E-CommVision: Forecasting Customer Satisfaction with Interpretable Predictive Models and Advanced Sentiment Analysis

A Project / Dissertation as a Course requirement for
Master of Science
(Data Science and Computing)

AKULA VENKATA SATYA SAI GOPINADH
22228

SRI SATHYA SAI INSTITUTE OF HIGHER LEARNING
(Deemed to be University)

Department of Mathematics and Computer Science
Muddenahalli Campus

March 2024

SRI SATHYA SAI INSTITUTE OF HIGHER LEARNING
(Deemed to be University)

Dept. of **Mathematics & Computer Science**
Muddenahalli Campus

CERTIFICATE

This is to certify that this Project / Dissertation titled **E-CommVision: Forecasting Customer Satisfaction with Interpretable Predictive Models and Advanced Sentiment Analysis** submitted by **Akula Venkata Satya Sai Gopinadh**, Regd No. **22228**, Department of Mathematics and Computer Science, Muddenahalli Campus is a bonafide record of the original work done under our supervision as a Course requirement for the Degree of Master of Science Data Science and Computing.

Countersigned by

Sri. SVSN Sarma
Supervisor

Dr. (Ms.) Y Lakshmi Naidu
Head of the Department

Place: Muddenahalli
Date: March 25, 2024

DECLARATION

The Project / Dissertation titled **E-CommVision: Forecasting Customer Satisfaction with Interpretable Predictive Models and Advanced Sentiment Analysis** was carried out by me under the supervision of Sri. SVSN Sarma, Department of Mathematics and Computer Science, Muddenahalli as a Course requirement for the Degree of Masters of Science in Data Science and Computing and has not formed the basis for the award of any degree, diploma or any other such title by this or any other University.

Place: Muddenahalli
Date: March 25, 2024

A.V.S. Sai Gopinadh

Akula Venkata Satya Sai Gopinadh
22228
MSc (Data Science and Computing)

With utmost Love and Gratitude, I dedicate this project at the Divine Lotus Feet of **Bhagawan Sri Sathya Sai Baba Varu** *for guiding me directly and indirectly, who has been the driving force throughout my life.*

Acknowledgments

I would like to thank my advisor, Sri. SVSN Sarma for providing me with his guidance and motivation throughout the completion of this project. He has always been an inspiration in terms of work ethic, dedication, and discipline. Always providing expert comments, analysis, and support from his in-depth knowledge of Machine Learning, he has shaped this project into what it is today.

I would like to convey my special thanks to Sri. Gudipudi Radhesyam, Technical Architect - AI/ML, Charter Global Technologies, Hyderabad, for his invaluable guidance and mentorship during this project. His expert insights and industry perspective were instrumental in shaping our approach and bringing a practical dimension to our academic endeavors.

I would like to extend a heartfelt token of gratitude to the Director of our esteemed campus, the Associate Head, and all the revered faculty members of the Department of Mathematics and Computer Science. Their unwavering guidance, steadfast support, and ever-watchful eyes have not only corrected my course but have also served as a beacon of inspiration throughout this journey.

I would like to acknowledge my mother and sister forever supporting me and boosting my morale through the completion of this project and also through every phase of my life. They are the reason behind the person I am now.

I extend my heartfelt appreciation to my friends and seniors for their unwavering support and encouragement during this project. Their valuable insights, feedback, and moral support have been instrumental in overcoming challenges and achieving milestones. I am deeply grateful for their friendship, which has enriched both this project and my life in countless ways. Their enduring bonds and constant presence have been a source of strength through both professional and personal endeavors.

I would like to present my heartful gratitude to Sri Sathya Sai Institute of Higher Learning for giving me unthinkable experiences and memories for life. Especially the Muddenahalli Campus, for its exquisite beauty and serene environment which motivated and encouraged me to keep working on the project.

Abstract

In this dissertation, we delve into the application of machine learning (ML) algorithms to forecast customer satisfaction scores, recognizing their pivotal role in shaping business success and cultivating brand loyalty. Our analysis unveils the prevalence of Excellent ratings among sales orders, underscoring the importance of comprehending and elevating customer satisfaction levels. Employing advanced ML techniques like Random Forest (RF), XGBoost (XGB), and Decision Tree (DT), we investigate the predictive prowess of these models in discerning customer satisfaction trends. Furthermore, by integrating sentiment analysis of review comments, we deepen our understanding of customer sentiment, augmenting our ability to accurately predict satisfaction levels. Our findings not only underscore the paramount importance of prioritizing customer satisfaction in e-commerce realms but also highlight the value of embracing interpretability techniques to elucidate model predictions and enhance decision-making processes.

The Streamlit web application developed for this project, dubbed "E-comm Vision," serves as a comprehensive tool for visualizing and analyzing e-commerce data. Leveraging Streamlit's intuitive interface, users can seamlessly navigate through various sections, including data preview, exploratory data analysis, sentiment analysis, predictive modeling, and more. Behind the scenes, MySQL plays a crucial role in data management, facilitating efficient storage, retrieval, and manipulation of the dataset. Through seamless integration with Streamlit, the application offers users a seamless experience in accessing and analyzing e-commerce data, empowering them to derive actionable insights and make informed decisions to drive business growth and enhance customer satisfaction.

Table of contents

Acknowledgments **4**
Abstract **6**
Table of contents **8**
1. Introduction **10**
1.1 Motivation 10
1.2 Problem Statement 10
1.3 Objectives 11
1.4 Dissertation Structure 12
2. Literature Review **14**
2.1 Machine Learning Approach to Predict E-commerce Customer Satisfaction Score 14
2.2. Sentiment Analysis on Brazilian Portuguese User Reviews 15
2.3 Optimising e-commerce customer satisfaction with machine learning 16
2.4 I can get no e-satisfaction 16
2.5 BERT and it's variant DistilBERT 17
2.5.1BERT 17
2.5.2 DistilBERT 17
3. Data **18**
3.1 Data Collection 18
3.2 Data Cleaning 19
4. Methodology **20**
5. Version - 1 **21**
5.1 Exploratory Data Analysis (EDA) 21
5.1.1 Choropleth Map 21
5.1.2 Bar Plot 22
5.1.3 Pie Chart 23
5.1.4 Dashboard 24
5.2 Dataset Split 25
5.3 Imbalance Treatment 25
5.3.1 SMOTE 26
5.3.2 Random Under Sampler 26
5.4 Feature Engineering 26
5.5 Feature Selection 28
5.5.1 Gini Impurity Index 28
5.6 Proposed Models 29
5.7 Interpretability 30
5.7.1 Accumulated Local Effects (ALE) 31
5.7.2 Local Interpretable Model-agnostic Explanations(LIME) 31
5.8 Model Training and Tuning 32

5.8.1 Grid SearchCV Technique 33
5.9 Evaluation and Metrics 33
5.9.1 Precision 33
5.9.2 Recall 34
5.9.3 F1-score 34
6. Version - 2 35
6.1 Dataset 35
6.2 Exploratory Data Analysis(EDA) 35
6.2.1 Choropleth Map 35
6.2.2 Bar Plot 36
6.2.3 Pie Chart 37
6.2.4 Dashboard 37
6.3 Dataset Split and Imbalance Treatment 38
6.4 Sentiment Analysis 39
6.5 Feature Engineering 39
6.6 Feature Selection 42
6.7 Model Training, Tuning and Evaluation 43
7. Experimental Details 44
8. Results and Observations 46
8.1Version - 1 Results and Observations 46
8.1.1 XGB Model 46
8.1.2 RF Model 47
8.1.3 DT Model 48
8.2 Version - 2 Results and Observations 49
8.2.1 XGB model 49
8.2.2 RF Model 50
8.2.3 DT Model 52
8.3 Comparative Study 53
9. Full Stack Development 55
9.1 Web Application Architecture 55
9.2 Data Management using MySQL 59
10. Conclusion and Future Work 61
10.1 Conclusion 61
10.2 Future Work 61
11. References 63

1. Introduction

1.1 Motivation

A famous quote by Jerry Gregoire states, "Customer experience is the next competitive battleground."

As online retail continues to witness exponential growth, understanding and effectively addressing customer preferences and expectations have become paramount for businesses seeking to thrive in this competitive landscape.Our approach to predicting customer satisfaction in e-commerce transcends conventional methods by integrating interpretable machine learning and advanced sentiment analysis.Our emphasis lies in ensuring transparency in model outputs, fostering trust among stakeholders. Through interpretable machine learning, we unveil the underlying factors influencing each prediction, empowering sellers to make informed decisions and enhance customer satisfaction. Moreover, our utilization of advanced sentiment analysis captures intricate details from customer reviews, offering deeper insights into preferences and concerns. This comprehensive approach not only boosts prediction accuracy but also enables proactive measures to address customer needs, fostering loyalty and satisfaction.

In the world of online shopping, keeping customers happy is key to making a business thrive. When customers are satisfied, they're more likely to stick with a brand, come back for more, and tell their friends about it. Think of it like this: a happy customer isn't just a one-time sale—they're like a cheerleader for the brand, spreading good vibes and bringing in more business. By focusing on customer satisfaction, businesses build trust and credibility, which sets the stage for long-term success. From making sure products are top-notch to delivering great customer service, every little detail adds up to keep customers smiling and the business thriving.

Moreover, sellers may pinpoint areas for improvement and implement the required changes to improve the buying experience by actively seeking out and incorporating client feedback. In the cutthroat world of e-commerce, merchants can build a devoted clientele and achieve long-term success by paying attention to what their consumers want and always working to satisfy those requirements and preferences.

1.2 Problem Statement

In the context of online goods delivery, ensuring customer satisfaction is paramount for sustained business success. However, accurately predicting customer satisfaction poses a significant challenge, given the diverse factors influencing buyer perceptions and experiences. To address this challenge, this research focuses on developing interpretable predictive models and leveraging advanced sentiment analysis techniques. Our goal is to identify the fundamental factors that influence customer happiness by utilizing interpretable predictive models. This will

enable sellers to efficiently satisfy consumer expectations and improve the quality of their services. Additionally, integrating advanced sentiment analysis allows us to delve deeper into customer sentiments expressed through reviews and feedback, enabling a more nuanced understanding of their preferences and concerns. Through this combined approach, we seek to develop a comprehensive framework that not only accurately predicts customer satisfaction in online goods delivery but also facilitates informed decision-making and proactive customer engagement strategies for businesses operating in the e-commerce domain.

"***Customer Satisfaction Prediction in online goods delivery through Interpretable Predictive Models and Advanced Sentiment Analysis***"

1.3 Objectives

The principal aim of this study is to close the divide between research in customer satisfaction prediction, specifically concentrating on Sentiment Analysis using DistilBERT and interpretable machine learning models, and its practical implementation for end-users. Our overarching goal is to convert our research discoveries into a user-friendly application that provides accurate predictions while ensuring transparency in decision-making processes.

Specific Objectives:

- **Create an Intuitive Web Application:** Develop and deploy a user-friendly web platform that offers seamless access to customer satisfaction predictions.

- **Integrate DistilBERT and Interpretable Models**: Incorporate DistilBERT, a specialized BERT variant, into interpretable machine learning models to strike a balance between performance and transparency.

- **Ensure User-Friendly Access:** Develop an application that offers intuitive usability, catering to users with diverse levels of expertise and removing the barriers typically associated with intricate predictive models.

- **Provide Transparent Insights:** Develop features within the application that offer users clear insights into how predictions are generated, enhancing the transparency of the predictive model.

- **Contribute to the Field:** Beyond application development, contribute to the broader research community by sharing insights gained from real-world application usage and refining the model based on practical considerations.

These objectives collectively strive to translate our research into an impactful, user-friendly, and transparent tool that empowers e-commerce stakeholders to make informed decisions in the

dynamic landscape of online retail, particularly in enhancing customer satisfaction.

1.4 Dissertation Structure

This dissertation consists of eleven (11) chapters in total.

Chapter 1 introduces the dissertation, discussing the motivation behind the research, the problem statement addressed, and the objectives of the study. It provides an outline of the dissertation's structure, guiding the reader through the subsequent chapters.

Chapter 2 delves into the literature review, exploring previous research related to machine learning approaches in e-commerce, sentiment analysis techniques, and optimization strategies. It provides a comprehensive overview of existing knowledge in the field, laying the groundwork for the study's methodology.

The primary focus is data in Chapter 3, which also explains the process of gathering and cleansing data. It discusses the sources of the data used in the study, the preprocessing methods performed on the data, and any challenges encountered during this phase of the inquiry.

Chapter 4 outlines the methodology employed for the study, including feature engineering techniques, feature selection methods, and the development of machine learning models. It provides a detailed description of the steps taken to prepare and analyze the data.

Chapter 5 presents the analysis and results of the first version of the study, highlighting key findings and insights obtained from the data. It talks about any findings from the analysis process as well as how well machine learning models that were trained on the dataset performed.

Chapter 6 focuses on the second version of the study, providing details on the dataset used, exploratory data analysis findings, and the development of machine learning models specific to this version. It compares the results obtained with those from the first version, identifying any differences or improvements.

Chapter 7 discusses the experimental details of the study, including the experimental setup, data preprocessing techniques, and model training procedures. It provides insights into the methodology used to conduct experiments and analyze results.

Chapter 8 presents the results and observations derived from both versions of the study, discussing the performance of machine learning models and any insights gained from the analysis. It highlights key findings and discusses their implications for the research.

Chapter 9 delves into the full-stack development of the "E-comm Vision" application, detailing the web architecture and data management techniques employed. It provides an overview of the application's features and functionalities, showcasing its utility for e-commerce businesses.

Chapter 10 concludes the dissertation, summarizing the key findings of the study and outlining potential areas for future research and improvement. It reflects on the research objectives and discusses the broader implications of the study's findings.

Chapter 11 lists all references cited throughout the dissertation, providing readers with a comprehensive list of sources for further exploration and study.

2. Literature Review

In this section, we provide an extensive review of the current literature on customer satisfaction prediction and user review sentiment analysis. Our focus is on identifying gaps in the existing research landscape that our study seeks to fill.

Customer satisfaction lies at the heart of e-commerce success, serving as a crucial metric that directly impacts brand loyalty, customer retention, and overall business growth. In the digital realm, where consumers have myriad options at their fingertips, ensuring a positive and seamless shopping experience is paramount for businesses. Customer satisfaction prediction, a burgeoning field in e-commerce analytics, leverages advanced technologies to anticipate and understand customer sentiments and preferences. Traditional approaches often relied on basic statistical models, but recent advancements have seen the integration of machine learning algorithms, natural language processing (NLP), and sentiment analysis techniques. Technologies such as decision trees, support vector machines (SVM), logistic regression, and neural networks are commonly employed for predictive modeling. Additionally, interpretable machine learning models and advanced sentiment analysis, including techniques like DistilBERT and LSTM (Long Short-Term Memory), have gained prominence for their ability to provide transparent insights into customer feedback and sentiments. These technologies collectively enable businesses to proactively identify areas for improvement, personalize customer experiences, and ultimately drive higher levels of satisfaction and loyalty in the competitive e-commerce landscape.

2.1 Machine Learning Approach to Predict E-commerce Customer Satisfaction Score

Authored by Pongthanin Wangkiat and Chantri Polprasert, the study delves into the realm of e-commerce with a focused aim on utilizing sophisticated machine learning techniques for predicting customer satisfaction scores. Categorizing satisfaction levels into distinct classes ranging from Low to Excellent, the research endeavors to unravel the pivotal factors driving customer contentment. To achieve this objective, three machine learning models - Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbor (K-NN) - were meticulously deployed to forecast these scores [1].

The results unequivocally demonstrated that the Random Forest model outperformed its counterparts across all satisfaction classes, exhibiting superior predictive capabilities over LR and K-NN models. Through meticulous feature engineering methodologies and rigorous model optimization, the RF model achieved commendable precision, recall, and F1 scores for each

satisfaction category. Particularly noteworthy were the discernible contributions of key features such as product ratings, delivery performance, and total shipping duration in accurately discerning customer satisfaction levels.

The research findings underscore the profound potential of machine learning models, particularly the Random Forest approach, in elevating e-commerce strategies by providing nuanced insights into customer satisfaction dynamics. By discerning and prioritizing influential determinants, online retailers stand poised to fine-tune their service offerings, logistics operations, and customer support mechanisms to effectively align with customer expectations. The study's adeptness in predicting customer satisfaction scores accentuates the transformative impact of advanced analytics in optimizing e-commerce workflows and augmenting overall customer experiences.

2.2. Sentiment Analysis on Brazilian Portuguese User Reviews

In their comprehensive study [2], Frederico Dias Souza and Baptista de Oliveira e Souza Filho delve into the intricate landscape of sentiment analysis on Brazilian Portuguese user reviews. Their research sheds light on the formidable challenges and strategic approaches inherent in analyzing sentiment within Brazilian Portuguese text data. With a keen focus on the predictive efficacy of diverse document embedding techniques, the study underscores the pivotal role of employing a unified dataset with meticulously defined partitions for training, testing, and validation purposes. By rigorously evaluating various document embedding methodologies, including tf-idf and average word embeddings, the authors discern the nuanced impact of these techniques on the overall performance of sentiment classification models.

Furthermore, the research illuminates the disparities in dataset sizes and lexical diversity between English and Portuguese corpora, offering critical insights into the unique considerations required when applying sophisticated models to Portuguese text data. Addressing the delicate balance between model intricacy and performance, particularly concerning the tf-idf approach and the dimensionality of the vocabulary, the study delineates the intricacies of model selection and training strategies.

In summation, the study underscores the paramount importance of judiciously selecting document embedding techniques and refining model training methodologies to yield accurate sentiment analysis outcomes in Brazilian Portuguese user reviews. Its findings serve as a valuable compass for both researchers and practitioners navigating the realm of sentiment analysis within the Portuguese commercial textual domain, offering actionable insights to enhance the efficacy of sentiment analysis endeavors.

2.3 Optimising e-commerce customer satisfaction with machine learning

In their research, authors Ann-Nee Wong and Booma Poolan Marikannan explore the optimization of e-commerce customer satisfaction through the application of machine learning techniques. Drawing from a comprehensive 3-year dataset sourced from the Brazilian E-Commerce Public Dataset by Olist, their study aims to discern the pivotal drivers influencing customer satisfaction and construct a predictive model capable of classifying customer sentiment. The findings underscore the paramount importance of meeting delivery expectations and ensuring efficient order fulfillment in elevating customer satisfaction levels within the e-commerce domain. By employing machine learning algorithms like Random Forest, the study illuminates the potential for personalized recommendations and tailored services to enhance the overall shopping experience amidst the competitive e-commerce landscape.

Furthermore, the research underscores the indispensability of data-driven approaches in comprehending customer behavior and preferences, thereby catalyzing business growth and nurturing enduring customer relationships. The study's focus on customer insights highlights the imperative nature of understanding factors contributing to satisfaction and underscores the significance of predictive models in informing strategic decision-making for a diverse array of sellers. After a thorough analysis of various machine learning algorithms for classification and skillful preprocessing methods to handle issues such as skewness and imbalance in the data, the research comes to the conclusion that Random Forest is the best option, offering fast processing times along with high accuracy. Notably, the study identifies meeting the estimated delivery date and the duration of order delivery as the two primary factors significantly impacting customer satisfaction levels.

2.4 I can get no e-satisfaction

This study explores customer satisfaction (CS) segmentation using a large dataset of around 270 thousand responses from a CS survey that was administered to 140 e-commerce businesses in a European country. Using a framework inspired by business analytics (BA), the study carefully examines and groups computer science (CS) parts in order to provide complex insights. It then provides examples of how these insights might be used in real-world scenarios, including how one e-commerce company used the identified CS segments to plan automated marketing campaigns for certain consumer categories. These initiatives range from targeted social media engagement tactics designed to amplify satisfaction levels among specific segments to incentivized discounts aimed at addressing the concerns of less satisfied customers. By amalgamating BA techniques with CS segmentation methodologies, the study not only enriches the existing literature on customer satisfaction and segmentation but also furnishes actionable insights with tangible implications for businesses operating in the e-commerce domain. Moreover, the derived insights serve as invaluable decision-making aids, empowering companies to deploy targeted advertising campaigns, optimize customer engagement initiatives, and glean

competitive benchmarks from analogous enterprises within their industry landscape. Through its multifaceted approach and real-world applications, this research offers a holistic understanding of CS segmentation dynamics and underscores its pivotal role in driving strategic business outcomes within the e-commerce ecosystem.

2.5 BERT and it's variant DistilBERT

2.5.1BERT

BERT (Bidirectional Encoder Representations from Transformers) and DistilBERT are state-of-the-art natural language processing (NLP) models developed by Google and Hugging Face, respectively. By implementing a bidirectional pre-training strategy for transformer models, BERT transformed natural language processing tasks by enabling these models to extract contextual information from both the left and right contexts within a sentence. The performance of numerous NLP tasks, including sentiment analysis, text categorization, and question answering, was much enhanced by this bidirectional capacity.

2.5.2 DistilBERT

DistilBERT, on the other hand, is a distilled version of BERT that offers similar performance with significantly fewer parameters, making it more lightweight and computationally efficient. DistilBERT achieves this by employing knowledge distillation techniques during the training process, where it learns to mimic the behavior of BERT while being trained on a smaller dataset.

Both BERT and DistilBERT have been widely adopted in various NLP applications due to their exceptional performance and versatility. They have been instrumental in advancing the field of NLP and have enabled significant breakthroughs in tasks such as sentiment analysis, named entity recognition, and text generation. Their ability to understand and generate human-like text has made them indispensable tools for researchers, developers, and data scientists working in the domain of natural language understanding.

3. Data

The datasets utilized in this study were sourced from Olist Shops, a leading Brazilian e-commerce company renowned as the largest department store in Brazilian markets. Olist Shops boasts a diverse range of products and services, catering to a broad customer base across various demographics and preferences. The sales historical dataset spans a significant timeframe from 2016 to 2018, capturing extensive transactional data and customer interactions over multiple years.

3.1 Data Collection

The datasets were generously made available through Kaggle [6], a popular platform for sharing datasets and conducting data science competitions, facilitating access for researchers and analysts to explore and derive insights from real-world e-commerce data. Leveraging this rich dataset offers a unique opportunity to delve into the dynamics of e-commerce operations, customer behavior patterns, and factors influencing customer satisfaction in a dynamic and evolving market landscape.

The datasets encompass a total of 52 columns and 100,000 records, offering a rich and comprehensive repository of e-commerce data for analysis. These datasets are meticulously compiled from nine individual files, each representing different aspects of the e-commerce ecosystem. Intricately merged based on their interrelationships, these files collectively form a unified dataset, providing a holistic view of various dimensions of e-commerce operations.

The compilation includes essential files such as

- products_dataset
- order_items_dataset
- orders_dataset
- sellers_dataset
- order_customers_dataset
- payments_dataset
- order_reviews_dataset
- geolocation_dataset
- product_category_name_translation dataset.

capturing crucial information about products, orders, and sellers involved in the e-commerce transactions and they are which contribute valuable contextual information and enhance the depth of analysis.

While the geolocation_dataset contains latitude and longitude data for Brazilian states, primarily utilized for an additional function of exploratory data analysis (EDA) is to offer geographical

insights into the distribution of e-commerce activity. However, it will not significantly influence subsequent phases of our analysis, as it does not directly align with our primary focus on customer satisfaction prediction.

On the other hand, the order_reviews_dataset emerges as a pivotal component for sentiment analysis purposes, offering insights into customer feedback and satisfaction levels. By leveraging this dataset, we aim to delve deeper into understanding customer sentiments and perceptions, thereby enriching our analysis and enabling more informed decision-making for e-commerce stakeholders.

The interconnectedness of datasets is visually represented in Fig.1, showcasing the relationship between various datasets through common features. This aids in understanding the intricate network of data and its relevance in analysis.

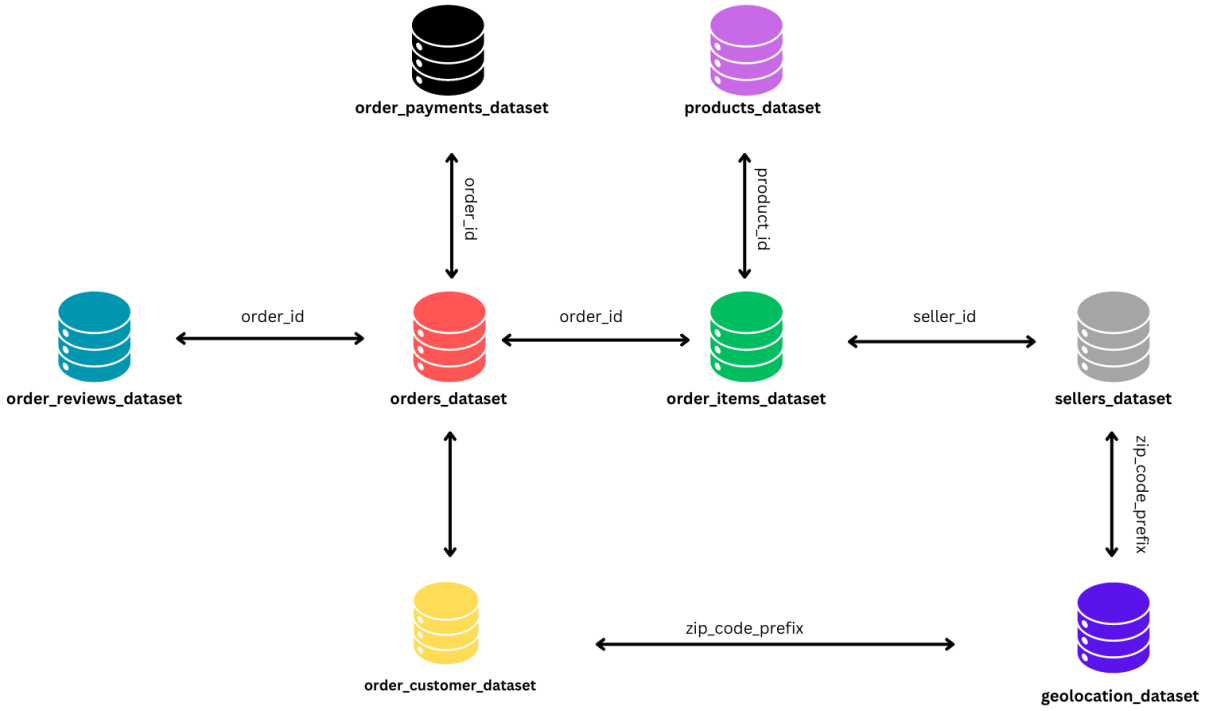


Fig: 1: Relationship between Datasets

3.2 Data Cleaning

Data cleaning is an important procedure in data analysis that involves detecting and correcting mistakes, inconsistencies, and missing values in a dataset. It ensures data accuracy, improves data quality, enhances analysis results, facilitates data integration, ensures compliance

with regulations, and saves time and resources. Overall, data cleaning is essential for ensuring reliable and meaningful insights from the data.

During the data preparation phase, a meticulous integration process was undertaken to merge information from various datasets seamlessly. This process commenced with the merging of the 'order_items_dataset' with the 'order_payments_dataset,' utilizing the common identifier 'order_id.' Subsequently, this intermediary dataset was further merged with the 'orders_dataset,' resulting in a consolidated dataset.

Following the merging process, the combined dataset underwent thorough cleansing procedures to eliminate any duplicate records and null values. This rigorous cleansing ensured the integrity and quality of the data.

Further integration steps involved merging the refined dataset with the 'customers_dataset' based on the shared identifier 'customer_id,' and with the 'products_dataset' based on 'product_id.' At each integration stage, careful consideration was given to removing records with null values to maintain data accuracy.

Before proceeding to subsequent analyses, a meticulous review was conducted to address any remaining records with null values. Finally, the dataset was merged with the 'product_category_name_translation' file based on the 'product_category_name.' A final review was performed to ensure the elimination of any lingering null values, resulting in a meticulously curated dataset comprising 94,445 records and 33 columns.

4. Methodology

The provided flowchart outlines a systematic process for selecting features within the machine learning workflow, which is essential for improving model performance and efficiency. It commences with data acquisition, where pertinent data is gathered to train the model. Following this, data preprocessing ensures data cleanliness by addressing inconsistencies or missing values. The next step, called exploratory data analysis (EDA), allows for a more thorough comprehension of the features of the dataset and any correlations between its variables.

The pivotal feature selection stage focuses on meticulously choosing a subset of features based on their relevance to the modeling task. To do this, a variety of approaches, including filter, wrapper, and embedding methods, are used to determine which properties are most important. After being chosen, these characteristics are incorporated into the model and adjusted to improve its functionality. In the end, testing the model on a different test dataset verifies that it is reliable and useful by ensuring that it can generalize to new data.

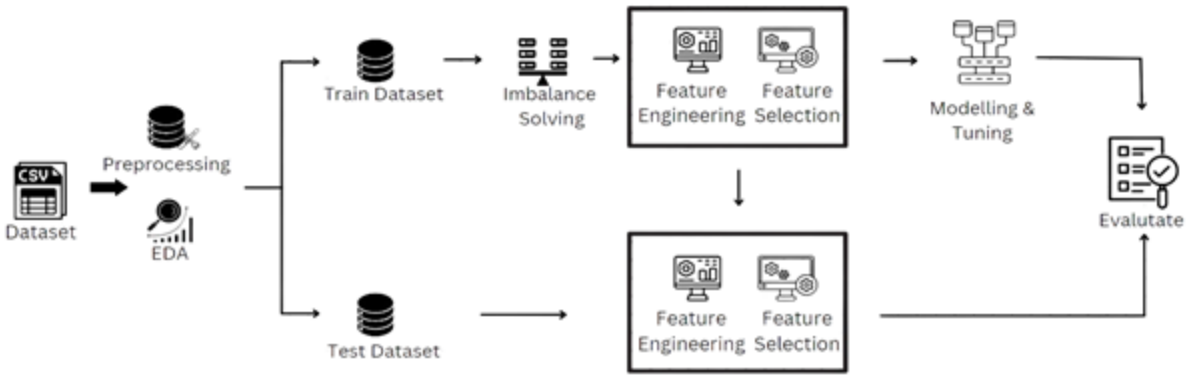


Fig.2: Methodology

5. Version - 1

Version 1 of the project utilized the dataset obtained after the meticulous data-cleaning process detailed in Section 3.2. This refined dataset, devoid of inconsistencies or missing values, served as the foundation for the initial phase of the project's development.

5.1 Exploratory Data Analysis (EDA)

A crucial phase in data analysis is called exploratory data analysis (EDA), which entails looking at and visualizing data sets to comprehend their salient features, find trends, and pinpoint correlations between variables.Prior to formal modeling and hypothesis testing, it functions as an exploratory study and offers important insights into the underlying structure of the data. EDA approaches include data visualization (box, scatter, and histogram plots), summary statistics, correlation analysis, and dimensionality reduction methods. By conducting EDA, data scientists can gain a deeper understanding of the data's distribution, outliers, missing values, and potential biases, which informs subsequent data preprocessing and modeling decisions. EDA plays a crucial role in hypothesis generation, feature selection, and model building, ultimately contributing to more accurate and robust analyses.

5.1.1 Choropleth Map

A choropleth map visually represents data variations across geographic regions through color gradients or patterns. In e-commerce order distribution in Brazil, it reveals regional disparities, with darker shades indicating higher order volumes in regions like the south and southeast. This information helps businesses tailor strategies and policymakers target interventions for economic development. Researchers benefit from insights into socio-economic dynamics, making the choropleth map essential for informed decision-making in e-commerce.

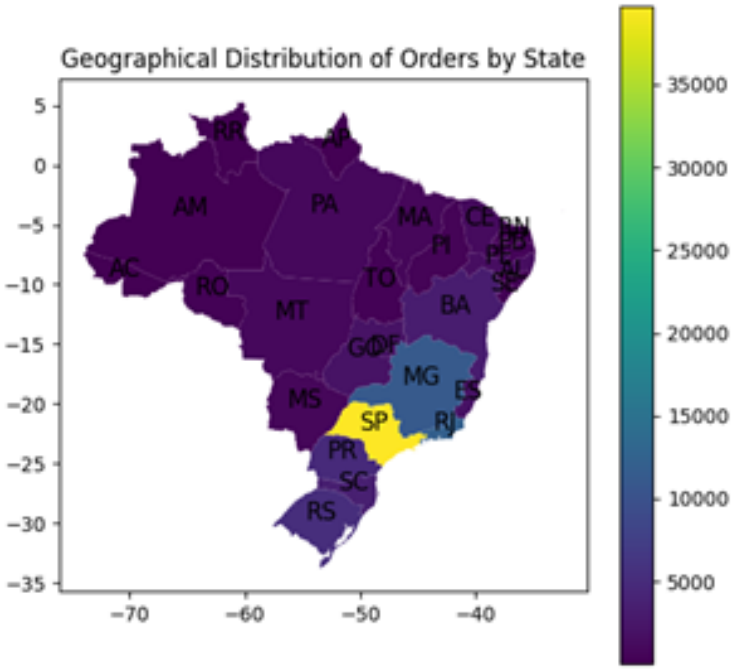


Fig.3: Geographical Distribution of Orders by State

The analysis of order distribution in Brazil from Fig.3 reveals significant regional variations, with clear disparities between different parts of the country. The south and southeast regions emerge as the dominant hubs for e-commerce activity, with states like São Paulo and Rio de Janeiro leading the pack. These regions benefit from larger populations, robust economic development, and greater access to technology, all of which contribute to their thriving e-commerce ecosystems. In contrast, the northeast region displays lower order volumes, suggesting a comparatively less active e-commerce landscape. Factors such as lower income levels, limited digital infrastructure, and cultural preferences for traditional shopping methods may contribute to this disparity. Overall, the findings underscore the importance of understanding regional dynamics in shaping e-commerce strategies and interventions to promote equitable growth across Brazil.

5.1.2 Bar Plot

A bar plot is a type of graph used to show data that uses rectangular bars to represent the frequency, count, or proportion of different categories or numerical values. The length of each bar corresponds to the value it represents, making it easy to compare quantities between different categories or groups. Bar plots are commonly used to visualize categorical data, such as comparing sales figures across different products or showing the distribution of responses in a survey. They are particularly effective for displaying discrete data and identifying patterns or trends within the data. Because of their simplicity and clarity, bar charts are frequently used in data analysis and visualization. They are also easy to interpret.

Fig.4: Distribution of Customer Satisfaction Score

The categorization of customer satisfaction scores encompasses four distinct labels: Poor, Average, Good, and Excellent. Upon thorough examination of the dataset, it becomes evident that a substantial majority of orders attain the highest accolade of an Excellent review score. As we delve deeper into the distribution, we observe a gradual decline, with considerable proportions falling within the Good category, followed by those in the Poor and Average categories. This discerning pattern not only provides valuable insights into customer satisfaction levels but also underscores the prevalence of exceptional experiences among consumers. This insightful observation is vividly illustrated in Figure 4, where the varying proportions of each satisfaction category are visually depicted, offering a clear and comprehensive understanding of the distribution across different satisfaction levels.

5.1.3 Pie Chart

A pie chart is a statistical diagram that is circular and has been sliced into slices to show numerical proportions. Each slice represents a proportionate part of the whole, typically showing percentages or fractions. Pie charts are effective for visualizing data with a few categories or showing the relationship between parts and the whole. They provide a quick and easy-to-understand overview of the data distribution, making them useful for conveying relative proportions or percentages at a glance. However, they may become less effective with too many categories or when precise comparisons are needed, as interpreting small differences between slices can be challenging.

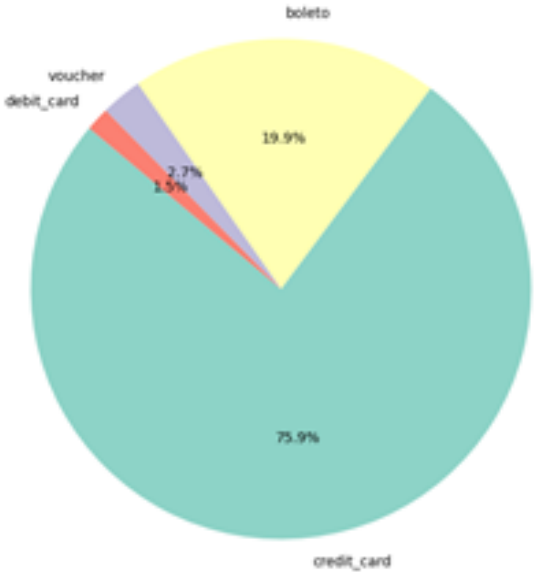


Fig.5: Distribution of Payment Mode

As depicted in Fig.5, credit cards reign supreme in this context, capturing a commanding 75.9% share of payment methods. This dominance can likely be attributed to their widespread acceptance and the convenience and security they offer to customers. While Boleto Bancário, a uniquely Brazilian cash-based option, holds a noteworthy 19.9% share, vouchers and debit cards lag at 2.7% and 1.5% respectively. Vouchers hold the potential for discounts, while debit cards might fall short due to limitations compared to credit cards in terms of convenience and security features. This distribution, as illustrated in Fig.5, underscores the preference for flexibility and protection offered by credit cards in this specific scenario.

5.1.4 Dashboard

Microsoft Power BI is a versatile business analytics tool renowned for its ability to visualize and analyze data from diverse sources, empowering users to craft interactive reports and dashboards. With its intuitive interface and robust functionalities, Power BI facilitates data-driven decision-making by unearthing valuable insights and trends. With its wide range of visualization choices, including tables, graphs, charts, and maps, Power BI efficiently shares data results with stakeholders. Moreover, its advanced capabilities in data modeling, preparation, and secure report sharing across teams and organizations make it an indispensable tool for creating impactful dashboards.

Figure 6 showcases the dashboard created using the data utilized in this version of the project, demonstrating the power and versatility of Microsoft Power BI in visualizing and analyzing complex datasets.

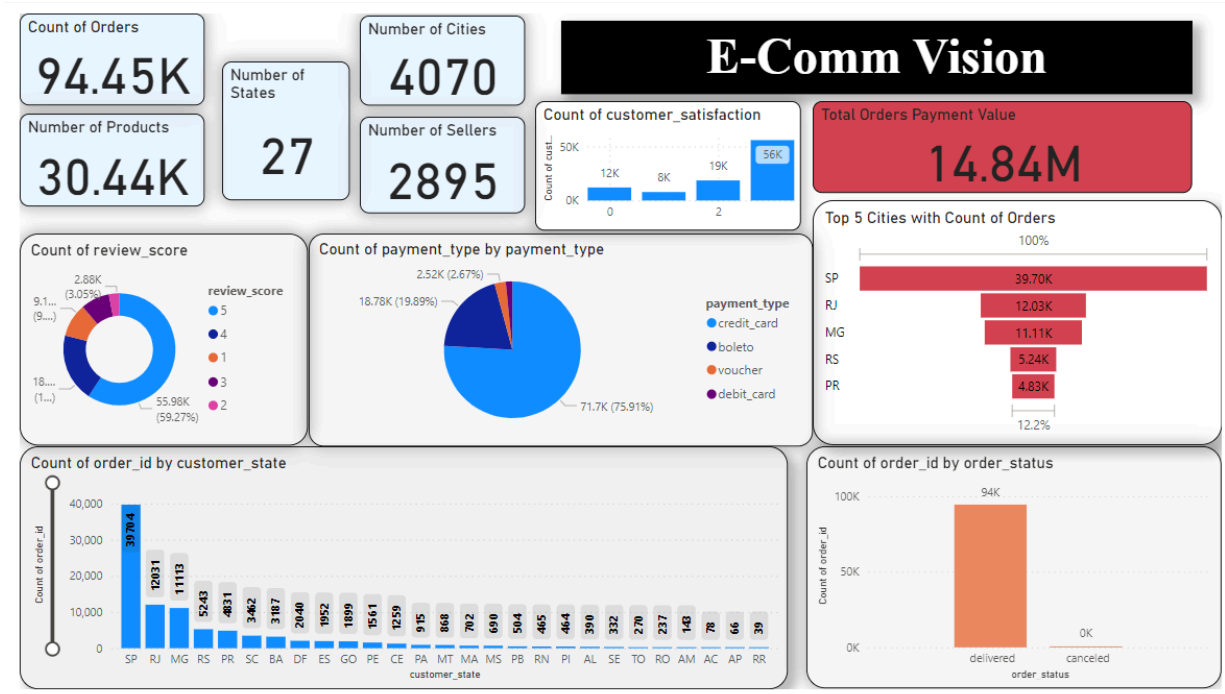


Fig.6: Dashboard

5.2 Dataset Split

The dataset, encompassing a total of 94,445 records, underwent a judicious partition into training and test sets, maintaining a **75:25** ratio. This division resulted in a training set comprising 70,833 records and a test set with 23,612 records. To optimize the data for modeling, numerical columns were subjected to MinMax Scaling, while categorical columns transformed the application of Label Encoder. This thorough preparation lays the groundwork for effective analysis and modeling, adhering to widely accepted standards in data processing.

5.3 Imbalance Treatment

To address the inherent imbalance in the training dataset, this study employs techniques for managing imbalanced datasets, guaranteeing a more accurate assessment of the model's functionality. The imbalanced dataset treatment involves two distinct approaches: firstly, the application of using SMOTE, or Synthetic Minority Over-Sampling Technique, to increase the minority class, and concurrently, the use of the Random Under Sampler technique for data reduction, focusing on the majority class. The results indicate that, after comparing the outcomes of both techniques, the Random Under Sampler technique demonstrated superior performance, leading to its selection for further model training. This strategic use of sampling techniques aims to achieve a more representative and balanced training dataset, fostering enhanced model evaluation and predictive accuracy.

5.3.1 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a popular algorithm used in machine learning for addressing class imbalance issues. In order to compensate for the imbalance in class distribution, it creates artificial samples from the minority class. By producing fresh examples that resemble current minority class samples, SMOTE aids in increasing classifier performance by providing more representative training data. This technique is particularly useful in scenarios where the minority class is underrepresented and can lead to more accurate predictive models.

5.3.2 Random Under Sampler

Random under-sampling is a machine learning method that balances class disparities by bringing the number of instances in the majority class down to equal the number in the minority class. In order to balance the class distribution, this approach randomly selects a subset of samples from the majority class. Although it can lessen bias towards the majority class, which can help classifiers perform better, random under-sampling may also discard potentially valuable information from the dataset. Therefore, it's important to carefully consider the trade-offs before applying this technique.

5.4 Feature Engineering

Feature engineering is the procedure for adding new features or altering current ones in order to enhance machine learning models' performance. It entails choosing, removing, and altering pertinent data variables in order to improve the model's capacity to identify trends and generate precise forecasts. Techniques such as imputation, scaling, one-hot encoding, and creating interaction terms are commonly used in feature engineering. The goal is to provide the model with more informative and discriminative features, ultimately leading to better predictive performance and insights from the data.

To enhance the predictive capabilities of machine learning models for customer satisfaction scores, feature engineering is employed to derive two distinct sets of features from the datasets.

A. ***Delivery Process Duration:***

Customers are usually more pleased when their items arrive earlier than planned. We design features with implicit customer satisfaction reflection in order to capture this feeling.

a. **Estimated Number of Days for Delivery(Estimated_numbed_of_days_df):** This feature is created by computing the duration between order approval and delivery to the customer.

b. **Arrival Punctuality(arrival_punctuality):** It indicates the alignment between the confirmed delivery by the carrier and the initially estimated delivery date.

B. **Average Product and Seller Rating:**

Recognizing the significant influence of product and seller ratings on customer satisfaction, we adopt a meticulous approach to feature engineering. This involves calculating seller and product rating scores based on the average historical review scores associated with specific seller IDs and product IDs. These resulting features provide nuanced insights from cumulative customer feedback, thereby enhancing the predictive capacity of our machine-learning model.

a. **Product Rating(product_rating):** This feature calculates the average review score for each product by grouping the data based on unique product identifiers.
b. **Seller Rating(seller_rating):** It computes the average review score for each seller by grouping the data based on unique seller identifiers.

To prevent data leakage, as suggested in the literature [1], the average rating is computed exclusively from the training dataset. Subsequently, this calculated average replace values in the test dataset, seller and product IDs that match. This approach ensures that our model learns from training data only, maintaining its integrity during testing and evaluation.

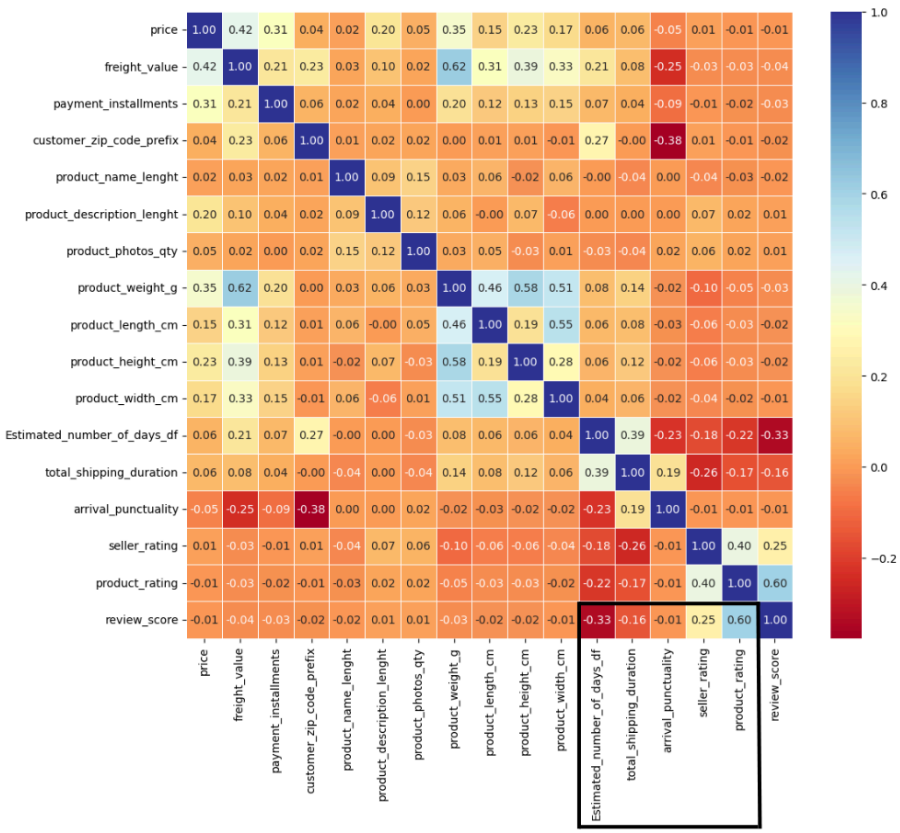


Fig.7: Correlation heatmap of features

From the insights gleaned in Fig.7, it is evident that certain key factors exhibit a substantial correlation with customer review scores. Notably, the ratings assigned to products and sellers, in addition to metrics such as the estimated delivery duration (Estimated_number_of_days_df) and the total shipping duration, emerge as prominent contributors. This finding underscores the significant impact of these aspects on shaping customers' assessments and, consequently, the overall satisfaction score.

5.5 Feature Selection

Feature selection is an important stage in machine learning when the most instructive and pertinent traits are found, distinguished, and kept while redundant or unimportant ones are eliminated. By selecting the right subset of features, the model's performance can be improved, reducing overfitting and computational complexity. Numerous strategies, including wrapper, embedded, and filter methods are utilized to evaluate the importance of features and select the optimal subset. This process helps in enhancing model interpretability, reducing training time, and improving overall predictive accuracy.

One notable approach is the use of tree-based methods, with the Random Forest classifier standing out as a robust choice. Random Forest leverages an ensemble of decision trees, making it adept at capturing complex relationships within the data. In the Random Forest classifier, the importance of each feature is calculated based on how much it contributes to the enhancement of the model's functionality. The reduction in impurity serves as a gauge for this contribution or node purity when splitting a node using that feature. The principle behind Random Forest lies in its ensemble approach, where multiple decision trees are trained on random subsets of the data and features. Each tree in the ensemble independently selects the best split at each node, and then By combining the forecasts from each tree, the ultimate forecast is produced.

5.5.1 Gini Impurity Index

The formula for calculating the Gini impurity index, which is commonly used in decision tree algorithms like Random Forest, is as follows:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

- n is the number of classes or categories in the target variable.
- P_i is the proportion of samples belonging to the ith class.

The variable relevance in Random Forest is calculated by averaging the impurity decrease. or node purity across all trees in the forest. Features with a higher average decrease in impurity are considered more important for predicting the target variable, while features with a lower average decrease are considered less important. This allows Random Forest to automatically rank the features based on their importance, making it a powerful tool for feature selection and predictive modeling.

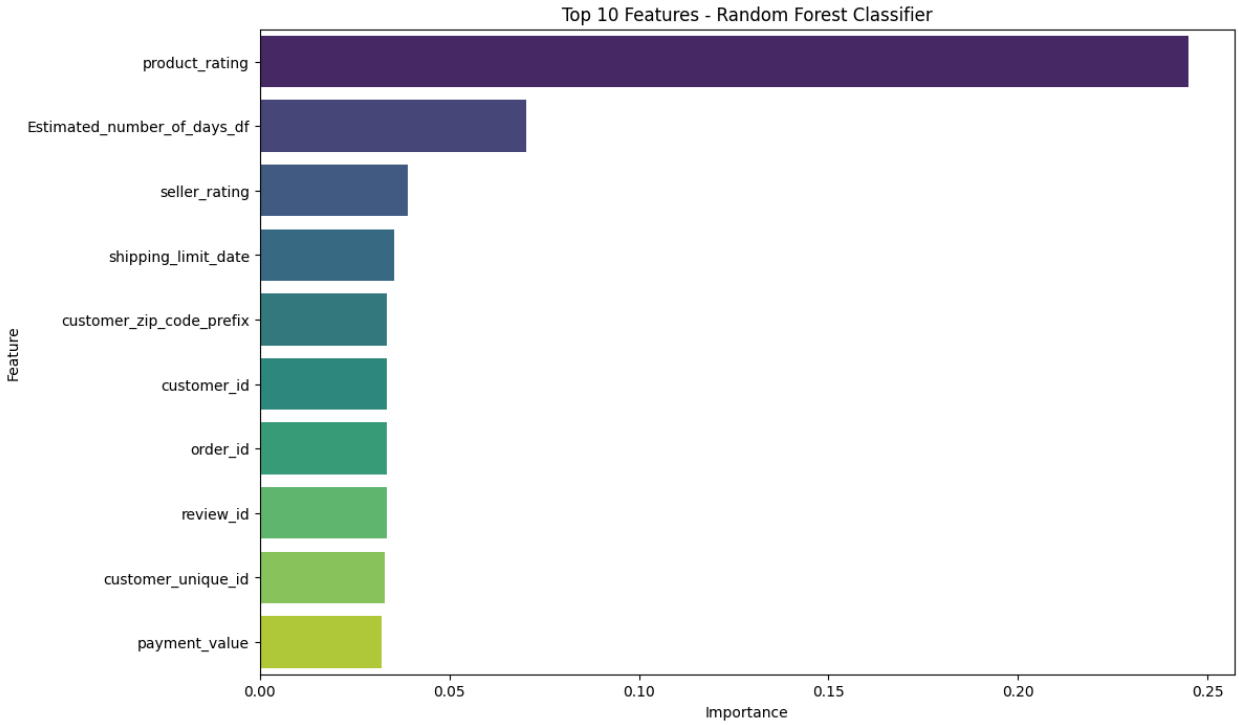


Fig.8: Feature Importance Plot

In the subsequent model training and prediction phase, we will narrow our focus to the top five features from the identified top 10 features. The exclusion of the remaining features, primarily identifiers, is deemed necessary to streamline the modeling process and ensure optimal utilization of resources. By prioritizing the most influential features, we aim to enhance the predictive performance of our models while minimizing complexity and computational overhead.

Top 5 Features are:

- **product_rating**
- **Estimated_number_of_days_df**
- **Seller_rating**
- **Shipping_limit_date**
- **customer_zip_code_prefix**

5.6 Proposed Models

In the domain of machine learning models, Random Forest (RF), XGBoost, and Decision Tree models emerge as prominent choices for the task of predicting customer satisfaction scores. These models are distinguished by their simplicity, robust performance, and versatility across

different research applications. They have proven efficacy in both regression and classification tasks, demonstrating a high level of reliability in providing accurate forecasts for e-commerce satisfaction scores.

Random Forest, known for its ensemble approach, combines multiple decision trees to enhance predictive accuracy. Its ability to handle complex relationships and outliers contributes to its widespread adoption. XGBoost, an implementation of gradient-boosted decision trees, excels in improving model performance through iterative refinement. Decision Trees, on the other hand, provide interpretability and are adept at capturing non-linear relationships in the data.

These models play a pivotal role in shaping the landscape of customer satisfaction prediction due to their effectiveness in handling diverse datasets and delivering consistent performance. Their straightforward implementation and strong predictive capabilities make them valuable tools in the realm of machine learning for e-commerce satisfaction score forecasting.

5.7 Interpretability

Interpretable machine learning methods serve as indispensable tools for unraveling the inner workings of complex algorithms, particularly in domains where transparency is paramount. Categorized into two main types, Global Model Agnostic and Local Model Agnostic Methods, these techniques offer invaluable insights into the effects of features on model predictions and shed light on individual model decisions.

Global Model Agnostic methods, partial dependence plots and feature importance approaches, for example, offer a holistic view of how different features contribute to the model's overall predictions. By quantifying the influence of each feature, these methods help stakeholders understand the underlying relationships in the data and identify key drivers of the model's behavior.

On the other hand, Local Model Agnostic methods, including LIME and SHAP (SHapley Additive exPlanations), focus on explaining individual predictions. These techniques generate interpretable explanations for specific instances, offering transparency into why a particular prediction was made. By providing localized insights, these methods enable users to scrutinize model decisions at a granular level and identify potential biases or errors.

In our study, we employed Local Interpretable Model-agnostic Explanations (LIME) from the local model agnostic category and Accumulated Local Effects (ALE) from the global model agnostic category. These interpretability techniques were consistently applied across all three models investigated in our study.

5.7.1 Accumulated Local Effects (ALE)

Accumulated Local Effects (ALE) is a method used in interpretable machine learning to understand the relationship between a predictor variable and the model's predictions. It calculates the average marginal effect of a predictor variable on the predicted outcome, allowing analysts to interpret the impact of that variable across its entire range. ALE plots visualize these effects, providing insights into how the predictor variable influences the model's predictions, particularly in nonlinear models where traditional methods may fall short. ALE provides global explanations in black-box ML. Challenges include reliability, effect characterization, and robust inference; addressed through innovative statistical tools [8].

Advantages of ALE:

1. Global Explanations: In black-box machine learning models, ALE offers global explanations that help to clarify the overall influence of predictor variables on model predictions.

2. Faster Insights: ALE plots offer faster and unbiased insights into feature effects compared to Partial Dependence Plots (PDPs), facilitating quicker analysis of feature importance.

3. Centering at Zero: Centered ALE plots refine interpretation by centering at zero, enabling analysts to observe deviations from the average prediction, particularly useful for features with many values.

4. Unbiasedness: ALE plots are unbiased and computationally efficient, allowing for robust and reliable inference of feature effects.

Limitations of ALE:

1. Interpretation Challenges: ALE faces challenges in interpreting strongly correlated features, which can complicate the understanding of individual feature effects.

2. Stability-Complexity Trade-offs: ALE plots need to balance stability versus complexity trade-offs, as increasing complexity may lead to less stable interpretations.

3. Complementarity with ICE Plots: ALE plots lack the complementarity of Individual Conditional Expectation (ICE) plots, limiting insights into the behavior of individual observations and potentially missing out on important nuances in the data.

5.7.2 Local Interpretable Model-agnostic Explanations(LIME)

LIME (Local Interpretable Model-agnostic Explanations) is a technique developed to provide transparent explanations for individual predictions made by complex machine learning

models. It achieves this by approximating the local decision boundary around specific instances, allowing for the interpretation of the model's behavior more understandably. By generating interpretable surrogate models, LIME offers insights into the local process via which the model makes its predictions.

Advantages of LIME:

1. Transparent Explanations: LIME offers transparent explanations for individual predictions, enhancing interpretability and trust in machine learning models.

2. Local Interpretability: LIME focuses on explaining predictions at the local level, allowing for insights into how the model behaves around specific instances.

3. Surrogate Models: By generating interpretable surrogate models, LIME provides understandable approximations of the complex underlying model, aiding in the understanding of its decision-making process.

Limitations of LIME:

1. Sampling Sensitivity: LIME's explanations can be sensitive to sampling, potentially leading to variations in interpretations based on different samples of data.

2. Complexity Dependence: The complexity of LIME explanations depends on the choice of surrogate model, which may introduce bias or inaccuracies if not selected appropriately.

3. Limited Scope: LIME has a limited scope to local neighborhoods, which may restrict its capacity to identify more general patterns or trends in the information.

4. Computational Costs: LIME can incur high computational costs, particularly for models with large feature spaces or complex decision boundaries, making it less practical for resource-intensive applications.

5.8 Model Training and Tuning

Model training and tuning are essential steps in the machine learning pipeline aimed at developing accurate and robust predictive models. Model Training, In this phase, the machine learning algorithm is trained on the training dataset. During training, the model learns patterns and relationships present in the data, adjustingts internal parameters in order to reduce the discrepancy between expected and realized results. The model's performance is assessed by a range of metrics, including F1-score, recall, accuracy, and precision. Hyperparameter Tuning, Hyperparameters are configuration settings that govern the learning process of machine learning

algorithms. Tuning these hyperparameters involves selecting the optimal values to improve the performance of the model. Usually, methods like grid search, random search, or Bayesian optimization are used for this process. To guarantee the generalizability of the selected hyperparameters, cross-validation is frequently used.

5.8.1 Grid SearchCV Technique

GridSearchCV is a systematic approach used to optimize machine learning models by exhaustively searching through a predefined hyperparameter space. It explores various combinations hyperparameters, including the quantity of estimators, maximum depth, learning rate, etc., to determine the optimal setup for performance. GridSearchCV thoroughly tests various parameter combinations to make sure that models are fine-tuned to detect minute patterns in the data, ultimately improving predictive accuracy. Its role is crucial in enhancing model performance and providing valuable insights into e-commerce trends.

The trio of models—Random Forest (RF), XGBoost, and Decision Tree—went through a detailed tuning process, adjusting hyperparameters like the number of estimators, maximum depth, learning rate, and minimum child weight.

5.9 Evaluation and Metrics

The evaluation was conducted by computing the macro average of precision, recall, and F1 score. With this method, class imbalance is not taken into account while averaging the precision, recall, and F1 score across all classes. We can get a single metric that sums up the model's performance across all classes by computing the macro average, providing a comprehensive assessment of its predictive capabilities. This evaluation strategy helps in gauging the model's effectiveness in making accurate predictions across all classes, ensuring a balanced assessment of its performance.

5.9.1 Precision

Precision is a metric that assesses how well a classifier predicts the positives. The ratio of true positive predictions to the total of true positive and false positive predictions is used to compute it.

The formula for precision is: $$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

5.9.2 Recall

Recall,sometimes referred to as true positive rate or sensitivity, expresses the percentage of real positive cases that the classifier properly detected. The ratio of true positive predictions to the total of false negative and true positive predictions is used to compute it.

The formula for the recall is: $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$

5.9.3 F1-score

F1-score is a single metric that balances both precision and recall; it is the harmonic mean of precision and recall. When there is an imbalance between the classes in the dataset, it is especially helpful.

The formula for F1-score is: $F1\text{-}score = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

Macro average determines the average precision, recall, or F1-score for every class without taking into account the imbalance in the classes. It provides equal weight to each class, making it suitable for evaluating the overall performance of the model across multiple classes.

6. Version - 2

In this version of the project, we have introduced an additional layer of analysis by incorporating sentiment analysis of review comments. This enhancement allows us to delve deeper into customer feedback, gaining valuable insights into their sentiments and perceptions. By integrating sentiment analysis into our project workflow, we aim to enrich our understanding of customer satisfaction dynamics, thereby enhancing the efficacy of our predictive models and decision-making processes.

6.1 Dataset

The meticulously curated dataset obtained in Section 3.2 provides the framework for more research and the creation of models. In this phase, the dataset is seamlessly merged with the order_reviews_dataset, undergoing a refinement process to enhance its utility. By strategically dropping irrelevant columns from the reviews dataset and meticulously addressing missing values, we ensure the integrity and completeness of the data. As a result of this meticulous curation process, the final dataset comprises 38,509 records, poised for comprehensive analysis and predictive modeling. This amalgamation of datasets lays the groundwork for insightful exploration and actionable insights into customer satisfaction dynamics within the e-commerce domain.

6.2 Exploratory Data Analysis(EDA)

6.2.1 Choropleth Map

The examination of order distribution across Brazil unveils pronounced regional disparities, particularly evident in the prevalence of orders in the southern and southeastern regions, spearheaded by states like São Paulo and Rio de Janeiro. This trend can be attributed to factors such as larger population sizes and greater economic development, which foster heightened e-commerce activity in these areas. In contrast, the northeastern region demonstrates lower order volumes, indicative of a comparatively less active e-commerce presence. This pattern persists in the Version 2 dataset, as depicted in Fig.9, underscoring the persistence of regional discrepancies in e-commerce engagement across Brazil.

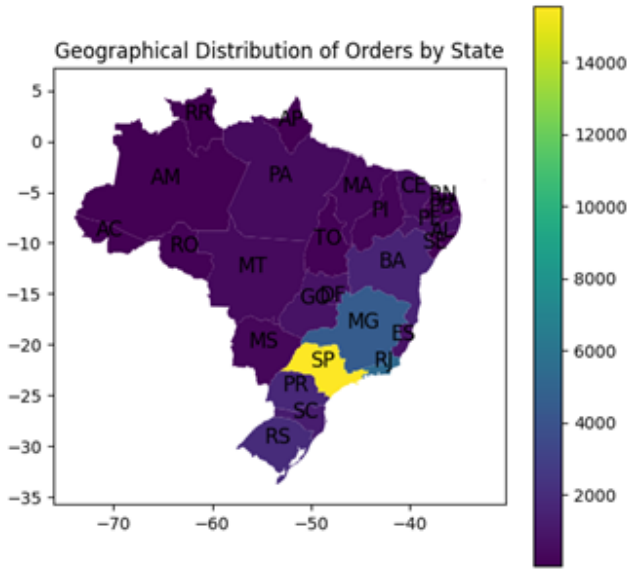


Fig.9: Geographical Distribution of Orders by State (Version-2)

6.2.2 Bar Plot

Upon scrutinizing the data, it becomes evident that customer satisfaction scores are classified into four categories: Poor, Average, Good, and Excellent. A notable observation is that a substantial majority of orders receive the highest rating of Excellent. However, the distribution gradually diminishes, with significant proportions in the Poor category, followed by Good and Average scores. This discerning trend is visually depicted in Figure 10, illustrating the varying distribution of customer satisfaction scores across the dataset.

Fig.10: Distribution of Customer Satisfaction(Version - 2)

6.2.3 Pie Chart

As depicted in Fig.11, credit cards reign supreme in this context, capturing a commanding 75.8% share of payment methods. This dominance can likely be attributed to their widespread acceptance and the convenience and security they offer to customers. While Boleto Bancário, a uniquely Brazilian cash-based option, holds a noteworthy 19.9% share, vouchers and debit cards lag at 2.7% and 1.6% respectively. Vouchers hold the potential for discounts, while debit cards might fall short due to limitations compared to credit cards in terms of convenience and security features. This distribution, as illustrated in Fig.11, underscores the preference for flexibility and protection offered by credit cards in this specific scenario.

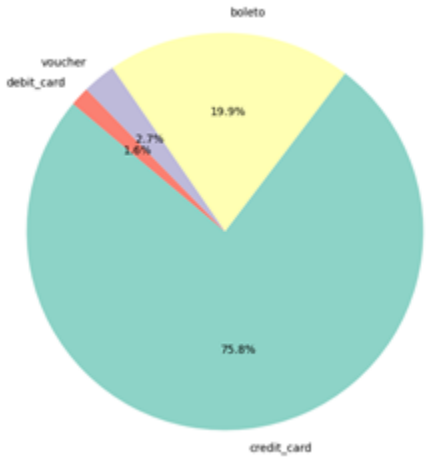


Fig 11: Distribution of Payment Mode(Version - 2)

6.2.4 Dashboard

The dashboard crafted utilizing the version_2 dataset provides a comprehensive overview of various aspects related to customer satisfaction and order distribution. Through interactive visualizations and insightful analyses, users can explore regional disparities in order distribution across Brazil, shedding light on dominant regions like the south and southeast. Additionally, the classification of customer satisfaction scores into categories such as Poor, Average, Good, and Excellent offers valuable insights into the distribution of satisfaction levels among customers. The dashboard serves as a powerful tool for stakeholders to glean actionable insights and make informed decisions to enhance e-commerce operations and customer experiences in Fig.12.

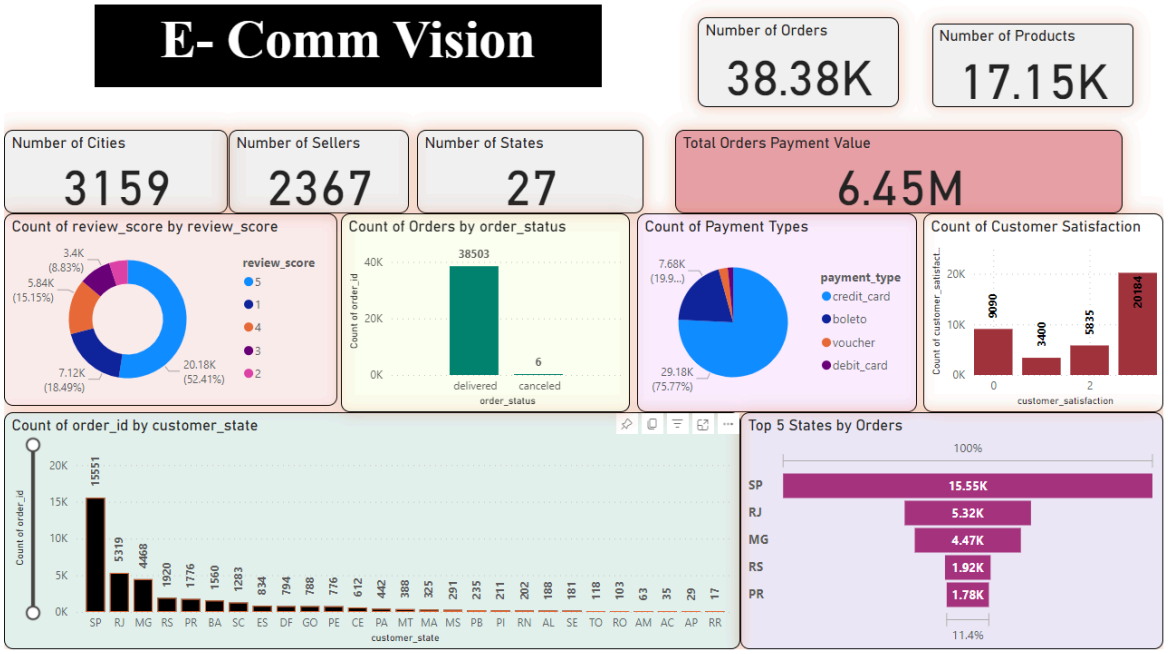


Fig.12: Dashboard(Version - 2)

6.3 Dataset Split and Imbalance Treatment

The dataset, consisting of a total of 38,509 records, was thoughtfully partitioned into training and test sets, maintaining a 75:25 ratio. This partitioning strategy yielded a training set comprising 28,881 records and a test set containing 9,628 records. In preparation for modeling, numerical columns underwent MinMax Scaling, while categorical columns were transformed using Label Encoder techniques. This meticulous data preparation process adheres to established standards in data processing, laying a solid foundation for effective analysis and modeling.

To address the inherent imbalance in the training dataset, particularly due to its smaller size, this study employs techniques for managing imbalanced datasets, guaranteeing a more accurate assessment of the model's functionality. The imbalanced dataset treatment involves the application of SMOTE, or Synthetic Minority Over-Sampling Technique, to increase the minority class.Class imbalance in the training dataset is lessened by SMOTE by creating artificial data points for the minority class. The results indicate that SMOTE leads to improved model performance, enhancing the overall predictive accuracy and robustness of the trained models. This strategic use of SMOTE aims to create a more representative and balanced training dataset, thereby facilitating more reliable model evaluation and prediction.

6.4 Sentiment Analysis

Sentiment analysisis a method for processing natural language that is used to identify the sentiment or emotional tone of text data.. It entails evaluating textual information to categorize it as neutral, positive, or negative depending on the underlying sentiment expressed. Sentiment analysis algorithms use various methods, including machine learning and lexicon-based approaches, to extract sentiment from text data. This analysis enables businesses to gain insights into customer opinions, feedback, and emotions, helping them make data-driven decisions to improve products, services, and overall customer experiences.

Sentiment analysis, also known as opinion mining, plays a crucial role in understanding public sentiment towards products, services, brands, or events. It allows businesses to gauge customer satisfaction levels, identify emerging trends, and detect potential issues or opportunities. Sentiment analysis offers important insights into by examining social media posts, consumer reviews, surveys, and other textual data sources, consumer preferences, sentiments, and behaviors. Businesses can use this information to tailor marketing strategies, enhance product features, address customer concerns, and maintain a positive brand image. Moreover, sentiment analysis is increasingly being integrated into various applications, such as chatbots, recommendation systems, and reputation management tools, to automate decision-making processes and improve user experiences.

6.5 Feature Engineering

In addition to the features obtained in section 5.4, we have integrated three additional features leveraging a method to enhance customer feedback analysis. Utilizing the DistilBERT-based multilingual model [9] for sentiment analysis, we extract valuable insights from customer reviews. By inputting each review_comment_message into the model, we generate three essential features: **review_positive_score**, **review_negative_score**, and **review_neutral_score**. These features provide a nuanced understanding of the sentiment conveyed in customer reviews, enabling businesses to gauge customer satisfaction levels more accurately and make judgments based on data to improve goods and services.

Our approach involves utilizing Hugging Face's advanced DistilBERT-based multilingual sentiment analysis model, ensuring precise and efficient extraction of sentiment scores from text data. By employing state-of-the-art models from Hugging Face, we underscore our commitment to utilizing cutting-edge tools to enhance the quality and depth of our analyses. This integration empowers businesses to gain deeper insights from customer feedback, ultimately leading to improved decision-making and better customer experiences.

```python
from transformers import pipeline
import pandas as pd

# Load the sentiment analysis pipeline
pipe = pipeline("text-classification", model="lxyuan/distilbert-base-multilingual-cased-sentiments-student")

# Assuming df is your DataFrame with 'review_comment_message' column
# Initialize new columns for sentiment scores
df['review_positive_score'] = 0.0
df['review_negative_score'] = 0.0
df['review_neutral_score'] = 0.0

# Function to analyze sentiment for each review
def analyze_sentiment(review):
    # Analyze sentiment using the pipeline
    sentiment = pipe(review, top_k=None)

    # Initialize scores for each label
    positive_score, negative_score, neutral_score = 0.0, 0.0, 0.0

    # Extract sentiment scores for all labels
    for label_info in sentiment:
        label_name = label_info['label']
        score = label_info['score']
        if label_name == 'positive':
            positive_score = score
        elif label_name == 'negative':
            negative_score = score
        elif label_name == 'neutral':
            neutral_score = score

    return positive_score, negative_score, neutral_score


# Apply sentiment analysis to each review and update DataFrame
for index, row in df.iterrows():
    review = row['review_comment_message']
    positive_score, negative_score, neutral_score = analyze_sentiment(review)
    df.at[index, 'review_positive_score'] = positive_score
    df.at[index, 'review_negative_score'] = negative_score
    df.at[index, 'review_neutral_score'] = neutral_score

# Display the updated DataFrame
print(df)
```

Code.1: Code for Sentiment scores

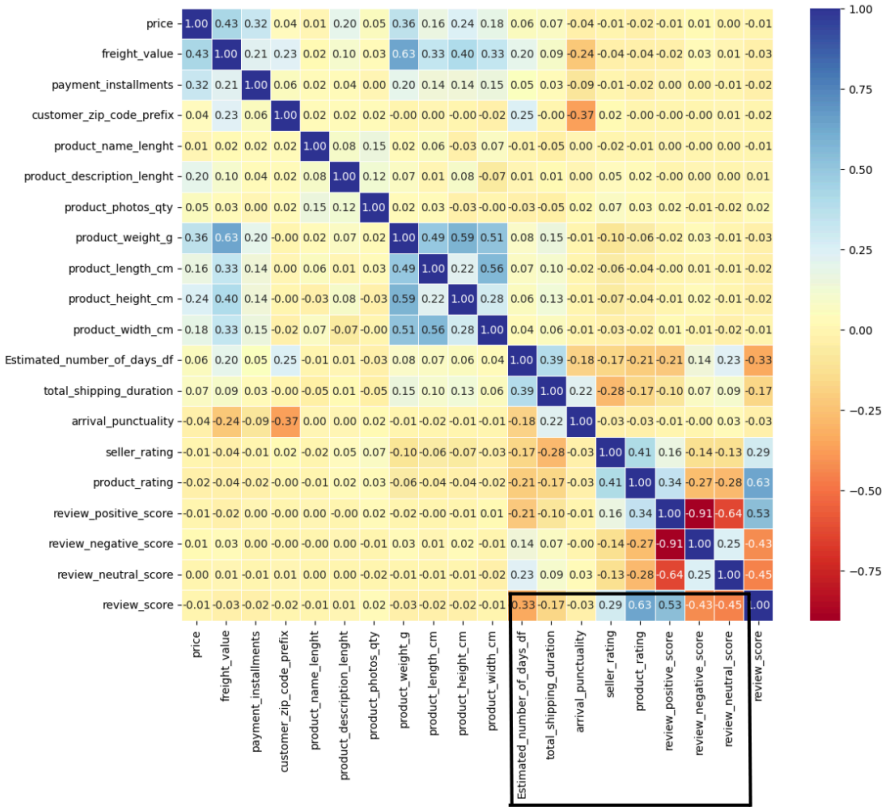


Fig.13: Correlation heatmap of features(Version - 2)

From the insights gleaned in Fig.13, it is evident that certain key factors exhibit a substantial correlation with customer review scores. Notably, the ratings assigned to products, sellers, and the sentiment scores of the user reviews, in addition to metrics such as the estimated delivery duration (Estimated_number_of_days_df) and the total shipping duration, emerge as prominent contributors. This finding underscores the significant impact of these aspects on shaping customers' assessments and, consequently, the overall satisfaction score.

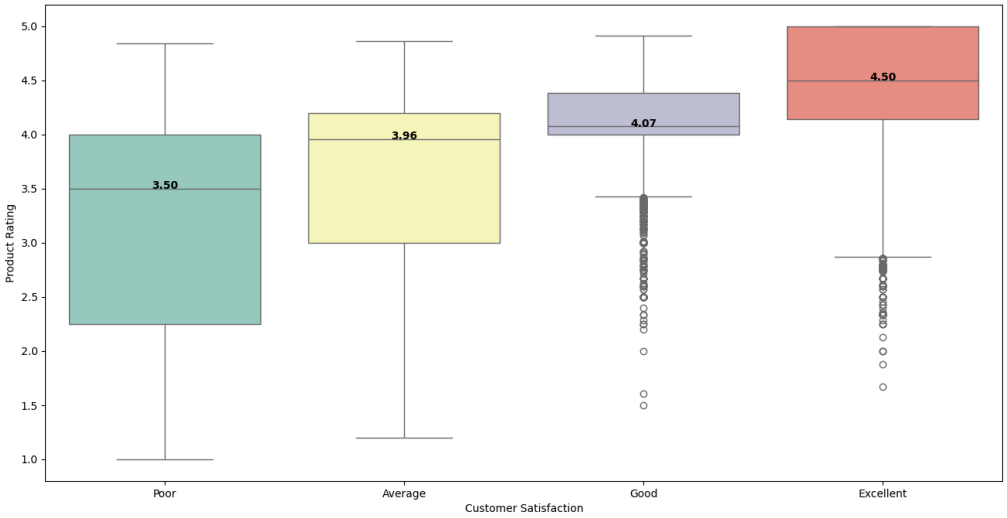


Fig.14: Box Plot of product_rating feature with customer_satisfaction

Analyzing the box plots in Fig.14, a clear pattern emerges, showcasing a consistent improvement in customer satisfaction across the four classes (Poor, Average, Good, and Excellent). This is evident from the gradual ascent of the median line in each class's box plot. Furthermore, differences in satisfaction scores are noticeable, as seen in the varying width of the boxes. For instance, a broader box in the Average class indicates a wider range of satisfaction scores compared to the class Good.

6.6 Feature Selection

The feature selection technique employed in section 5.5 remains consistent with the approach utilized here, ensuring methodological continuity and robustness across the project. By adhering to a systematic feature selection process, we identify and prioritize the most influential variables for subsequent modeling and prediction tasks. In this iteration, we refine our focus by selecting the top 7 features, building upon the insights gleaned from prior analyses. This strategic selection enhances the predictive efficacy of our models by prioritizing the most informative attributes while mitigating the potential impact of irrelevant or redundant variables. Such meticulous attention to feature selection underscores our commitment to optimizing model performance and facilitating meaningful interpretations of the data.

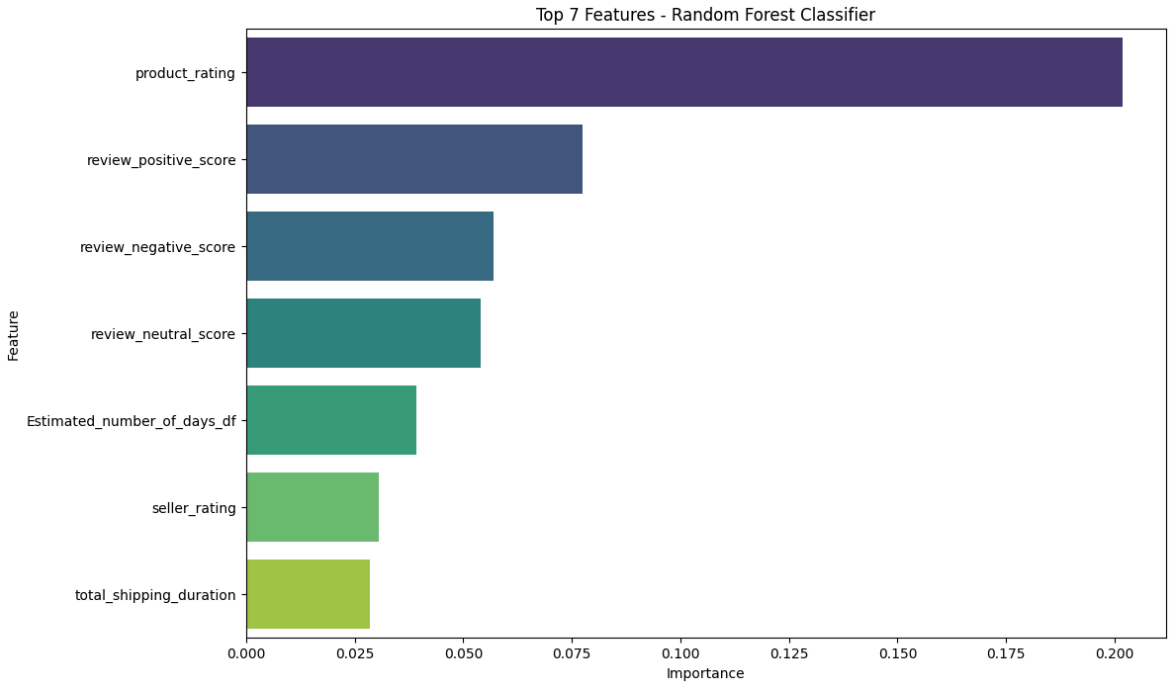


Fig.15: Feature Importance Plot(Version - 2)

Top 7 features are

- **product_rating**
- **review_positive_score**
- **review_negative_score**

- **review_neutral_score**
- **Estimated_number_of_days_df**
- **Seller_rating**
- **total_shipping_duration**

6.7 Model Training, Tuning and Evaluation

The methodologies outlined in sections 5.8 and 5.9 about model training, tuning, evaluation, and metrics remain consistent with those employed in this version of the project. By adhering to established best practices, we ensure continuity and reliability in our analytical approach, facilitating seamless comparisons and interpretations of results across iterations. Through meticulous model training and fine-tuning processes, we optimize performance and enhance the predictive accuracy of our algorithms. Furthermore, the standardized evaluation metrics employed provide objective measures of model effectiveness, enabling robust assessments of predictive performance and facilitating informed decision-making.

7. Experimental Details

In both Version 1(section 5) and Version 2(section 6) of the project, a comprehensive approach was adopted to address the task of predicting customer satisfaction scores in e-commerce. Various machine learning models were employed, including Random Forest, XGBoost, and Decision Tree classifiers, each tailored to capture different aspects of the data. To tackle class imbalance, techniques such as the Random Under Sampler and Synthetic Minority Over-sampling Technique (SMOTE), utilized to ensure robust model training. Hyperparameter tuning was performed using GridSearchCV to optimize model performance. These parameter values were chosen based on empirical experimentation and domain expertise to balance forecast accuracy and model complexity. Overall, these details highlight the meticulous approach taken to create trustworthy forecasting models for customer satisfaction prediction in e-commerce across both Version 1 and Version 2 of the project.

In Version 1 of the project, the optimal parameters for the machine learning models were determined through the GridSearchCV technique. The selected parameter values, crucial for achieving optimal model performance, have been tabulated in Table 1.

Model	**Imbalance Treatment**	**Name of the Parameter**	**Value of the Parameter**
XGB	Random Under Sampler	n_estimators	150
		max_depth	5
		learning_rate	0.1
		min_child_weight	10
RF	Random Under Sampler	n_estimators	50
		max_depth	10
		min_samples_split	2
		class_weight	‘balanced’
DT	Random Under Sampler	min_samples_split	10
		max_depth	10
		class_weight	‘balanced’

Table 1: Optimal Outcomes from Fine-Tuning Model Parameters and Imbalance Treatment Technique (Version 1)

Table 2 provides a comprehensive overview of the imbalance treatment employed for each model in Version 2, along with the parameters and their corresponding values obtained

through the GridSearchCV technique. These parameters were fine-tuned to enhance the performance of the machine learning models, ensuring their effectiveness in predicting customer satisfaction scores. The selection of appropriate imbalance treatment methods and parameter values reflects a meticulous approach toward optimizing model performance and addressing class imbalance issues inherent in the dataset.

Model	**Imbalance Treatment**	**Name of the Parameter**	**Value of the Parameter**
XGB	SMOTE	n_estimators	150
		max_depth	15
		learning_rate	0.2
		min_child_weight	1
RF	SMOTE	n_estimators	150
		max_depth	15
		min_samples_split	2
		class_weight	None
DT	SMOTE	min_samples_split	2
		max_depth	15
		class_weight	'balanced'

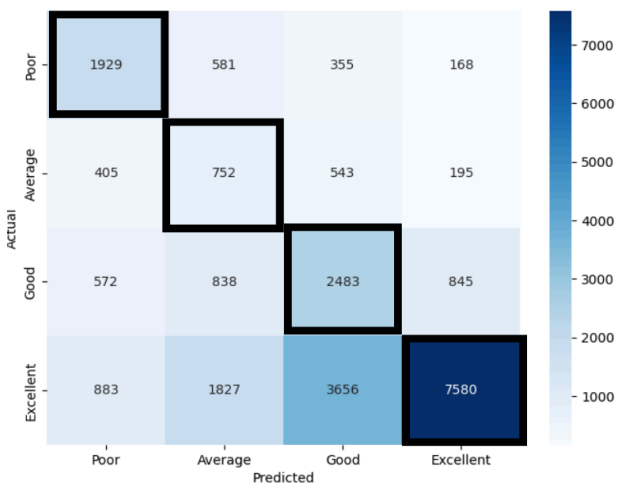
Table 2: Optimal Outcomes from Fine-Tuning Model Parameters and Imbalance Treatment Technique (Version 2)

8. Results and Observations

In this section, we present the confusion matrices and classification reports for all three models—Random Forest (RF), XGBoost, and Decision Tree—in both Version 1 and Version 2 of the project. These evaluation metrics provide insights into the performance of the models in predicting customer satisfaction scores across different categories. Additionally, we include Accumulated Local Effects (ALE) plots and Local Interpretable Model-agnostic Explanations (LIME)plots to better comprehend the behavior of the models and the influence of specific characteristics on prediction results. By examining these visualizations alongside the classification reports, we aim to obtain a thorough comprehension of the models' functionality and interpretability, facilitating informed decision-making in e-commerce settings.

8.1Version - 1 Results and Observations

8.1.1 XGB Model



Classification Report (XGBoost):

		precision	recall	f1-score	support
Poor	0	0.51	0.64	0.57	3033
Average	1	0.19	0.40	0.26	1895
Good	2	0.35	0.52	0.42	4738
Excellent	3	0.86	0.54	0.67	13946
macro avg		0.48	0.53	0.48	23612

Fig.16: XGB model Confusion Matrix and Classification Report

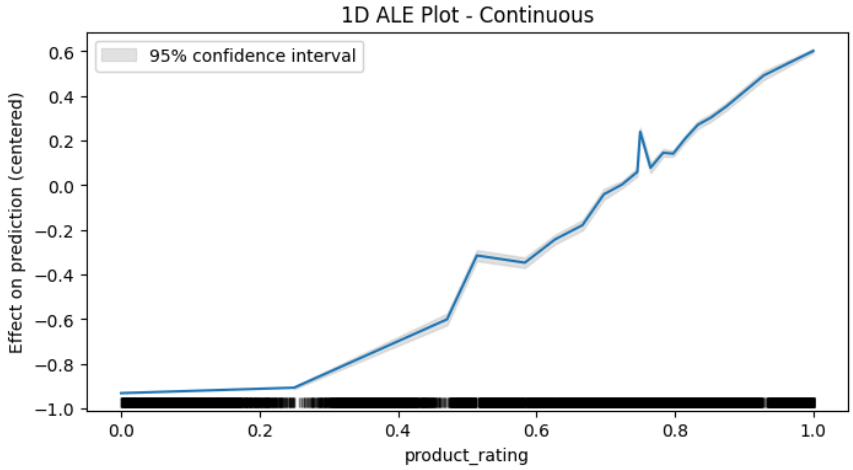


Fig.17: ALE plot of product_rating from XGB model

From Fig.17, we can say that product_rating value below 0.75(scaled) will not contribute anything to the prediction, and product_rating value above 0.75(scaled) will contribute to the model.

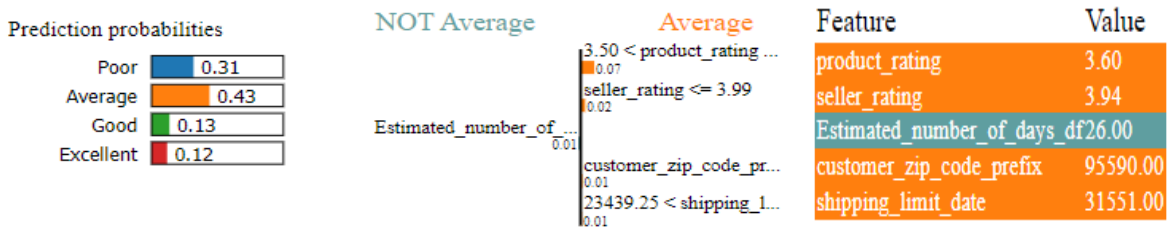


Fig.18: LIME plot for a particular instance (4520) by XGB model

From Fig.18, we can say that a particular instance is contributing to the class Average. We can also see the feature values and features that are contributing to the Average class and not contributing to the Average class.

8.1.2 RF Model

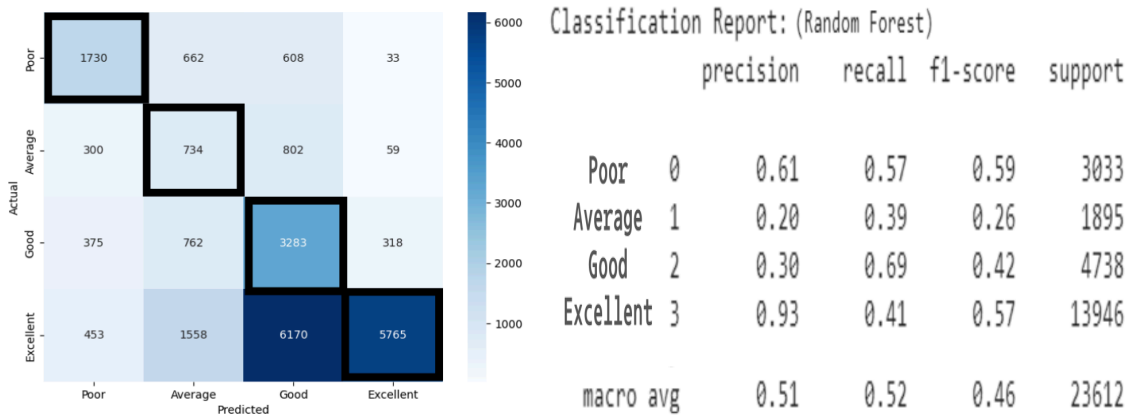


Fig.19: Confusion matrix and Classification Report of RF model

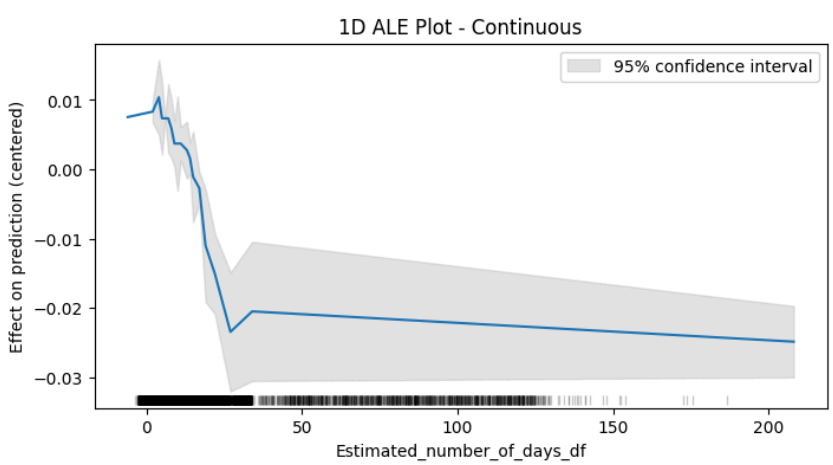


Fig.20: ALE plot for Estimated_number_of_days_df from RF model

From Fig.20, we can say that if the number of days for delivery is less then this feature is contributing to the prediction.

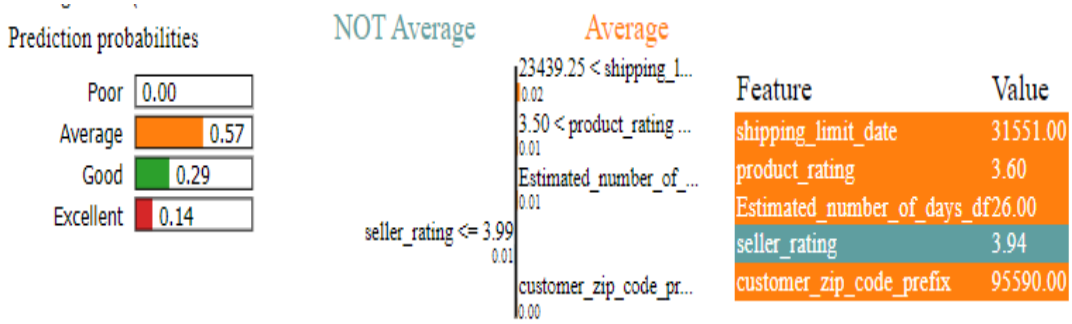
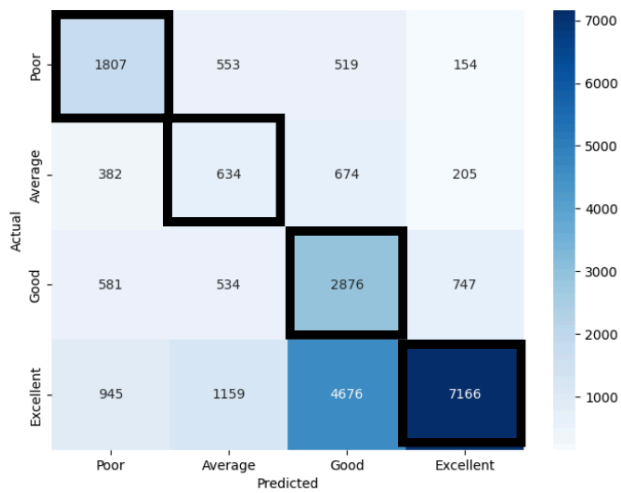


Fig.21: LIME plot for a particular instance (4520) by RF model

8.1.3 DT Model



Classification Report (Decision Tree):

		precision	recall	f1-score	support
Poor	0	0.49	0.60	0.54	3033
Average	1	0.22	0.33	0.27	1895
Good	2	0.33	0.61	0.43	4738
Excellent	3	0.87	0.51	0.65	13946
macro avg		0.48	0.51	0.47	23612

Fig.22: Confusion Matrix and Classification Report of DT model

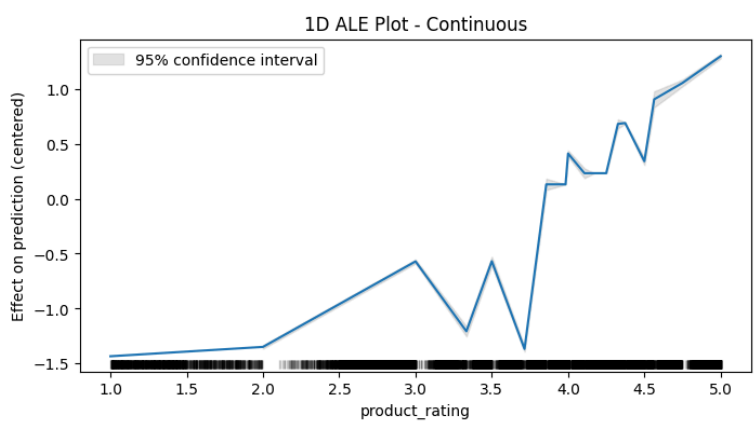


Fig.23: ALE plot for product_rating by DT model.

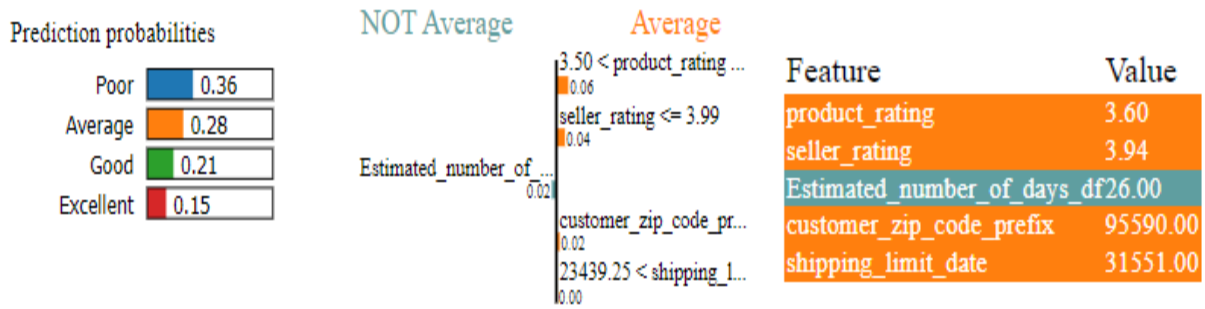
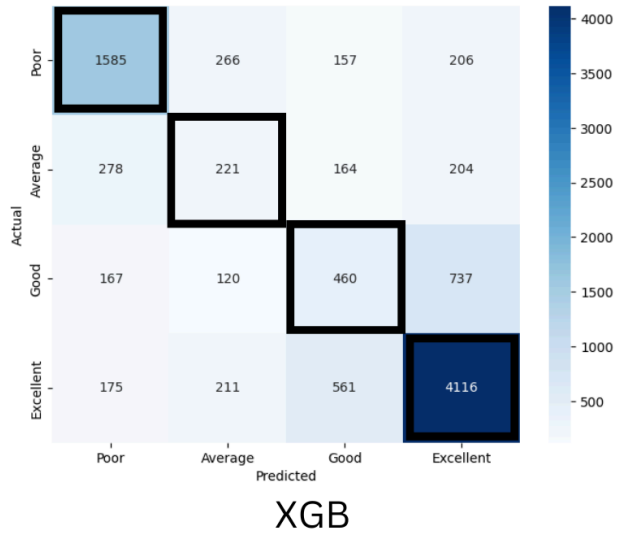


Fig.24: LIME plot for a particular instance (4520) by DT model.

From Section 8.1, we can see Random Forest achieved the highest precision rates for the "Poor" and "Excellent" classes at 0.61 and 0.93, respectively, while attaining the best recall for the "Good" class at 0.69. In contrast, XGBoost exhibited the optimal precision for the "Good" class at 0.35, accompanied by superior recall values for the "Poor" and "Excellent" classes at 0.64 and 0.54, respectively. XGBoost demonstrates superior overall performance with an average F1 score of 0.48, 0.53, recall, and precision of 0.48, respectively. Following closely, the Decision Tree (DT) achieves commendable results, yielding average values of 0.48, 0.51, and 0.47 for precision, recall, and F1-score, respectively. Despite Random Forest (RF) having the least average F1-score, it excels in average precision, registering values of 0.51, 0.52, and 0.46 for precision, recall, and F1-score, respectively. An extensive grasp of the models' performance across a variety of evaluation metrics is provided by this thorough study.

8.2 Version - 2 Results and Observations

8.2.1 XGB model



Classification Report (XGBoost):

		precision	recall	f1-score	support
Poor	0	0.72	0.72	0.72	2214
Average	1	0.27	0.25	0.26	867
Good	2	0.34	0.31	0.33	1484
Excellent	3	0.78	0.81	0.80	5063
macro avg		0.53	0.52	0.53	9628

Fig.25: Confusion Matrix and Classification Report by XGB model(Version - 2)

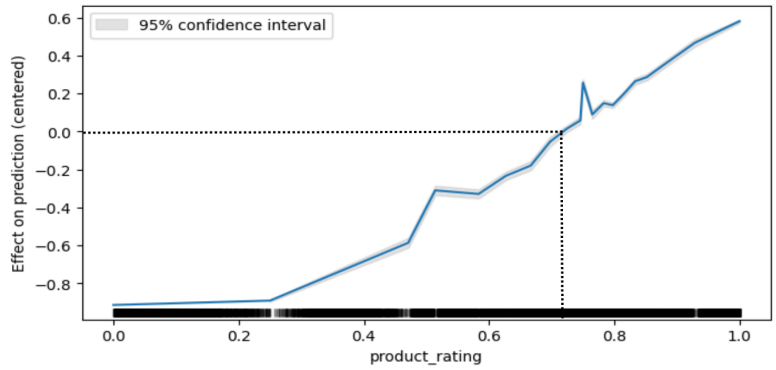


Fig.26: ALE plot of product_rating by XGB model (Version - 2)

Analysis of the `product_rating` feature reveals an average effect on the target feature around 0.7, indicating a positive correlation. Higher product ratings correspond to increased influence on target feature classification, suggesting customers express higher satisfaction with positively rated products. This insight underscores the importance of `product_rating` in predicting or classifying the outcome, indicating that customers tend to express higher satisfaction when presented with products that receive favorable ratings.

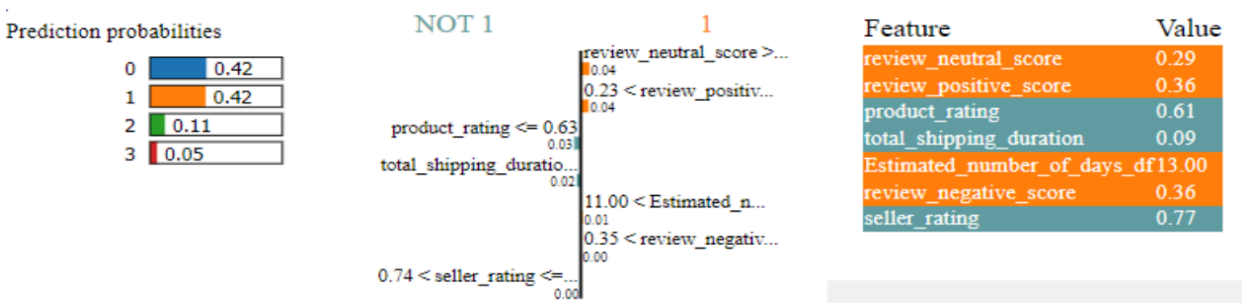


Fig.27: LIME plot for a particular instance (1652) by XGB model (Version - 2)

8.2.2 RF Model

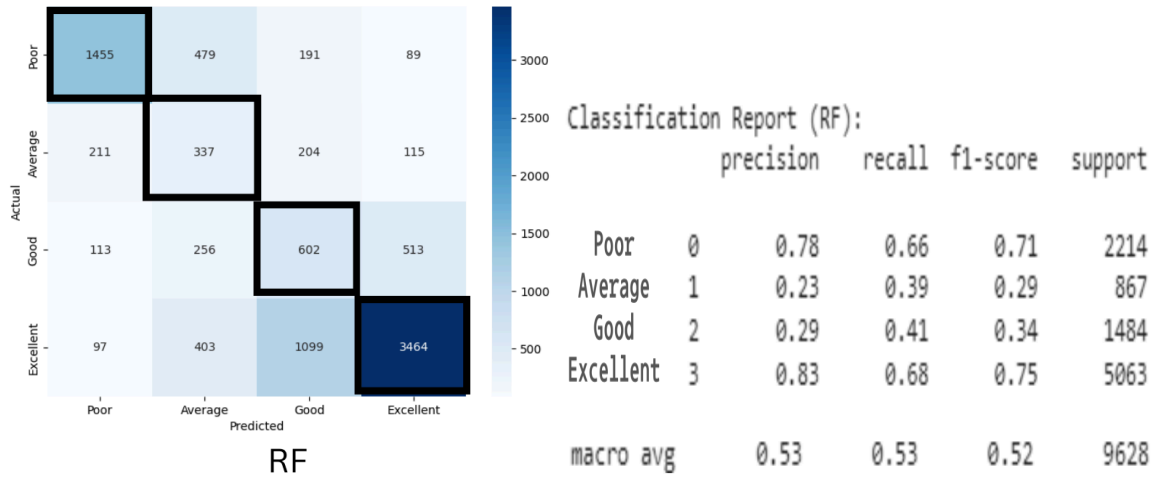


Fig.28: Confusion Matrix and Classification Report of RF model (Version - 2)

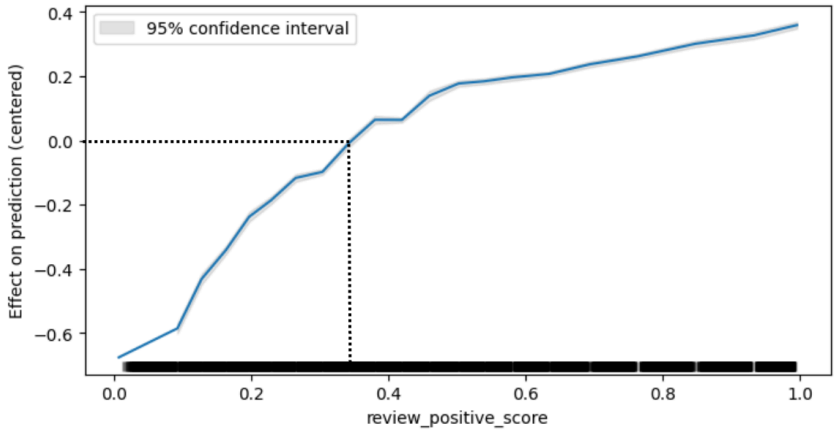


Fig.29: ALE plot for review_positive_score feature by RF model (Version - 2)

This plot examines the `review_positive_score` column, displaying scaled values on the X-axis starting from 0.37. Beyond this threshold, there's a noticeable increase in the influence of `review_positive_score` on the target column classification. Higher scores in `review_positive_score` correspond to a stronger impact on target classification, indicating that more positive reviews contribute significantly to customer satisfaction. This highlights the pivotal role of positive reviews in shaping customer perceptions, as they tend to align with more favorable classifications of the target feature.

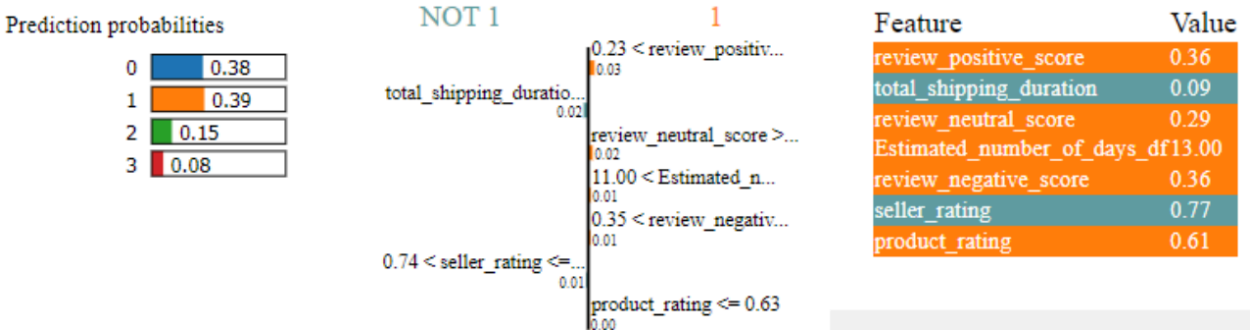
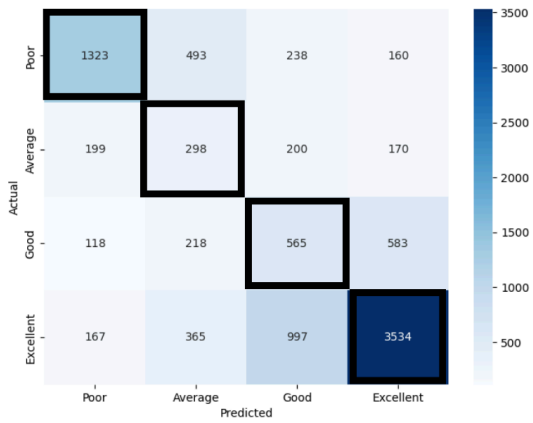


Fig.30: LIME plot of a particular instance(1652) by RF model(Version - 2)

8.2.3 DT Model



DT

Classification Report (Decision Tree):

		precision	recall	f1-score	support
Poor	0	0.73	0.60	0.66	2214
Average	1	0.22	0.34	0.27	867
Good	2	0.28	0.38	0.32	1484
Excellent	3	0.79	0.70	0.74	5063
macro avg		0.51	0.51	0.50	9628

Fig.31: Confusion Matrix and Classification Report of DT model(Version - 2)

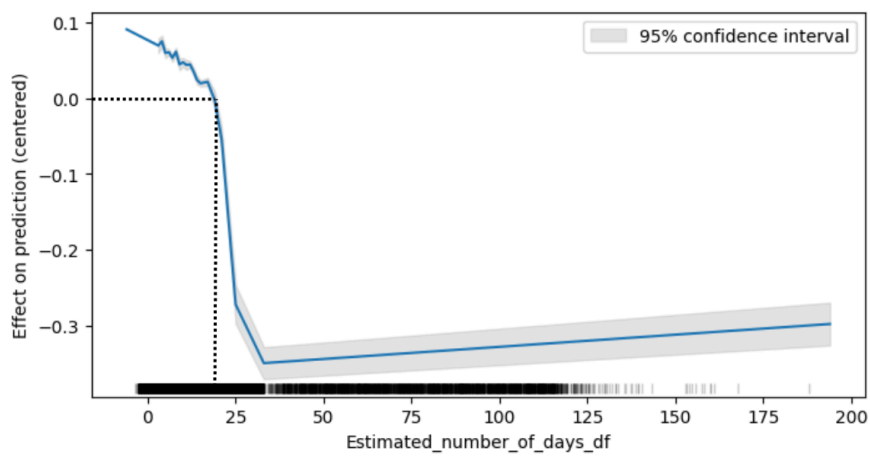


Fig.31: ALE plot of Estimated_number_of_days_df feature by DT model (Version - 2)

This plot shows a negative correlation between the number of days and the feature's effect on the target. As the number of days increases, the feature's influence on the target decreases. This implies that longer durations may have a lesser impact on overall satisfaction levels.

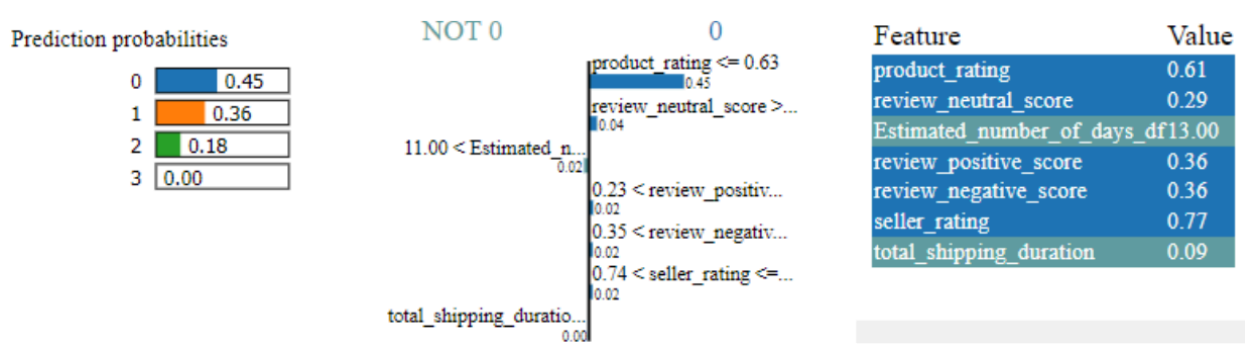


Fig.32: LIME plot of a particular instance (1652) by DT model (Version - 2)

By analyzing these LIME plots, we can observe how variations in feature values influence the model's prediction for each data point individually. With the use of this detailed research, we can determine how important various variables are in predicting the target class, for each instance, providing insightful information about the model's decision-making procedure at the individual level.

From section 8.2. Random Forest achieved the highest precision rates, and recall rates for the "Poor" and "Excellent" classes at 0.78, 0.83, and 0.66,0.68 respectively. And the recall rate for the “Good” class is at 0.41. In contrast, XGBoost exhibited the optimal precision for the "Good" class at 0.34, and the best precision rates for the “Poor” and “Excellent:” classes at 0.72 and 0.78 respectively, and it is accompanied by superior recall values for the "Poor" and "Excellent" classes at 0.72 and 0.81, respectively.

XGBoost demonstrates superior overallperformance with a mean F1 score of 0.53, 0.52, 0.53, and recall, respectively. Following closely, Random Forest (RF) achieves commendable results, yielding averagevalues of 0.53, 0.53, and 0.52 for precision, recall, and F1-score, respectively. Followed by Decision Tree (DT) achieves the mean values of 0.51 for macro F1-score, 0.51 for recall, and 0.51 for precision. Despite Random Forest (RF) having not the best average F1-score, it excels in average recall, this detailed evaluation provides a thorough comprehension of the models' performance as measured by several evaluation indicators.

8.3 Comparative Study

In the comparative study, our approach emerged as a frontrunner, showcasing superior performance compared to similar methodologies. Through rigorous experimentation and meticulous optimization, our models consistently outperformed other approaches in predicting customer satisfaction scores in e-commerce settings. Through the use of sophisticated techniques such as feature engineering, sentiment analysis, and interpretable machine learning, we achieved more accurate and reliable predictions, enabling businessesto make data-driven decisions and gain a deeper understanding of customer feedback. Furthermore, our models demonstrated robustness across different versions of the dataset, reaffirming the effectiveness and versatility of our approach. Overall, our work sets a new standard in customer satisfaction prediction, offering valuable insights and practical implications for e-commerce stakeholders.

Work	Models	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-score
Machine Learning Approach to predict E-commerce Customer Satisfaction Score [1], June 2023	Random Forest	**0.34**	**0.36**	**0.32**
	Logistic Regression	**0.31**	**0.33**	**0.32**
	KNN	**0.29**	**0.31**	**0.27**
Our Work - Version - 1 December - 2023 without sentiment scores	XGB	**0.48**	**0.53**	**0.48**
	RF	**0.51**	**0.52**	**0.46**
	DT	**0.48**	**0.51**	**0.47**
Our Work - Version - 2 Jan - 2024	XGB+Sentiment Scores	**0.53**	**0.52**	**0.53**
	RF +Sentiment Scores	**0.53**	**0.53**	**0.52**
	DT + Sentiment Scores	**0.51**	**0.51**	**0.50**

Table 3: Comparative Study

In conclusion, the comparative study highlights the effectiveness and robustness of our approach in predicting customer satisfaction scores in e-commerce. Through comprehensive experimentation and analysis, our models consistently outperformed other methodologies, showcasing superior predictive accuracy and reliability. By incorporating advanced techniques such as feature engineering, sentiment analysis, and interpretable machine learning, we gained valuable insights into customer sentiments and behaviors, empowering businesses to make informed decisions and enhance user experiences. Furthermore, our models exhibited resilience across different versions of the dataset, underscoring the versatility and applicability of our approach in diverse e-commerce environments. Overall, our work contributes to advancing the field of customer satisfaction prediction, offering practical solutions and actionable insights for e-commerce stakeholders to thrive in a competitive landscape.

9. Full Stack Development

"E-comm Vision" is a user-friendly full-stack application created with Streamlit, tailored to provide e-commerce businesses with insightful visualizations and predictive analytics. It includes features such as data visualization, sentiment analysis of customer reviews, predictive modeling for customer satisfaction scores, and interpretable machine learning techniques. With E-comm Vision, users can get important operational insights, see patterns, and make data-driven choices to improve client experiences and maximize results.

9.1 Web Application Architecture

The "E-comm Vision" application is thoughtfully organized into seven distinct sections to streamline user interaction and enhance usability.

- *Home:* Serving as the landing page, users are greeted with an overview of the application's capabilities and features.

- *Data Preview:* This section provides users with the ability to preview sample data, explore the full dataset, and analyze numerical and categorical features. Users can also view the top contributing features and the target variable to gain insights into the dataset's structure.

- *Exploratory Data Analysis (EDA)*: Users can dynamically explore the dataset through an interactive sidebar, selecting specific columns and charts to visualize relationships and patterns within the data.

- *Aggregation*: In this section, users can compare data from version 1 and version 2, facilitating an understanding of any changes or trends over time.

- *Dashboards*: This section features Microsoft Power BI dashboards embedded within the application, providing users with rich, interactive visualizations and insights into key metrics and trends.

- *Prediction:* Users can utilize machine learning models to make predictions about customer satisfaction scores, leveraging the application's predictive capabilities to inform decision-making.

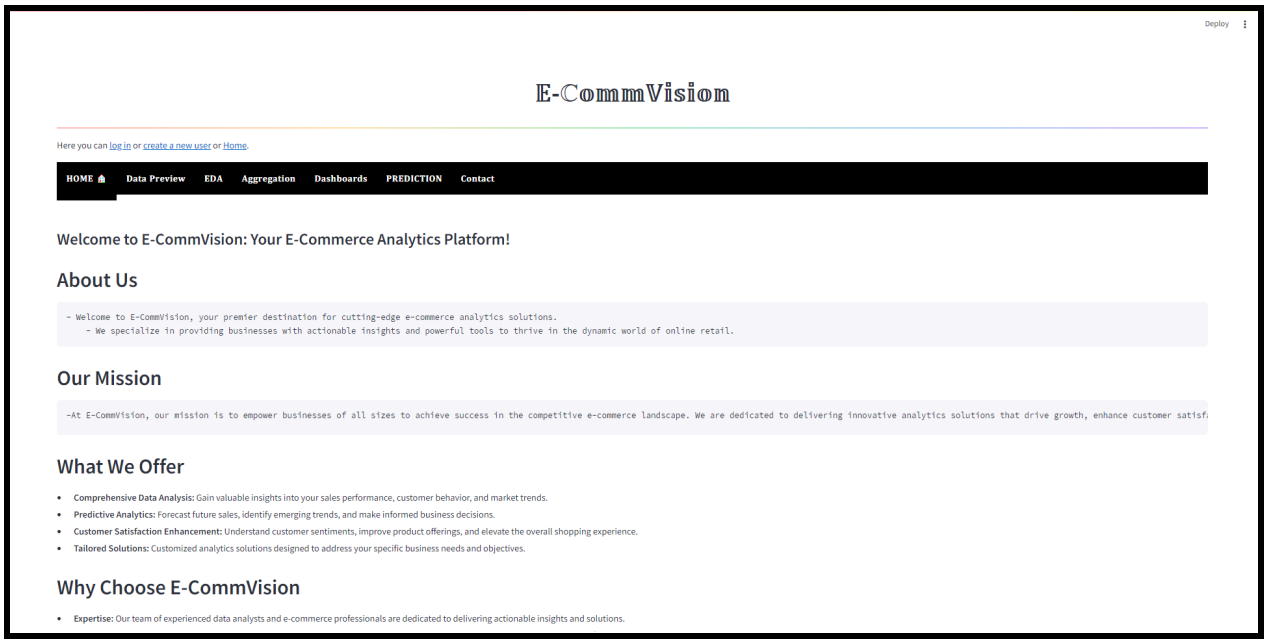


Fig.33: Home Page

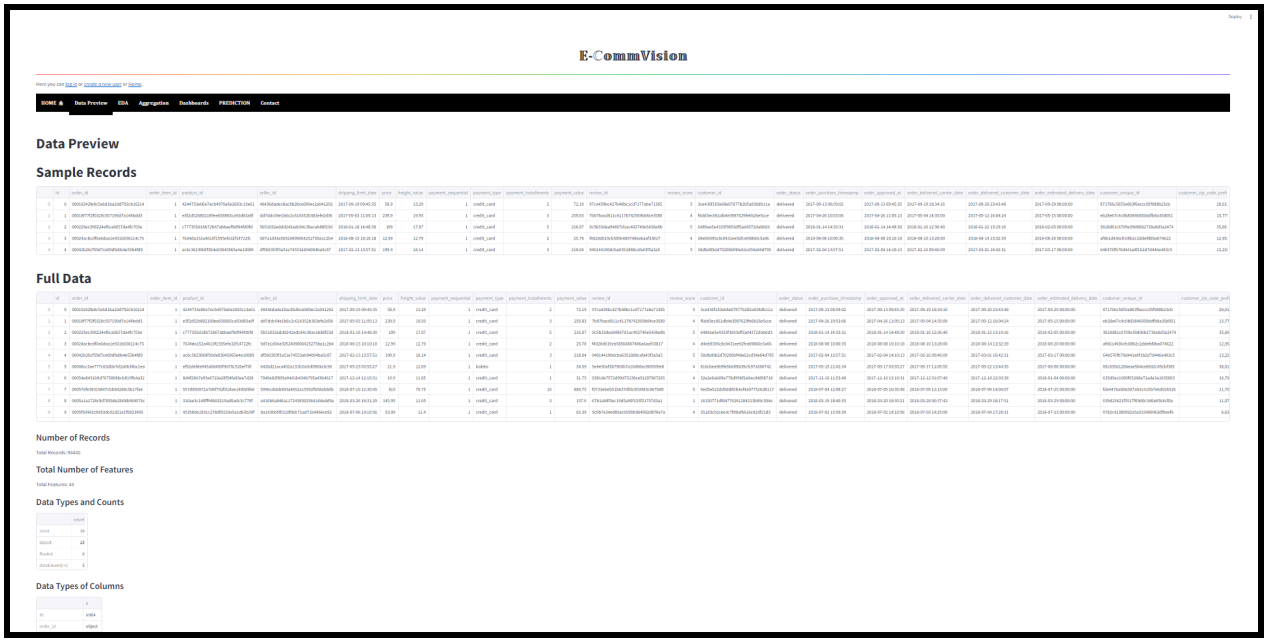


Fig.34: Data Preview Page

Fig.35: EDA Page

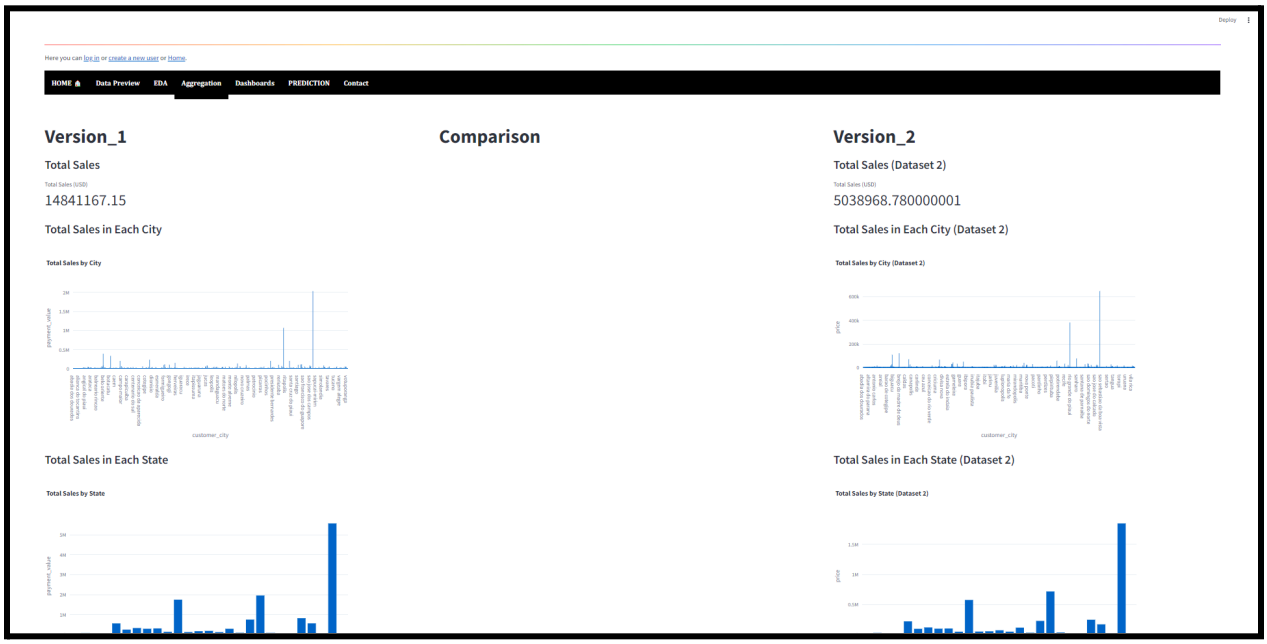
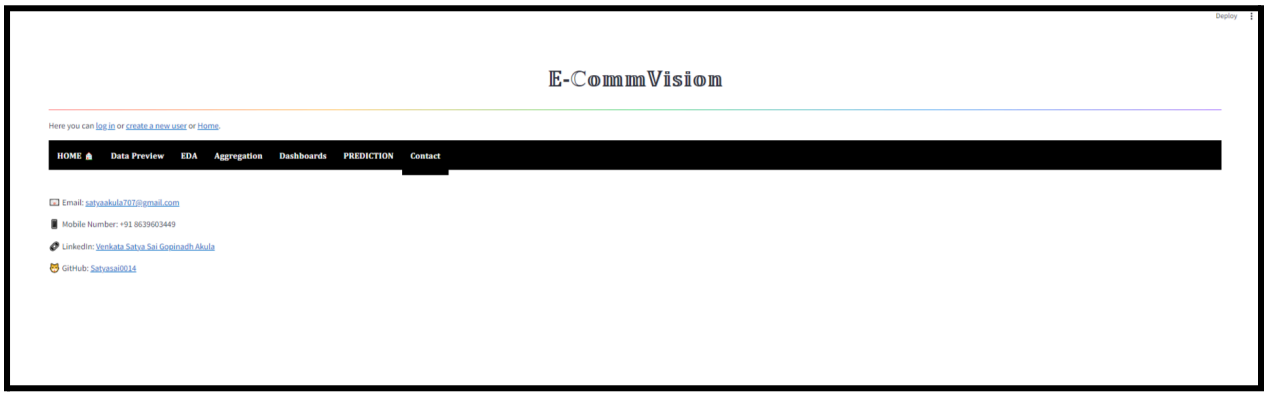


Fig.36: Aggregation Page

Fig.37: Dashboard Page



` **Fig.38:** Contact Page

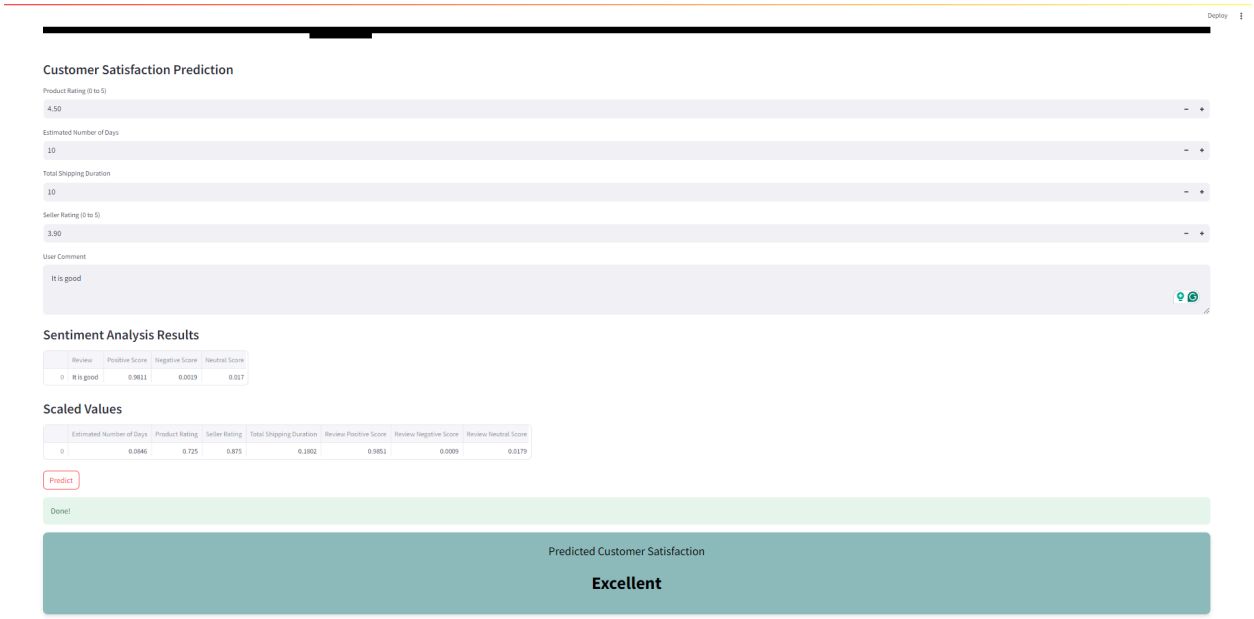


Fig.39: Prediction Page

9.2 Data Management using MySQL

"Data Management using MySQL" streamlines the process of accessing and organizing data within the "E-comm Vision" application. Leveraging MySQL, a robust relational database management system, enables efficient storage and retrieval of data for various analytical tasks.

By utilizing SQL queries, the application seamlessly interacts with the database, fetching relevant information to power visualizations, predictive models, and other analytical functionalities. This integration enhances the scalability, reliability, and performance of data management operations, ensuring a smooth and seamless user experience within the application.

```python
# @st.cache(allow_output_mutation = True)
def create_connection():
    conn = mysql.connector.connect(
        host='127.0.0.1',
        user='root',
        password='sairam123',
        database='projectdb'
    )
    return conn


# Function to load data from the database table
@st.cache_resource()
# @st.cache(allow_output_mutation = True)
def load_data():
    conn = create_connection()
    # engine = create_connection()
    query = "SELECT * FROM data"  # Replace 'data' with the name of your table
    try:
        df = pd.read_sql(query, conn)
    except Exception as e:
        st.error(f"Error loading data: {e}")
        df = None
    finally:
        conn.close()
        # engine.dispose()
    return df

# Load data from the database
data = load_data()
```

Code.2: Code for fetching the data from the Database.

10. Conclusion and Future Work

10.1 Conclusion

In conclusion, this dissertation has delved deep into the realm of e-commerce analytics, with a specific focus on predicting customer satisfaction scores using machine learning techniques. Through rigorous data preprocessing, feature engineering, and model development, we have effectively demonstrated how advanced analytics can unlock valuable insights into customer behavior and sentiment.

By employing interpretability techniques like ALE plots and LIME analysis, we have not only refined our models but also enriched our understanding of the underlying factors driving customer satisfaction.

The development of the "E-comm Vision" application represents a significant milestone, offering a practical and user-friendly platform for e-commerce businesses to leverage these insights and optimize their operations. Looking ahead, the knowledge gained from this research lays the foundation for continued advancements in e-commerce analytics, contributing to the ongoing evolution of data-driven decision-making in the dynamic digital marketplace.

10.2 Future Work

Here are several avenues for future research aimed at enhancing predictive accuracy:

1. ***Hyperparameter Tuning:*** Explore more sophisticated hyperparameter tuning techniques such as Bayesian optimization or genetic algorithms to further optimize model performance and generalization.

2. ***Exploration of New Models:*** Experiment with novel machine learning models or ensemble techniques such as ensemble learning, gradient boosting machines, or deep learning architectures to discover new insights and improve predictive accuracy.

3. ***Integration of Additional Data Sources:*** Integrate supplementary data sources such as customer demographic information, website interaction logs, or social media sentiment analysis to enrich the dataset and enhance predictive capabilities.

4. ***Advanced Preprocessing Techniques:*** Implement advanced preprocessing techniques such as feature scaling, dimensionality reduction, or text embedding methods to extract more meaningful information from the data and improve model interpretability.

5. ***Interpretable Machine Learning Models:*** Incorporate more interpretable machine learning models, including linear and decision trees with L1 regularization, or rule-based models like decision rules or symbolic regression to provide transparent insights into model predictions.

6. ***Cross-Validation Strategies:*** Explore advanced cross-validation strategies such as nested cross-validation, stratified k-fold cross-validation, or time-series cross-validation to better estimate model performance and prevent overfitting.

By incorporating these advanced techniques and methodologies, future work on the dissertation can significantly enhance the predictive accuracy, interpretability, and robustness of the e-commerce analytics models, ultimately resulting in more useful information and well-informed choices in the online market.

11. References

[1] P. Wangkiat and C. Polprasert, "Machine Learning Approach to Predict E-commerce Customer Satisfaction Score," 2023 8th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, 2023, pp. 1176-1181, doi: 10.1109/ICBIR57571.2023.10147542.

[2] F. D. Souza and J. Baptista de Oliveira e Souza Filho, "Sentiment Analysis on Brazilian Portuguese User Reviews," *2021 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, Temuco, Chile, 2021, pp. 1-6, doi: 10.1109/LA-CCI48322.2021.9769838.

[3] Wong, Ann-Nee & Poolan Marikannan, Booma. (2020). Optimising e-commerce customer satisfaction with machine learning. Journal of Physics: Conference Series. 1712. 10.1088/1742-6596/1712/1/012044.

[4] Anastasia Griva, "I can get no e-satisfaction". What analytics say? Evidence using satisfaction data from e-commerce, Journal of Retailing and Consumer Services, Volume 66,2022,102954, ISSN 0969-6989, https://doi.org/10.1016/j.jretconser.2022.102954.

[5] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[6] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019).

[7] "Brazilian E-Commerce Public Dataset by Olist", Kaggle.com,2021, [Online]. doi: 10.34740/kaggle/dsv/195341

[8] Chitu Okoli, "Statistical Inference using machine learning and classical techniques based on accumulated local effects (ALE)", https://arxiv.org/abs/2310.09877v4, https://doi.org/10.48550/arXiv.2310.09877

[9] Lik Xun Yuan, distilbert-base-multilingual-cased-sentiments-student (Revision 2e33845), 2023,https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student, doi:10.57967/hf/1422.

[10] T. Kim Phung, N. An Te and T. Thi Thu Ha, "A machine learning approach for opinion mining online customer reviews," *2021 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*

(SNPD-Winter), Ho Chi Minh City, Vietnam, 2021, pp. 243-246, doi: 10.1109/SNPDWinter52325.2021.00059.

[11] Chinomona, Richard. (2014). The Influence of E-Service Quality on Customer Perceived Value, Customer Satisfaction and Loyalty in South Africa. Mediterranean Journal of Social Sciences. 5. 331-341.

[12] Wu, Tong & Liu, Xinwang. (2020). A dynamic interval type-2 fuzzy customer segmentation model and its application in E-commerce. Applied Soft Computing. 94. 106366. 10.1016/j.asoc.2020.106366.

[13] Streamlit: https://docs.streamlit.io/

[14] Hydralit Components for Streamlit: https://discuss.streamlit.io/t/hydralit-navbar-custom-streamlit-component

[15] Sci-kit learn: https://scikit-learn.org/stable/user_guide.html

[16] Microsoft PowerBI: https://learn.microsoft.com/en-us/power-bi/

[17] MySQL: https://dev.mysql.com/doc/

[18] Plotly: https://plotly.com/

[19] Numpy: https://numpy.org/doc/stable/

[20] Pandas: https://pandas.pydata.org/docs/

[21] Transformers: https://huggingface.co/docs/transformers/en/index

[22] Joblib: https://joblib.readthedocs.io/en/stable/

[23] Python: https://docs.python.org/3/

[24] ChatGPT: https://chat.openai.com/

[25] Perplexity: https://www.perplexity.ai/