of the review form, being suggested as feedback a "suggestion for the business". This methodology results in unbalancing of the amount of text data available for negative and positive reviews. Negative reviews usually have much more detailed text explaining their reasons. The requirement to evaluate all items before submitting a positive textual review, may also contribute to loss of text about positive and neutral reviews. It may cause biases in the dataset toward negative feedback, shortening the amount of text available about positive feedback, and excluding neutral perspectives. It may have simplified the binary classification task for churn prediction. Models trained with transactional data, on the other hand, are not subject to this bias, suggesting the potential value of conducting experiments using review and transactional feature inputs. However, much more investigation is required to fully analyze and understand the impact of the review data collecting methodology on churn classification results and the proper use of the collected data. It is beyond the scope of this work, but we plan go further in theses issues in future work, which will consider data collected by using different methodologies.

Finally, the combined use of conventional data (e.g., profile and transaction data, numeric scores of reviews) with textual data (e.g., review texts from several sources) may lead to more gains. Furthermore, analyzing the content of the reviews could show possible recurring customer problems, such as delays in deliveries. With the possibility of creating new metrics, it would be possible to incorporate them into the data sets. Enabling possible featurings with more significant correlations with the churn target.

7 CONCLUSIONS AND FUTURE WORK

This work compared alternative ML and language models for churn prediction on datasets of distinct HCT enterprises, using the CP-HCT framework introduced in this work. This framework, described in detail in section 4, is available for reuse by the ML community. It allows efficiently training and selecting ML models by supporting an extension of the CRISP-DM process. Its main extension is a quick test, between the Data Preparation and the Hyperparameter Setting phases of CRISP-DM, to select models trained with default parameter configurations, based on their performance. It allows the speed-up of the subsequent Hyperparameter Setting, Modeling and Evaluation phases, by concentrating efforts on the most promising models for the datasets at hand. The CP-HCT framework also drives the data preparation phase for leveraging RFM segmentation of transactional data to create categorical data for training.

The goal of our experiments was to identify models and data features that provide the best results for churn prediction in industries with HCT. In our first experiment, purchase transaction data were the basis for model training and evaluation in all the HCT scenarios considered. The quick test enabled the fast selection of the most promising models. After careful hyperparamenter setting and modeling in our first experiment, the Random Forest consistently presented the best accuracy in most datasets considered. Other models used in related work (section 3), such as SVM and Multilayer Perceptron, also yielded good results. The best results presented in this paper were obtained by using RFM. Without it the results had 15% lower accuracy in average.

In our second experiment, it was observed that the MLP and SVM models, when employed with embeddings from BERT, produced highly satisfactory performances, registering more than 96% in F1-score. This finding emerges from the study of a specific subset within the food delivery service database, focusing on recently made customer reviews. It is noteworthy that, even by limiting the analysis to a single variable, the preliminary results obtained were considerably higher in comparison to those derived from transaction data whether or not RFM segmentation was used.

7.1 PUBLICATIONS TIMELINE

Figure 11 presents the timeline to develop this work and generate 4 articles:

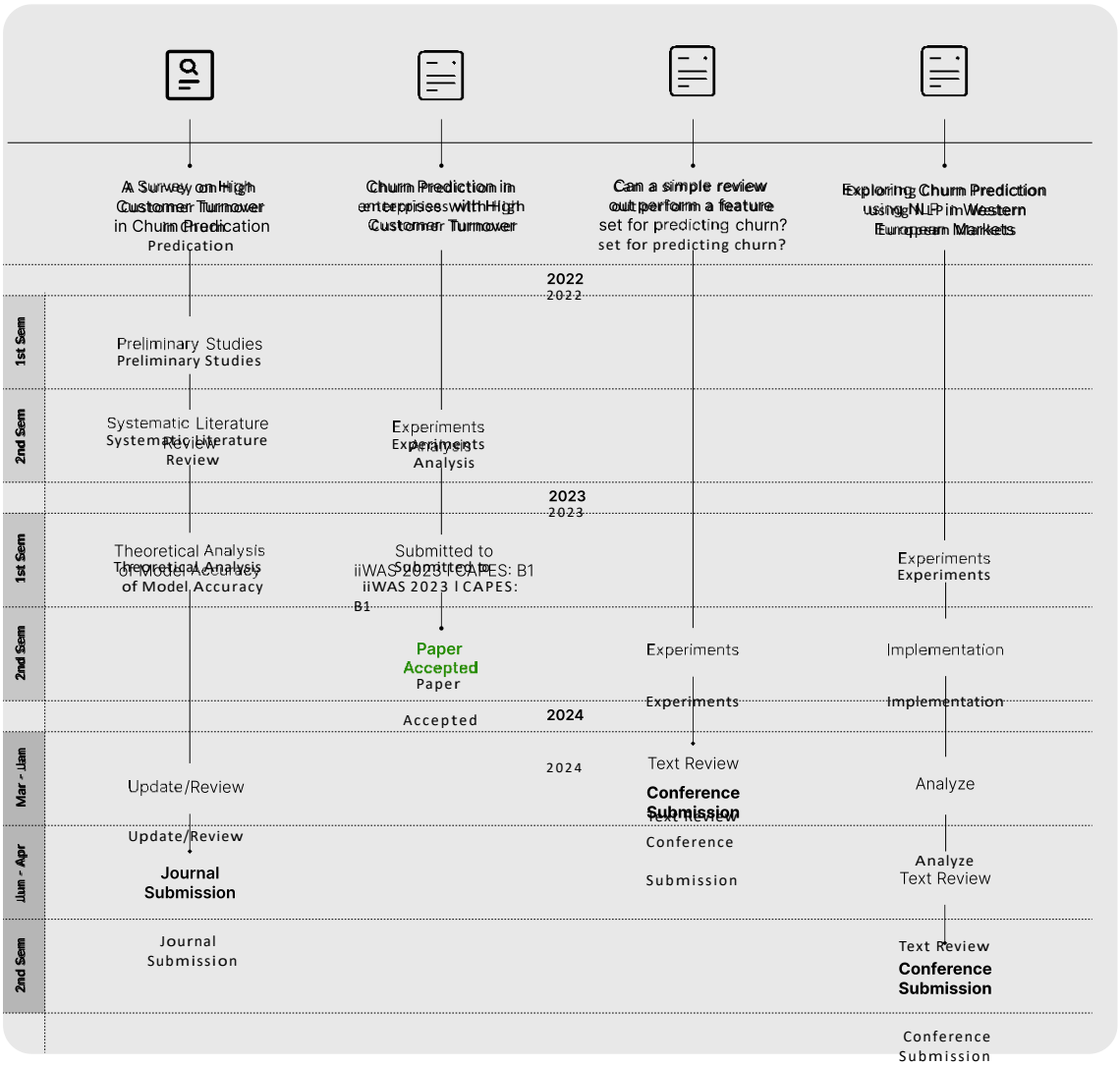
A Survey on High Customer Turnover in Churn Predication: a systematic review about churn prediction HCT environments;

Churn Prediction in enterprises with High Customer Turnover: reports experiments with 4 ML models, using 3 datasets from different sectors. (Accepted for publication at iiWAS 2023, B1 event in Qualis CAPES, (BECKHAUSER; FILETO, 2023));

Can a simple review out perform a feature set for predicting churn?: employs NLP techniques to exploit user reviews and features for churn prediction;

Churn Prediction using NLP in Western European Markets: Reports a case study of our proposal for churn prediction in an European Food Delivery company.

Figure 11 – Publications Timeline



7.2 FUTURE WORK

Our envision for future work beyond the scope of this work include: (i) expand research with NLP techniques for predicting churn; (ii) conducting multiclass classification studies on textual data to ascertain churn motivations; (iii) surveying related research on churn prediction in HCT enterprises to better determine key characteristics and differences from other areas; (iv) expanding research about churn prediction to more companies of distinct sectors with HCT; (v) updating research that aims to compare customer acquisition costs versus retention costs in contexts with and without HCT; (vi)

building a common database and benchmarks to compare alternative approaches for churn prediction in HCT enterprises.

REFERENCES

ABDULLAH, Sarini; PRASETYO, G V. EASY ENSEMMBLE WITH RANDOM FOREST TO HANDLE IMBALANCED DATA IN CLASSIFICATION. In.

AHN, Jaehyun; HWANG, Junsik; KIM, Doyoung; CHOI, Hyukgeun; KANG, Shinjin. A Survey on Churn Analysis in Various Business Domains. **IEEE Access**, v. 8, p. 220816–220839, 2020.

AHN, Jaehyun; HWANG, Junsik; KIM, Doyoung; CHOI, Hyukgeun; KANG, Shinjin. A survey on churn analysis in various business domains. **IEEE Access**, IEEE, v. 8, p. 220816–220839, 2020.

ALI, Rohaid et al. Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. **Neurosurgery**, LWW, p. 10–1227, 2022.

ALMEIDA, I.; VARGAS, M.; NOBRE, L.; SILVA, Bruno Légora Souza da; BULCÃO, André; LANDAU, Luiz; EVSUKOFF, Alexandre. Machine Learning applied in Swell Noise classification. **Proceedings of the 16th International Congress of the Brazilian Geophysical Society&Expogef**, 2019.

AMIRI, Hadi; DAUME III, Hal. Short Text Representation for Detecting Churn in Microblogs. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 30, n. 1, Mar. 2016.

AMIRI, Hadi; DAUME III, Hal. Target-Dependent Churn Classification in Microblogs. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 29, n. 1, Feb. 2015.

BECKHAUSER, William Jones; FILETO, Renato. Churn Prediction in Enterprises with High Customer Turnover. In: SPRINGER. INTERNATIONAL Conference on Information Integration and Web Intelligence. [S.l.: s.n.], 2023. P. 176–191.

BERRY, Michael JA; LINOFF, Gordon S. **Data mining techniques: for marketing, sales, and customer relationship management**. [S.l.]: John Wiley & Sons, 2004.

BURSTEINAS, Borisas; LONG, James Allen. Transforming supervised classifiers for feature extraction. **Proceedings 12th IEEE Internationals Conference on Tools with Artificial Intelligence. ICTAI 2000**, p. 274–280, 2000.

CHANG, Yupeng et al. A survey on evaluation of large language models. **arXiv preprint arXiv:2307.03109**, 2023.

CHAPMAN, Pete; CLINTON, Julian; KERBER, Randy; KHABAZA, Thomas; REINARTZ, Thomas; SHEARER, Colin; WIRTH, Rudiger. **CRISP-DM 1.0 Step-by-step data mining guide**. [S.l.], Aug. 2000.

CHENG, Ching-Hsue; CHEN, You-Shyang. Classifying the segmentation of customer value via RFM model and RS theory. **Expert systems with applications**, Elsevier, v. 36, n. 3, p. 4176–4184, 2009.

CHOI, Hyunjin; KIM, Judong; JOE, Seongho; GWON, Youngjune. Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks. In: 2020 25th International Conference on Pattern Recognition (ICPR). [S.l.: s.n.], 2021. P. 5482–5487.

COUSSEMENT, Kristof; DE BOCK, Koen. Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. **Journal of Business Research**, v. 66, n. 9, p. 1629–1636, 2013.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: PROCEEDINGS of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, June 2019. P. 4171–4186.

DU, Yingpeng; LUO, Di; YAN, Rui; LIU, Hongzhi; SONG, Yang; ZHU, Hengshu; ZHANG, Jie. **Enhancing Job Recommendation through LLM-based Generative Adversarial Networks**. [S.l.: s.n.], 2023. arXiv: 2307.10747 [cs.IR].

FLORIDI, Luciano; CHIRIATTI, Massimo. GPT-3: Its nature, scope, limits, and consequences. **Minds and Machines**, Springer, v. 30, p. 681–694, 2020.

GALLO, Amy. **The value of keeping the right customers**. [S.l.: s.n.], Nov. 2014. Available from: https://hbr.org/2014/10/the-value-of-keeping-the-right-customers.

GAO, Lily; HAAN, Evert de; MELERO-POLO, Iguácel; SESE, F Javier. Winning your customers' minds and hearts: Disentangling the effects of lock-in and affective customer experience on retention. **Journal of the Academy of Marketing Science**, Springer, v. 51, n. 2, p. 334–371, 2023.

GAO, Peng et al. Llama-adapter v2: Parameter-efficient visual instruction model. **arXiv preprint arXiv:2304.15010**, 2023.

GEILER, Louis; AFFELDT, Séverine; NADIF, Mohamed. A survey on machine learning methods for churn prediction. **International Journal of Data Science and Analytics**, Springer, v. 14, n. 3, p. 217–242, 2022.

GHAHRAMANI, Zoubin. Probabilistic machine learning and artificial intelligence. **Nature**, Nature Publishing Group UK London, v. 521, n. 7553, p. 452–459, 2015.

GRIDACH, Mourad; HADDAD, Hatem; MULKI, Hala. Churn Identification in Microblogs using Convolutional Neural Networks with Structured Logical Knowledge. In.

HASHMI, Nabgha; BUTT, Naveed Anwer; IQBAL, Dr.Muddesar. Customer Churn Prediction in Telecommunication A Decade Review and Classification. **IJCSI**, v. 10, p. 271–282, Sept. 2013.

HOCHSTEIN, Bryan; VOORHEES, Clay M; PRATT, Alexander B; RANGARAJAN, Deva; NAGEL, Duane M; MEHROTRA, Vijay. Customer success management, customer health, and retention in B2B industries. **International Journal of Research in Marketing**, Elsevier, v. 40, n. 4, p. 912–932, 2023.

HUANG, Shih-Chi. Multilayer perceptrons for image data compression and speech recognition. In.

IACOBUCCI, Dawn; HIBBARD, Jonathan D. Toward an encompassing theory of business marketing relationships (BMRS) and interpersonal commercial relationships (ICRS): An empirical generalization. eng. **Journal of interactive marketing**, Elsevier Inc, New York, v. 13, n. 3, p. 13–33, 1999. ISSN 1094-9968.

IBITOYE, Ayodeji; ONIFADE, Olufade. Sequentially Clustered Social Opinion for Improved Customer Management in Churn prediction. In.

INCITTI, Francesca; URLI, Federico; SNIDARO, Lauro. Beyond word embeddings: A survey. **Information Fusion**, Elsevier, v. 89, p. 418–436, 2023.

JAIN, Hemlata; KHUNTETA, Ajay; SRIVASTAVA, Sumit. Telecom churn prediction and used techniques, datasets and performance measures: a review. **Telecommunication Systems**, Springer, v. 76, p. 613–630, 2021.

KAGGLE. **Brazilian E-Commerce Public Dataset by Olist Kernel Description**. 2018. Available from: https://www.kaggle.com/olistbr/brazilian-ecommerce. Visited on: 20 Feb. 2023.

KANDIL, Farah Ahmed Hamed. Customer Churn in Internet Service Providers, 2023.

KIM, Daeyoung; LEE, Eunjung; RHEE, Wonjong. Churn prediction of mobile and online casual games using play log data. eng. **PloS one**, Public Library of Science, United States, v. 12, n. 7, e0180735–e0180735, 2017. ISSN 1932-6203.

KIM, Sulim; LEE, Heeseok. Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees. **Procedia Computer Science**, v. 199, p. 1332–1339, 2022. The 8th International Conference on Information Technology and Quantitative

Management (ITQM 2020 2021): Developing Global Digital Economy after COVID-19. ISSN 1877-0509.

KOROTEEV, MV. BERT: a review of applications in natural language processing and understanding. **arXiv preprint arXiv:2103.11943**, 2021.

LIANG, Jinwen; QIN, Zheng; NI, Jianbing; LIN, Xiaodong; SHEN, Xuemin (Sherman). Practical and Secure SVM Classification for Cloud-Based Remote Clinical Decision Services. **IEEE Transactions on Computers**, v. 70, p. 1612–1625, 2021.

LIU, Qi; KUSNER, Matt J.; BLUNSOM, Phil. **A Survey on Contextual Embeddings**. [S.l.: s.n.], 2020. arXiv: 2003.07278 [cs.CL].

LIU, Yinhan et al. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. [S.l.: s.n.], 2019. arXiv: 1907.11692 [cs.CL].

MAKOWSKI, Louis; OSTROY, Joseph M. Perfect Competition and the Creativity of the Market. eng. **Journal of economic literature**, American Economic Association, NASHVILLE, v. 39, n. 2, p. 479–535, 2001. ISSN 0022-0515.

MARTÍNEZ, Andrés; SCHMUCK, Claudia; PEREVERZYEV, Sergiy; PIRKER, Clemens; HALTMEIER, Markus. A machine learning framework for customer purchase prediction in the non-contractual setting. eng. **European journal of operational research**, Elsevier B.V, v. 281, n. 3, p. 588–596, 2020. ISSN 0377-2217.

MILOŠEVIĆ, Miloš; ŽIVIĆ, Nenad; ANDJELKOVIĆ, Igor. Early churn prediction with personalized targeting in mobile social games. **Expert Systems with Applications**, Elsevier, v. 83, p. 326–332, 2017.

MUSTAFA, Nurulhuda; SOOK LING, Lew; ABDUL RAZAK, Siti Fatimah. Customer churn prediction for telecommunication industry: A Malaysian Case Study [version 1; peer review: 2 approved]. eng. **F1000 research**, Faculty of 1000 Ltd, England, v. 10, p. 1274–1274, 2021. ISSN 2046-1402.

NEW framework for Improving Random Forest Classification Accuracy. **International Journal of Emerging Trends in Engineering Research**, 2022.

OLIVEIRA, Vera Lúcia Miguéis. **Analytical customer relationship management in retailing supported by data mining techniques**. 2012. PhD thesis – University of Porto (Portugal).

PAWAR, Atish. Semantic similarity between words and sentences using lexical database and word embeddings. In.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PENG, Yifan. A Survey of Kernel Methods in Relation Extraction. In.

PERIÁÑEZ, África; SAAS, Alain; GUITART, Anna; MAGNE, Colin. Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles. eng. In: ARXIV.ORG. Ithaca: Cornell University Library, arXiv.org, 2017.

PERIŠIĆ, Ana; PAHOR, Marko. RFM-LIR Feature Framework for Churn Prediction in the Mobile Games Market. **IEEE Transactions on Games**, v. 14, n. 2, p. 126–137, 2022.

RADHIKA, K; VALARMATHY, S; SELVARASU, S; BASHKARAN, K; SRINIVASAN, C. Predictive Road Sign Maintenance Using Random Forest Regression and IoT Data. In: IEEE. 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA). [S.l.: s.n.], 2023. P. 348–353.

RHEAULT, Ludovic; COCHRANE, Christopher. Word Embeddings for the Estimation of Ideological Placement in Parliamentary Corpora. In.

SANTOS, Luis Gustavo Moneda dos. Domain generalization, invariance and the Time Robust Forest. In.

SIERS, Michael J.; ISLAM, Md. Zahidul. Class Imbalance and Cost-Sensitive Decision Trees. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, v. 15, p. 1–31, 2020.

SOBREIRO, Pedro; MARTINHO, Domingos Dos Santos; ALONSO, Jose G.; BERROCAL, Javier. A SLR on Customer Dropout Prediction. eng. **IEEE access**, IEEE, Piscataway, v. 10, p. 14529–14547, 2022. ISSN 2169-3536.

SOUZA, Alexandre Renato Rodrigues de; FERREIRA, Fabrício Neitzke; LAMBRECHT, Rodrigo Blanke; REICHOW, Leonardo Costa; SANTOS, Helida Salles; REISER, Renata Hax Sander; YAMIN, Adenauer Correa. Mortality Risk Evaluation: A Proposal for Intensive Care Units Patients Exploring Machine Learning Methods. In: INTELLIGENT Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28 – December 1, 2022, Proceedings, Part I. Campinas, Brazil: Springer-Verlag, 2022. P. 1–14.

SUH, Youngjung. Machine learning based customer churn prediction in home appliance rental business. **Journal of big Data**, Springer, v. 10, n. 1, p. 41, 2023.

TORREGROSA, Javier; BELLO-ORGAZ, Gema; MARTÍNEZ-CÁMARA, Eugenio; SER, Javier Del; CAMACHO, David. A survey on extremism analysis using natural language processing: definitions, literature review, trends and challenges. **Journal of Ambient Intelligence and Humanized Computing**, Springer, v. 14, n. 8, p. 9869–9905, 2023.

TRAN, Hoang; LE, Ngoc; NGUYEN, Van-Ho. CUSTOMER CHURN PREDICTION IN THE BANKING SECTOR USING MACHINE LEARNING-BASED CLASSIFICATION MODELS. **Interdisciplinary Journal of Information, Knowledge & Management**, v. 18, 2023.

WEI, Mengxi; HE, Yifan; ZHANG, Qiong. **Robust Layout-aware IE for Visually Rich Documents with Pre-trained Language Models**. [S.l.: s.n.], 2020. arXiv: 2005.11017 [cs.CL].

WEI, Jo-Ting; LIN, Shih-Yen; WU, Hsin-Hung. A review of the application of RFM model. **African Journal of Business Management**, v. 4, p. 4199–4206, 2010.

WENGER, Michael. **Strategic business models in the online food delivery industry - detailed analysis of the -order and delivery- business model**. [S.l.: s.n.], Dec. 2021. Available from: http://hdl.handle.net/10362/140006.

WU, Shengqiong; FEI, Hao; QU, Leigang; JI, Wei; CHUA, Tat-Seng. Next-gpt: Any-to-any multimodal llm. **arXiv preprint arXiv:2309.05519**, 2023.

WU, Xing; LI, Pan; ZHAO, Ming; LIU, Ying; CRESPO, Rubén González; HERRERA-VIEDMA, Enrique. Customer churn prediction for web browsers. **Expert Systems with Applications**, v. 209, p. 118177, 2022. ISSN 0957-4174.

XIAHOU, Xiancheng; HARADA, Yoshio. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. **Journal of Theoretical and Applied Electronic Commerce Research**, v. 17, n. 2, p. 458–475, 2022. ISSN 0718-1876.

YIN, Deyun; WU, Zhao; YOKOTA, Kazuki; MATSUMOTO, Kuniko; SHIBAYAMA, Sotaro. Identify novel elements of knowledge with word embedding. **PLOS ONE**, Public Library of Science, v. 18, n. 6, p. 1–16, June 2023.

ZHONG, Junmei; LI, William. Predicting Customer Churn in the Telecommunication Industry by Analyzing Phone Call Transcripts with Convolutional Neural Networks. In: PROCEEDINGS of the 2019 3rd International Conference on Innovation in Artificial Intelligence. Suzhou, China: Association for Computing Machinery, 2019. (ICIAI 2019), p. 55–59.

Appendix

APPENDIX A – DESCRIPTION OF THE FEATURES PRESENT IN THE DATASETS

Figure A.1 – Description of the Features from Food Delivery Dataset

Name	Description	Distribution of Values (V)	Distinct V.
orderQuantity (float)	Quantity of items ordered in a transaction	(0-49: 2680), (50-99: 2634), (100-199: 735), (>200: 258)	128
areaName (object)	Name of the area associated with the order	(Hub 1: 4411), (Hub 2: 2838)	2
OrderDinner (bool)	Indicates if the order includes dinner items	(False: 5794), (True: 1455)	2
OrderLunch (bool)	Indicates if the order includes lunch items	(True: 6920), (False: 329)	2
dessertQuantity (float)	Quantity of desserts in the order	(0-9: 7547), (10-19: 43), (20-29: 22), (>30-39: 8)	33
discountQuantity (float)	Quantity of discount applied in the transaction	(0-9: 6319), (10-19: 109), (20-29: 100), (30-39: 42), (40-49: 52), (>50: 59)	136
drinkQuantity (float)	Quantity of drinks in the order	(0-9: 6729), (10-19: 119), (20-29: 42), (30-39: 20), (40-49: 13), (>50: 7)	49
plateQuantity (float)	Quantity of main dishes in the order	(0-9: 4322), (10-19: 2590), (20-29: 907), (30-39: 345), (40-49: 242), (>50: 161)	164
noplateQuantity (float)	Quantity of items ordered without a main dish	(0-9: 6627), (10-19: 1542), (20-29: 637), (30-39: 176), (40-49: 74), (>50: 17)	103
totalValue (float)	Total value of the transaction	(0-19: 7770), (20-39: 3788), (40-59: 3188), (>60: 1397)	629
totalValueAVG (float)	Average total value across transactions	(0-20: 1545), (21-40: 3589), (41-60: 1463), (>61: 223)	64
discountTotal (float)	Total value of applied discounts	(0-20: 3941), (21-40: 4287), (41-60: 2234), (>61: 722)	449
discountTotalAVG (float)	Average value of applied discounts	(0-500: 6433), (501-1000: 1291),(1001-1500: 190), (>1501: 27)	106
email_ok (bool)	Indicates permission to send marketing email	(True: 4661), (False: 2588)	2
language (object)	Language associated with the transaction	(pt: 6200), (en: 1049)	2
sms_ok (bool)	Indicates permission to send marketing SMS	(True: 3781), (False: 3468)	2
plan (bool)	Indicates whether you have a loyalty plan applied	False: 6554), (True: 695)	2
promocode (bool)	Indicates if a promotional code is applied	(False: 6915), (True: 334)	2
gender (object)	Customer gender	(F: 2956), (0: 2252), (M: 2041)	3
reviews (float)	Reviews associated with the transaction	(0: 5058), (5: 1102), (4: 757), (3: 217), (2: 75), (1: 40)	6
FirstOrder (bool)	Indicates if it's the customer's first order	(False: 7073), (True: 176)	2
DaysLastOrder (float)	Number of days since the last order	(0-50: 2255), (51-100: 2075), (101-150: 1283), (>151: 453)	132
churn (bool)	Indicates if the customer has churned	(False: 3628), (True: 3621)	2
RFM_segment (object)	RFM segment associated with the customer	(Switchers: 3586), (Not_Fan: 1927), (Loyal: 1439), (Champions: 297)	4

Figure A.2 – Description of the Features from e-commerce Dataset

Name	Description	Distribution of Values (V)	Distinctive V.
sum_orderid (float)	Sum of order IDs	(1.0, 2.0: 23815), (3.0, 4.0: 79), (5.0, 6.0: 5), (7.0, 17.0: 3)	8
sum_price (float)	Sum of prices for orders	(0.0, 1719.0: 13277),(1720.0, 5089.0: 5332),, (5090.0, 12796.0: 3322),(12799.0, 459990.0: 1969)	3786
avg_review_score (float)	Average review score for orders	(0.0, 23.0: 2728),(24.0, 33.0: 1732),(34.0, 42.0: 4484),,(43.0, 50.0: 14958)	32
avg_payment_installments (float)	Average number of payment installments	(0.0, 28.0: 15045),(29.0, 54.0: 5226),(55.0, 88.0: 2539),,(90.0, 240.0: 1092)	72
last_order_status (object)	Last order status	(delivered: 23493), (shipped: 196), (canceled: 102), (invoiced: 59), (unavailable: 45), (processing: 7)	6
last_payment_type (object)	Last payment type	(credit_card: 17916), (boleto: 4419), (voucher: 829), (debit_card: 737), (not_defined: 1)	5
cust_custom (float)	Custom customer attribute	(-6.0, 13.0: 5738),(14.0, 32.0: 14600),,(33.0, 51.0: 3382),(52.0, 72.0: 180)	78
avg_freight_value (float)	Average freight value for orders	(0.0, 1473.0: 9612),(1474.0, 2674.0: 10492), (2675.0, 4709.0: 2346),(4711.0, 32146.0: 1448)	4227
sum_payment_value (float)	Sum of payment values for orders	(187.0, 6917.0: 8543),(6918.0, 14347.0: 6953), (14348.0, 27722.0: 4895),(27736.0, 1945704.0: 3509)	12550
last_product_category_name (object)	Last product category name	(beleza_saude: 2718), (cama_mesa_banho: 2117), (...)	71
avg_review_score_item (float)	Average review score for orders (Items)	(0.0, 10.0: 2110),(15.0, 20.0: 623),,(25.0, 30.0: 1733), (35.0, 40.0: 4480)	10
last_seller_state (object)	State of the last seller in the order	(SP: 17001), (PR: 1659, (...))	20
last_customer_state (object)	State of the last customer in the order	(SP: 11126, RJ: 2774), (...))	27
avg_product_height_cm (float)	Average product height in centimeters	(0.0, 130.0: 12789),(133.0, 235.0: 7039),(238.0, 463.0: 3313),(465.0, 1040.0: 743)	237
DaysLastOrder (float)	Number of days since the last order	(0.0, 32.0: 6251),(33.0, 60.0: 6109),,(62.0, 89.0: 5483),,(90.0, 117.0: 5533)	114
churn (bool)	Indicates if the customer has churned	(False: 12360), (True: 11542)	2
RFM_segment (object)	RFM segment associated with the customer	(Switchers: 13948), (Not_Fan: 9447), (Champions: 339),, (Loyal: 168)	4

Figure A.3 – Description of the Features from Gambling Dataset

Name	Description	Distribution of Values (V)	Distinctive V.
RG_case (float)	Registration case indicator	(1: 950), (0: 851)	2
Missing_Daily_Transactions (float)	Indicator for missing daily transactions	(0: 1801)	1
CountryName (object)	Country of residence	(Germany: 885), (Poland: 126), (...)	37
LanguageName (object)	Preferred language of the user	(German: 1030), (Polish: 128), (...)	24
Gender (object)	Gender of the user	(M: 1660), (F: 141)	2
YearofBirth (float)	Year of birth of the user	(1920 - 1949: 34), (1950, 1961: 134), (1962, 1973: 443), (1974, 1985: 1021)	51
age_at_registration (float)	Age of the user at the time of registration	(16, 28: 1027), (29, 41: 570), (42, 54: 166),(55, 84: 38)	52
sum_stakes_fixedodds (float)	Total stakes in fixed odds betting	(0-479: 735),(480 -2114: 388)(2145-7896: 341) (7907-224216: 334)	1331
sum_bets_fixedodds (float)	Total bets placed in fixed odds betting	(0-214 : 1026),(215-569: 319),(572, 1532: 244), (1547, 48823: 211)	829
bettingdays_fixedodds (float)	Number of days engaged in fixed odds betting	(0, 109: 1199), (110, 224: 292), (225-419: 182), (422-1711: 126)	438
duration_fixedodds (float)	Duration of engagement in fixed odds betting	(0-626: 579), (628-1203: 388), (1205-1678: 469), (1680-3619: 362)	1183
frequency_fixedodds (float)	Frequency of fixed odds betting	(0: 1609), (1: 192)	2
bets_per_day_fixedodds (float)	Average number of bets per day in fixed odds	(0-11: 1674), (12-23: 87), (24-39: 23), (43-83: 14)	51
euros_per_bet_fixedodds (float)	Average amount in euros per bet in fixed odds	(0-33: 1590), (34-69: 121), (71-126: 51), (129-485: 37)	138
net_loss_fixedodds (float)	Net loss in fixed odds betting	(-7898-75: 778), (76-467: 467), (468-1705: 289), (1710-30693: 264)	1027
percent_lost_fixedodds (float)	Percentage of total stakes lost in fixed odds	(0: 1379), (1: 365), (-1: 49), (-2: 5), (-3: 2), (-6: 1)	6
churn (float)	Indicates if the customer has churned	(True: 990, False: 811)	2
DaysLastOrder (float)	Number of days since the last order	(-30-85: 1013),(86-215: 383),(216-377: 251), (380-703: 152)	454
RFM_segment (object)	RFM segment associated with the customer	(Switchers: 779), (Not_Fan: 483), (Loyal: 374), (Champions: 165)	4
(...)	(...)	(...)	(...)

APPENDIX B – HYPERPARAMETER VALUES

Table 13 – Values of hyperparameters for each model

Multilayer Perceptron	
Number of hidden layers	3, 5, 10, 16, 32, 64
Alphas (regularization strength)	0.001, 0.01, 0.1, 1
Solvers (optimization algorithm)	'sgd', 'adam'
Activations (hidden layers)	'relu', 'tanh'
Support Vector Machine	
Kernels (used for transformation)	'linear', 'rbf', 'poly'
Cs (regularization)	0.1, 1, 10
Gammas (kernel coefficient)	'auto', 'scale'
Random Forest	
N estimators (number of trees)	10, 50, 100
Criterion (split quality measure)	'gini', 'entropy'
Max depth (trees depth maximum)	None, 10, 20
Random state	None, 42
Decision Tree	
Criterion (split quality measure)	'gini', 'entropy'
Max depth (trees depth maximum)	None, 10, 20
Min samples split	2, 5, 10
Min samples leaf	1, 2, 4
BERT	
Number of hidden layers	1;5;10;15;16;17;18,19;20;25;30
Learning rate	0.01;0.001;0.0001
Batch size	16, 32
RoBERTa	
Number of hidden layers	1;5;10;15;16;17;18,19;20;25;30
Learning rate	0.01;0.001;0.0001
Batch size	16, 32

APPENDIX C – BEST COMBINATIONS OF HYPERPARAMETERS

Table 14 – Best combinations of hyperparameter values

	Gambling	**Food Delivery**	**E-commerce**
RF	criterion: 'entropy' max depth: None n estimators: 100 random state: None	criterion: 'entropy' max depth: 10 n estimators: 50 random state: None	criterion: 'entropy' max depth: None n estimators: 100 random state: None
SVM	kernel': 'rbf' C: 1 gamma: 'scale'	kernel: poly C: 10 gamma: 'scale'	kernel: 'linear' C: 10 gamma: 'auto'
DT	criterion: 'entropy' max_depth: 10 min_samples_split: 2 min_samples_leaf: 4 splitter: 'best'	criterion: 'entropy' max_depth: 10 min_samples_split: 10 min_samples_leaf: 4 splitter: 'best'	criterion: gini max_depth: 10 min_samples_leaf: 1 min_samples_split: 2 splitter: 'best'
MLP	hidden layer sizes: 5 alpha: 0.001 solver: 'adam' activation': 'relu'	hidden layer sizes: 5 alpha: 0.001 solver: 'adam' activation: 'relu'	hidden layer sizes: 5 alpha: 0.001 solver: 'adam' activation': 'relu'