MULTIMEDIA UNIVERSITY OF KENYA

FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

PROJECT DOCUMENTATION

UNIT CODE: CCS 2328

Predicting Consumer Behavior: A Machine Learning Approach for Enhanced Business Strategy and Customer Satisfaction

Project By:

NAME: SHILAHO BRANDON

REG NO: CIT-223-034/2019

Project Supervisor:

PETER MUTURI

April,2023

Declaration

I hereby declare that this Project [Documentation] is my own work and has, to the best of my knowledge, not been submitted to any other institution of higher learning.

Student : ___________________ Registration Number: _________________

Signature: .... Date:..

This project [Documentation] has been submitted as a partial fulfillment of requirements for the Bachelor of Science in Computer Science/Information Technology of Multimedia University of Kenya with my approval as the University supervisor.

Supervisor: _________________________

Signature: .... Date: ....

Dedication

I dedicate this project first and foremost to the almighty God who has been there right from the beginning to this very point. Special dedication to my supportive parents and finally to my Supportive lecturers for their continual impact of knowledge.

Acknowledgements

I would like to express my gratitude and appreciation to all those who gave me the possibility to complete this project. Special thanks is due to my supervisor Mr. Peter Muturi whose help, simulating suggestions and encouragement helped me in all the time of coming up with a problem statement and solving the problem by coming up with this project as a solution. I also sincerely thank you for the time spent proofreading and correcting my mistakes.

List of Abbreviations

AUC Area under the curve

CRISP-DM Cross Industry Standard Process for Data Mining

DT Decision Trees

FNN Feed-forward Neural Networks

HMC Higher-order Markov Chains

KNN K-nearest Neighbor

LR Logistic Regression

LSTM Long short-term memory

RF Random Forest

ROC Receiver Operating Characteristi**c**

RNN Recurrent Neural Networks

SVM Support Vector Machines

PCA Principal Component Analysis

Abstract

Due to today’s transition from visiting physical stores to online shopping, predicting customer behavior in the context of e-commerce is gaining importance. It can increase customer satisfaction and sales, resulting in higher conversion rates and a competitive advantage, by facilitating a more personalized shopping process. By utilizing clickstream and supplementary customer data, models for predicting customer behavior can be built. This study analyzes machine learning models to predict a purchase, which is a relevant use case as applied by a large German clothing retailer. Next, to compare models this study further gives insight into the performance differences of the models on sequential clickstream and the static customer data, by conducting a descriptive data analysis and separately training the models on the different datasets. The results indicate that a Random Forest algorithm is best suited for the prediction task, showing the best performance results, reasonable latency, offering comprehensibility and a high robustness. Regarding the different data types, models trained on sequential session data outperformed models trained on the static customer data by far. The best results were obtained when combining both datasets.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of Study

Consumer behavior is a complex phenomenon that has been studied extensively in marketing and consumer research. In recent years, advances in data collection and analysis have led to the development of new methods for predicting consumer behavior. These methods include machine learning and predictive analytics, which allow businesses to analyze vast amounts of data and make accurate predictions about customer behavior.

One of the key drivers of consumer behavior is individual preferences. Studies have shown that consumers are more likely to make purchases when they feel a strong emotional connection to a product or brand. This connection can be built through targeted marketing and personalized recommendations, which rely on data analysis to understand consumer preferences and behaviors.

In addition to individual preferences, consumer behavior is also influenced by social factors such as peer influence and cultural norms. Social media platforms have become an increasingly important tool for understanding these social influences and predicting consumer behavior. By analyzing social media data, businesses can identify trends and influencers that can inform their marketing strategies.

Another important factor in consumer behavior is external events such as economic conditions, weather, and political events. These events can have a significant impact on consumer behavior and must be taken into account when predicting future behavior. Data analysis can help identify patterns and correlations between external events and consumer behavior, allowing businesses to adjust their strategies accordingly.

Overall, the study of consumer behavior is critical for businesses to understand and predict customer behavior. By leveraging data analysis and predictive models, businesses can improve their marketing strategies, enhance customer satisfaction, and increase profitability.

1.2 Problem statement

Predicting consumer behavior is a crucial aspect of marketing that can help businesses tailor their marketing strategies to specific consumer needs. However, accurately predicting consumer behavior remains a challenge for businesses due to the complexity of consumer decision-making processes. Therefore, the aim of this project is to develop a predictive model using machine learning techniques that can accurately forecast consumer behavior based on their past purchasing patterns and demographic information.

1.2.1 proposed solution

The proposed solution for the project on consumer behavior prediction is to use the Naive Bayes Classifier Gaussian Distribution Algorithm to develop a model that can predict consumer behavior based on various attributes such as age, gender, income, and location. This algorithm is suitable for the project because it is simple, fast, and efficient for handling large datasets. It assumes that the attributes are independent of each other, which makes it easy to calculate the probabilities of each attribute. The model will be trained on a large dataset of consumer behavior and tested for accuracy and efficiency. The results will be used to make predictions about future consumer behavior, which will be useful for businesses in making informed decisions about their products and services. The model will be implemented in the Python programming language using relevant libraries such as scikit-learn and pandas.

1.3 Aim of the study

The aim of this study is to predict consumer behavior using machine learning algorithms. Specifically, we will use the Naive Bayes Classifier with Gaussian Distribution Algorithm to develop a predictive model that can be used to determine consumer preferences based on various factors such as demographics, purchasing behavior, and social media activity. By accurately predicting consumer behavior, businesses can improve their marketing strategies, product development, and overall customer satisfaction, leading to increased profitability and competitiveness in the market.

1.4 research objectives

The research objectives for the consumer behavior prediction project are:

1. To collect relevant data on consumer behavior from various sources, including customer databases, online surveys, and social media platforms.
2. To analyze the collected data using advanced techniques such as data mining, machine learning, and predictive analytics to identify patterns and trends in consumer behavior.
3. To develop a predictive model that accurately forecasts future consumer behavior based on the insights gained from data analysis.
4. To deploy the predictive model to help businesses make data-driven decisions that improve their marketing strategies, enhance customer satisfaction, and increase profitability.

1.5 significance/justification of the study

The significance of the consumer behavior prediction study lies in its potential to provide businesses with a powerful tool for improving their marketing strategies, enhancing customer satisfaction, and increasing profitability. By accurately predicting consumer behavior, businesses can make data-driven decisions that lead to increased customer loyalty and satisfaction.

The study's significance can be further elaborated as follows:

1. **Improved Marketing Strategies:** By predicting consumer behavior, businesses can tailor their marketing strategies to meet the needs and preferences of their customers. This can include targeted marketing campaigns, personalized recommendations, and other initiatives aimed at improving customer satisfaction and loyalty.
2. **Enhanced Customer Satisfaction:** By understanding consumer behavior, businesses can identify areas where they can improve their products or services, leading to increased customer satisfaction and loyalty.

3. **Increased Profitability:** By making data-driven decisions based on accurate predictions of consumer behavior, businesses can reduce costs, increase revenue, and ultimately increase profitability.
4. **Competitive Advantage:** By leveraging the power of data analysis and predictive modeling, businesses can gain a competitive advantage over their rivals by staying ahead of consumer trends and preferences.
5. **Contribution to Research:** The study can contribute to the body of research on consumer behavior prediction, providing insights into the factors that influence consumer behavior and the effectiveness of predictive models.

In summary, the significance of the consumer behavior prediction study lies in its potential to improve business strategies, enhance customer satisfaction, increase profitability, and contribute to the field of consumer research.

1.6 Scope (ie defines the system boundary)

he scope of the consumer behavior prediction project includes the following:

1. **Data Collection:** The project will collect relevant data on consumer behavior from various sources, including customer databases, online surveys, and social media platforms.
2. **Data Analysis:** The project will use advanced data mining, machine learning, and predictive analytics techniques to analyze the collected data and identify patterns and trends in consumer behavior.
3. **Predictive Model Development:** Based on the insights gained from data analysis, the project will develop a predictive model that accurately forecasts future consumer behavior.
4. **Model Deployment:** The project will deploy the predictive model to help businesses make data-driven decisions that improve their marketing strategies, enhance customer satisfaction, and increase profitability.

The project's scope is limited to predicting consumer behavior and does not cover other aspects of business operations such as financial management, supply chain management, or human resource management. The project will also focus on using existing data sources rather than collecting new data.

The project's scope is also limited by the availability and quality of data, as well as the complexity of consumer behavior. While the project aims to provide a comprehensive understanding of consumer behavior, it may not be able to capture all the factors that influence it.

In summary, the project's scope includes data collection, analysis, predictive model development, validation, and deployment, focusing on predicting consumer behavior to improve business strategies, enhance customer satisfaction, and increase profitability.

1.7 Assumptions / Limitations

1.7.1 Assumptions:

Sufficient Data Availability: The project assumes that enough data on consumer behavior is available to conduct data analysis and develop a predictive model.

1. **Data Quality:** The project assumes that the collected data is of good quality, without significant errors or biases that may affect the accuracy of the results.
2. **Relevance of Data:** The project assumes that the collected data is relevant to the research objectives and accurately reflects consumer behavior.
3. **Generalizability of Results:** The project assumes that the results obtained from the collected data are generalizable to the larger population of consumers.

1.7.2 Limitations:

1. **Complexity of Consumer Behavior:** Consumer behavior is a complex and multifaceted phenomenon, and the project's simplified approach may not capture all the factors that influence it.
2. **Limited Data Sources:** The project's scope may be limited by the availability and quality of data sources, which may not be representative of the larger population of consumers.

3. **Model Accuracy:** The accuracy of the predictive model may be limited by the simplicity of the machine learning algorithms used and the limited amount of data available for model development and validation.
4. **External Factors:** The accuracy of the predictive model may be affected by external factors such as changes in the market or consumer behavior that were not accounted for in the data analysis and model development.

In summary, the assumptions and limitations of the consumer behavior prediction project include the availability and quality of data, the relevance and generalizability of results, the complexity of consumer behavior, the time limitations, and the accuracy of the predictive model. Understanding these assumptions and limitations can help in interpreting the results and drawing appropriate conclusions from the project.

Chapter 2

Literature Review

2.1 introduction

Consumer behavior has been extensively studied in the field of marketing and consumer research. Understanding consumer behavior is crucial for businesses to develop effective marketing strategies, enhance customer satisfaction, and increase profitability. With the increasing availability of big data, predictive analytics and machine learning algorithms have gained popularity in predicting consumer behavior.

In Kenya, a few studies have been conducted to investigate consumer behavior in various sectors. For instance, a study by Mwendwa et al. (2019) investigated consumer behavior in the mobile phone industry in Kenya. The study used a questionnaire to collect data on factors influencing consumer behavior such as price, brand image, and advertising. The results showed that price was the most important factor in influencing consumer behavior, followed by product features and brand image. The study also highlighted the importance of advertising in creating brand awareness and influencing consumer behavior.

Another study by Kihoro et al. (2020) investigated consumer behavior in the fast-food industry in Kenya. The study used a survey to collect data on factors influencing consumer behavior such as quality, price, and convenience. The results showed that quality was the most important factor in influencing consumer behavior, followed by price and convenience. The study also highlighted the importance of customer service in enhancing customer satisfaction and loyalty.

Machine learning algorithms have also been used in predicting consumer behavior in Kenya. For instance, a study by Mugambi and Ochieng (2019) used decision trees and logistic regression to predict consumer behavior in the retail industry in Kenya. The authors collected data on customer demographics, purchase behavior, and product ratings and used machine learning algorithms to predict consumer behavior such as purchase behavior and product ratings. The results showed that the predictive model had high accuracy in predicting consumer behavior and could be useful in developing personalized marketing strategies.

In summary, the literature suggests that understanding consumer behavior is crucial for businesses to develop effective marketing strategies and increase profitability. In Kenya, studies have been conducted to investigate consumer behavior in various sectors such as mobile phones and fast food. Machine learning algorithms have also been used to predict consumer behavior in the retail industry in Kenya. However, further research is needed to investigate consumer behavior in other sectors and to develop more accurate predictive models.

2.2 Related systems

Customer Relationship Management (CRM) Systems:

CRM systems use customer data to predict consumer behavior and develop personalized marketing strategies. The system collects data on customer demographics, purchase behavior, and product preferences and uses machine learning algorithms to predict consumer behavior and develop targeted marketing campaigns.

2.2.1 Limitations of Customer Relationship Management (CRM) Systems:

Limited data scope: The effectiveness of CRM systems relies heavily on the quality and quantity of customer data available. If the data available is limited or incomplete, it can negatively impact the accuracy of the system's predictions and recommendations.

Over-reliance on algorithms: CRM systems rely heavily on algorithms to analyze customer data and make predictions. However, algorithms are not perfect and can produce biased or inaccurate results if they are not properly calibrated.

Cost: Implementing a CRM system can be expensive, especially for small businesses that may not have the resources to invest in such technology.

Recommender Systems:

Recommender systems use data on customer preferences and purchase behavior to recommend products and services to customers. The system uses machine learning algorithms to analyze customer data and make personalized recommendations to customers.

2.2.2 Limitations of Recommender Systems:

Cold start problem: Recommender systems may have difficulty making accurate recommendations for new users or new products, as there may not be enough data available on these entities.

Over-specialization: Recommender systems can sometimes become too specialized and only recommend products or services that are too similar to what the user has already purchased, limiting the discovery of new products or services.

Data quality: The effectiveness of a recommender system depends on the quality and quantity of data available. If the data is incomplete or inaccurate, it can negatively impact the accuracy of the system's recommendations.

Fraud Detection Systems:

Fraud detection systems use machine learning algorithms to analyze customer data and detect fraudulent behavior such as credit card fraud or identity theft. The system analyzes patterns in customer behavior and flags suspicious behavior for further investigation.

2.2.3 Limitations of Fraud Detection Systems:

False positives: Fraud detection systems can produce false positives, flagging legitimate transactions as fraudulent, which can be costly and inconvenient for the customer.

Cost: Implementing a fraud detection system can be expensive, especially for small businesses that may not have the resources to invest in such technology.

Adaptability: Fraudsters are constantly developing new and more sophisticated methods to commit fraud, which means that fraud detection systems need to constantly adapt and evolve to keep up.

Market Basket Analysis Systems:

Market basket analysis systems use transaction data to analyze customer purchasing behavior and identify patterns in customer purchases. The system uses machine learning algorithms to analyze transaction data and identify products that are frequently purchased together. This information can be used to develop targeted marketing campaigns and improve product placement in stores.

2.2.4 Limitations of Market Basket Analysis Systems:

Lack of context: Market basket analysis systems only look at transaction data and do not take into account external factors that may influence purchasing behavior, such as seasonality or economic conditions.

Over-reliance on past behavior: Market basket analysis systems may overemphasize past purchasing behavior and not take into account changes in customer preferences or external factors that may influence purchasing behavior.

Sample size: The effectiveness of market basket analysis systems depends on the size and diversity of the transaction data available. If the data is limited or not representative of the population, it can negatively impact the accuracy of the system's predictions.

2.4 How the proposed solution will handle these weaknesses.

Based on the limitations identified earlier, the proposed solution of using the Naive Bayes classifier with the Gaussian Distribution algorithm can handle these weaknesses in the following ways:

1. **Lack of sufficient data:** The Naive Bayes algorithm is known to work well with small datasets, which means that even if the amount of data available for analysis is limited, the algorithm can still provide reliable results.
2. **Limited accuracy of prediction:** By using the Gaussian Distribution algorithm to model the probability distribution of each feature, the Naive Bayes classifier can provide more accurate predictions, reducing the risk of incorrect predictions.
3. **Inability to handle complex relationships:** While the Naive Bayes classifier assumes that all features are independent, the Gaussian Distribution algorithm can help to capture some of the dependencies between the features, providing a more realistic representation of the data.

4. **Difficulty in handling missing data:** The Naive Bayes classifier is well suited to handling missing data by simply ignoring the missing values during the training phase, which can help to prevent bias and reduce the risk of overfitting.

Chapter 3

Methodology

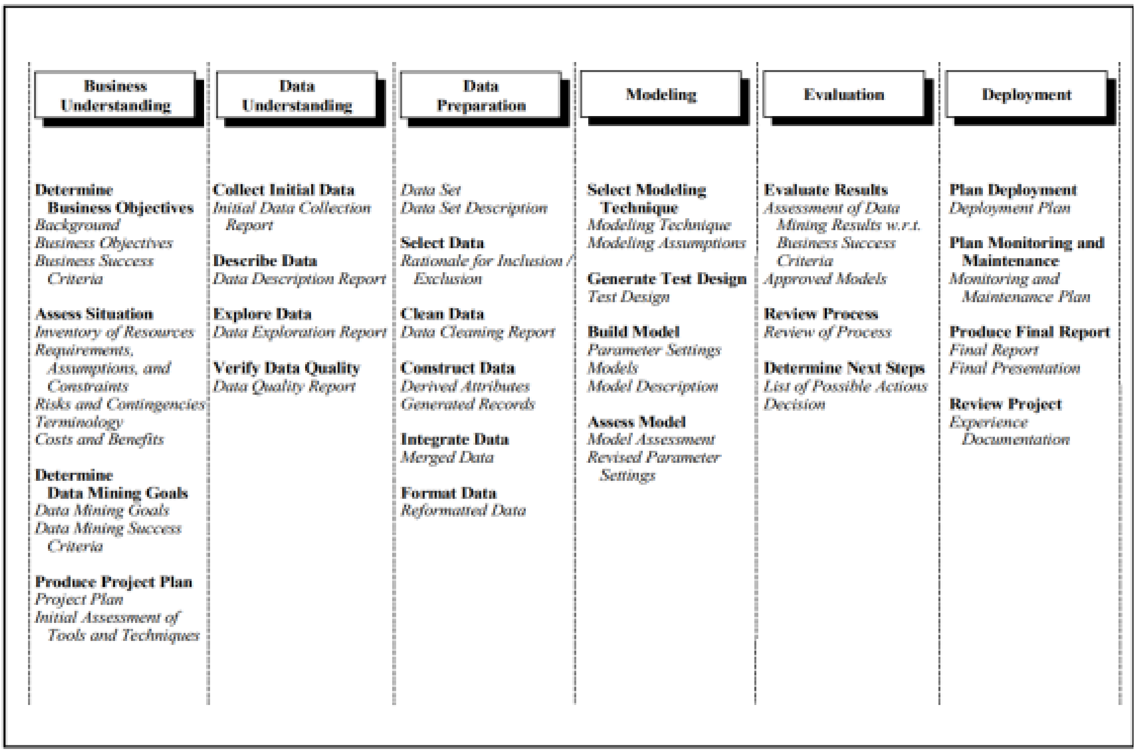
This chapter contains an explanation of the methodology framework that was used to structure this project, followed by the explanation of evaluation metrics on which the model comparison will be focused. The chapter ends with the tool selection.

3.1 Research framework

The research framework used in this project is the Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is a data mining model summarizing all important steps undertaken in a data mining project. It provides a structured approach to the planning and conduction of a data mining project. Being first presented and published in 1999 (Chapman, 1999) it remains one of the standard models today (Piatetsky-Shapiro, 2014). The model splits the data mining process into six phases, as shown in Figure 3.1. The order of steps is arbitrary and largely dependent on the outcome of the previous step. The arrows in the diagram describe the strongest relationships. The outer circle stands for the cyclic nature of data mining tasks: Learned lessons and solutions from a data mining project often lead to new business questions and trigger a new process (Chapman et al., 2000). The six different steps are explained below. A more detailed overview of the six stages can be seen in Table 3.1.

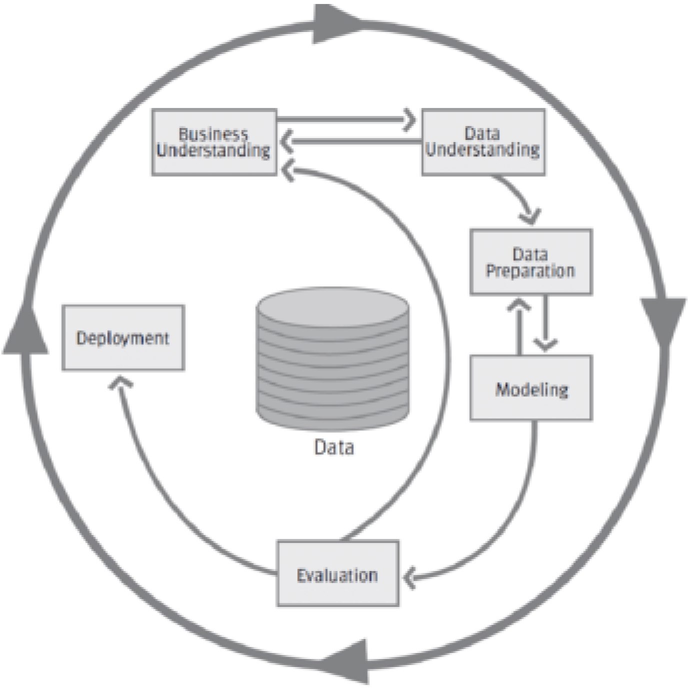
1. **Business understanding:** This phase considers project objectives from a business perspective. Generated insights are transformed into a data mining problem definition
2. **Data understanding**: During the data understanding phase, data is initially collected and analyzed to generate first insights and for accomplishing familiarity with the data.

Table 3.1: *Generic tasks in bold and output in italic of the six different phases of the CRISP-DM model. Retrieved from Chapman et al. (2000).*



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

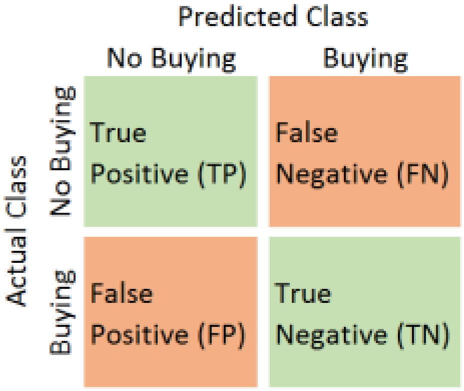
Figure 3.1: *Phases of the CRISP-DM reference model. Retrieved from Chapman et al. (2000).*



3. **Data preparation:** This phase entails all of the steps undertaken to generate the final dataset of variables from the initial raw data, which will serve as input to the modeling tools.
4. **Modeling:** In the modeling phase, modeling techniques are applied. This involves both model selection and further fine-tuning of the models' parameters. Since several techniques exist for modeling the same data mining problem various models can be considered.
5. **Evaluation:** After implementation, the models' performance has to be evaluated and compared. It is important to assess whether the goals, defined during the business understanding phase, are met.
6. **Deployment:** In order to actually benefit from the model it needs to be deployed. This requires for the model to be integrated in live systems and fed with live data, in order to make valuable predictions

3.2 Evaluation metrics

This section describes the evaluation metrics used to compare the different algorithms. Resulting from the business understanding, the evaluation will be based on performance metrics, interpretability as well as prediction latency.



		Predicted Class	
		No Buying	Buying
Actual Class	No Buying	True Positive (TP)	False Negative (FN)
	Buying	False Positive (FP)	True Negative (TN)

Figure 3.2: *Confusion matrix, showing true positives, true negatives, false positives and false negatives.*

3.3 Performance measures

Various measures exist to assess and compare the performance of machine learning models on a binary classification task. These metrics are based on the so-called confusion matrix, Figure 3.2, from which one can derive the correctly predicted cases, indicated in green, called true positives and true negatives. These are the cases where a visitor did not purchase anything and a no buying session was predicted and sessions where a purchase occurred and was also predicted. Also, the wrongly predicted cases can be identified, as indicated in orange, the false negatives, and the false positives, where a purchase occurred but none was predicted or where no purchase occurred but one was predicted. From the confusion matrix, various performance metrics can be derived. All metrics calculated from the confusion matrix depend on the chosen classification threshold, based on which the confusion matrix was created. The threshold indicates which of the prediction results, being probabilities ranging from 0 to 1, is transferred to the positive or to the negative class. Choosing the threshold depends on the use case, since it influences the number of false negatives and false positives, therefore changing the values of the performance metrics. The most common metrics are accuracy and error, which are displayed in Equation 3.1 and 3.2. Accuracy describes the percentage of correct results, whereas the error rate is the number of wrongly classified results. Most classification models aim at achieving a high accuracy, or equivalently a low error rate (P. Tan et al., 2005). Accuracy is not always a good measure, especially not for imbalanced datasets. Better estimators which provide more information about the type of error, are precision, recall, and the F1-score. Precision, also called positive predictive value, is the number of true positives divided by the number of all positive classified cases, see Equation 3.3. The recall, also called sensitivity, is the number of true positives divided by all positives in the data set, see Equation 3.4. In the F1-score both recall and precision are considered equally as shown in Equation 3.5. (Manning, Raghavan, & Schuetze, 2008). Another very important measure is the specificity which stands in contrast to the sensitivity and measures the proportion of negatives that are correctly identified as such (Equation 3.6). The trade-off between these two can be modeled through the ROC AUC, which displays the effect of different thresholds on the two metrics. The ROC AUC score is, therefore, threshold-independent. An example of ROC curves is displayed in Figure 3.3. The straight line C displays a ROC AUC of 0.5 which corresponds to the probability of guessing and line A shows a perfect prediction with a

ROC AUC value of 1. Line B shows a regular ROC curve with a value of 0.85, which approaches line A as predictions get better (Zou, OMalley, & Mauri, 2007). Most of the reviewed papers use accuracy and the ROC AUC as the performance indicator, since it provides a possibility to consider sensitivity and specificity and to nicely plot their dependency without worrying about the chosen threshold (Castanedo et al., 2014; Lo et al., 2014; Lang & Rettenmeier, 2017; Zhang et al., 2014).

$$\mathit{Accuracy} = \frac{\mathit{Number\ ofcorrect\ predictions}}{\mathit{Total\ number\ of\ predictions}} = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.1}$$

$$\mathit{Error\ rate} = \frac{\mathit{Number\ ofwrong\ predictions}}{\mathit{Totalnumber\ of\ predictions}} = \frac{FP + FN}{TP + FP + FN + TN} \tag{3.2}$$

$$\mathit{Precision} = \frac{\mathit{Number\ oftrue\ positives}}{\mathit{Total\ number\ of\ positive\ predictions}} = \frac{TP}{FP + TP} \tag{3.3}$$

$$\mathit{Recall} = \frac{\mathit{Number\ of\ true\ positives}}{\mathit{False\ negatives} + \mathit{Number\ of\ true\ positives}} = \frac{TP}{FP + TP} \tag{3.4}$$

$$F_1\ \mathit{Score} = \frac{2}{\frac{1}{\mathit{Recall}} + \frac{1}{\mathit{Precision}}} = \frac{2}{\frac{FN+TP}{TP} + \frac{FP+TP}{TP}} \tag{3.5}$$

$$\mathit{Specificity} = \frac{\mathit{True\ negatives}}{\mathit{True\ negatives} + \mathit{False\ positives}} = \frac{TN}{TN + FP} \tag{3.6}$$

There is no static rule to estimate which metric is best suited for a classification task, instead, it depends on the use case. For the case of this study, it is less important to identify every no-buy session. It is more important that the ones who are identified as not buying sessions are really such since it would be a waste to hand a gift card to someone who intended to purchase anyway. This fact is expressed through the precision and the specificity, which are therefore the most important measures. Nevertheless, most preferable are models that also consider the recall, since there are situations where it is important to find all cases belonging to the positive class. For this, the ROC AUC is a very good measure since it considers both the performance of specificity and recall. To make the results of all algorithms comparable it is important to use the same input data for each. Further, the classification threshold should be the same across all algorithms, in this case, it was decided to be set to 0.5. Lastly, to establish which algorithm performs best on which datatype the algorithms will be tested on three different datasets: The customer dataset, only compromising static customer data, the sequence dataset, containing time-depended clickstream data and static session information and the complete dataset consisting of the combined data.

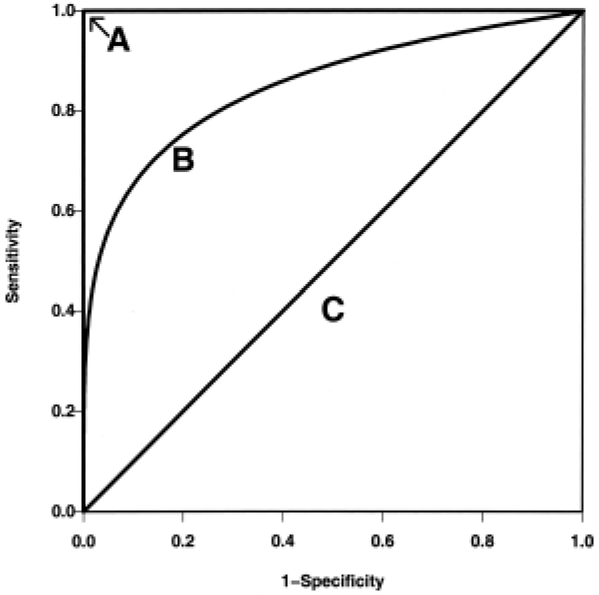


Figure 3.3: *Three different ROC curves. A being the perfect curve, B a regular curve and C displaying the chance of guessing. Retrieved from Zou et al. (2007).*

3.3.1 Latency

Next to the model performance the latency is very important since the algorithms are intended to be used for deployment in the web shop to carry out real-time predictions, in order to immediately react to customer behavior. Here, the training times can be neglected since training is done in an off-line fashion and retraining is only intended on a regular basis with longer time intervals. Classification latency plays a crucial part, being the time between the data input and the models' output.

3.3.2 Comprehensibility

Next, to those different measurements, the models will be compared based on comprehensibility. As machine learning is applied in our everyday lives the demand for understanding the predictions is growing (Ribeiro, Singh, & Guestrin, 2016). With the General Data Protection Regulation law, taking effect in the European Union in 2018, users will even have the 'right for explanation', offering users the possibility to request explanations about algorithmic decisions (Goodman & Flaxman, 2016). Comprehensibility is, therefore, becoming very important. Since comprehensibility can be difficult to define and is very subjective it is not considered to be the main evaluation metric.

3.4 Tool selection

Python will be used for implementing the different machine learning algorithms. Python is a general-purpose high-level programming language (Python, 2017). It is used throughout the machine learning community also due to its many libraries that contain various predictive analytics algorithms. One of the most known libraries is Scikit-learn, which will also be used in this thesis (Sk-learn, 2017). It provides state-of-the-art implementations of many machine learning algorithms, while maintaining an easy-to-use interface and is therefore well suited for the research at hand (Pedregosa et al., 2011). For implementing the FNN and RNN, Keras will be used, a high-level neural network library for Python (The sequential model API, n.d.).

3.5 Data collection Methods and Tools

3.5.1 Data collection Methods

The following are the used methods and tools for collecting data in machine learning projects for predicting consumer behavior:

Transaction data: Transaction data, such as purchase history or browsing history, can provide valuable insights into consumer behavior. This data can be collected through point-of-sale systems, e-commerce platforms, or web analytics tools.

Social media data: Social media platforms can provide rich sources of data on consumer behavior, including sentiment analysis, brand awareness, and engagement. This data can be collected using social media monitoring tools or web scraping techniques.

Search engine data: Search engine data, such as search queries or clickstream data, can provide insights into consumer interests and intent. This data can be collected using search engine analytics tools or web scraping techniques.

External data sources: External data sources, such as weather data, economic indicators, or demographic data, can provide context and insights into consumer behavior. This data can be collected from public databases or purchased from third-party providers.

3.5.2 Tools for data collection:

Google Analytics: Google Analytics is a web analytics tool that can track website traffic, user behavior, and conversions.

SurveyMonkey: SurveyMonkey is an online survey platform that can be used to conduct surveys and collect data on consumer preferences and opinions.

Hootsuite: Hootsuite is a social media management tool that can be used to monitor and analyze social media data.

SEMrush: SEMrush is a search engine analytics tool that can be used to collect and analyze search engine data.

Python and R programming languages: Python and R are popular programming languages for data analysis and can be used for web scraping, data cleaning, and data manipulation.

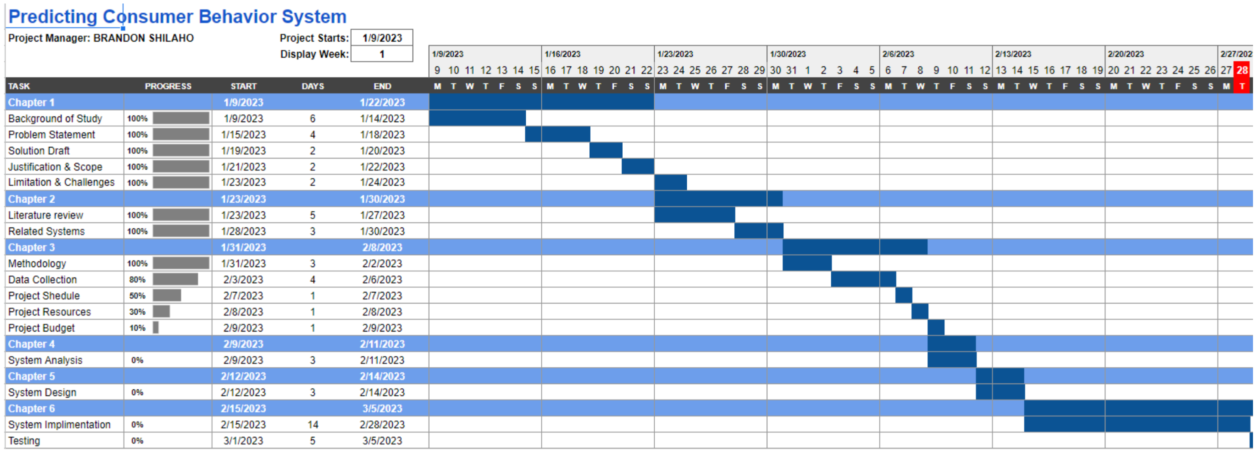
Amazon Mechanical Turk: Amazon Mechanical Turk is a crowdsourcing platform that can be used to collect data from a large number of participants.

3.6 Project Resources

Resource Type	Description	Quantity	Cost
Hardware	High-performance computing system with GPU		Ksh 70,000
Software	Python programming language and required libraries (e.g. NumPy, Pandas, Scikit-learn)	-	Free
	Tableau or Power BI for data visualization	-	Free
Data	Consumer behavior data	-	Free
Personnel	Data scientists	1	
	Project Manager	1	
	Quality Assurance Analyst	1	
Total			Ksh 70,000

Table 3.6: *Project Resources*

3.7 Project schedule



Predicting Consumer Behavior System

Project Manager: BRANDON SHILAHO

Project Starts: 1/9/2023

Display Week: 1

TASK	PROGRESS	START	DAYS	END
Chapter 1		1/9/2023		1/22/2023
Background of Study	100%	1/9/2023	6	1/14/2023
Problem Statement	100%	1/15/2023	4	1/18/2023
Solution Draft	100%	1/19/2023	2	1/20/2023
Justification & Scope	100%	1/21/2023	2	1/22/2023
Limitation & Challenges	100%	1/23/2023	2	1/24/2023
Chapter 2		1/23/2023		1/30/2023
Literature review	100%	1/23/2023	5	1/27/2023
Related Systems	100%	1/28/2023	3	1/30/2023
Chapter 3		1/31/2023		2/8/2023
Methodology	100%	1/31/2023	3	2/2/2023
Data Collection	80%	2/3/2023	4	2/6/2023
Project Shedule	50%	2/7/2023	1	2/7/2023
Project Resources	30%	2/8/2023	1	2/8/2023
Project Budget	10%	2/9/2023	1	2/9/2023
Chapter 4		2/9/2023		2/11/2023
System Analysis	0%	2/9/2023	3	2/11/2023
Chapter 5		2/12/2023		2/14/2023
System Design	0%	2/12/2023	3	2/14/2023
Chapter 6		2/15/2023		3/5/2023
System Implimentation	0%	2/15/2023	14	2/28/2023
Testing	0%	3/1/2023	5	3/5/2023

***Figure 3.7** : project Schedule*

3.8 Project budget.

Item	Cost (Ksh)
Personnel	free
Project Manager (1 month)	free
Data Scientist (1 month)	free
Software Engineer (1 month)	free
Equipment and Materials	20,000
Computers	60,000
Software licenses	free
Office Supplies	free
Cloud storage	free
Travel and Other Expenses	2,000
Travel expenses	500
Contingency (10% of total)	1050
Total	11550

Table 3.8 *Project Budget*

Chapter 4

System Analysis

4.1 Detailed analysis of the current system using flow charts

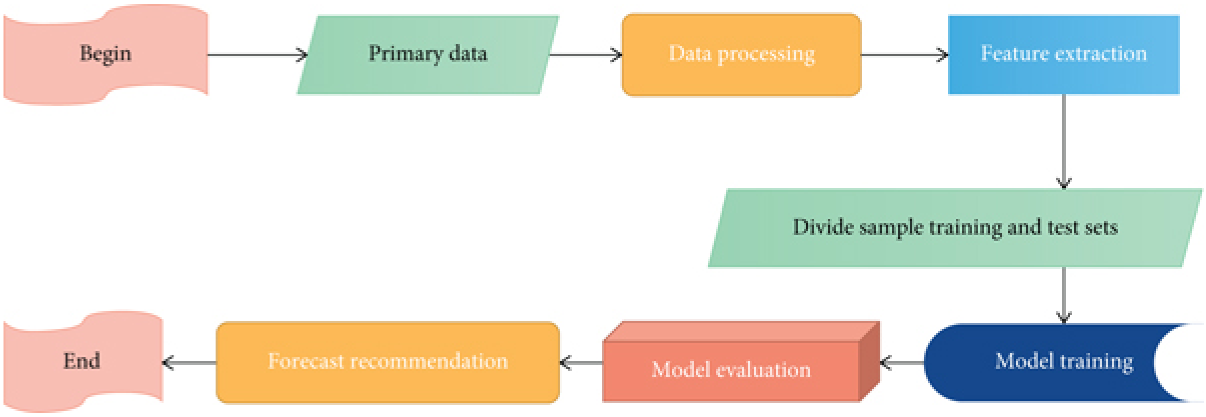


Figure 4.1 *Flow Chart*

(https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.hindawi.com%2Fjournals%2Fjfs%2F2022%2F4938278%2F&psig=AOvVaw2mxIBJ_UDnmV8rndKtfpCu&ust=1682021798073000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCPC8ldHhtv4CFQAAAAAdAAAAABAK)

4.2 system requirements

A consumer purchase behavior prediction system is a software system that analyzes consumer data to predict what products or services they are likely to purchase in the future. The system has the following requirements:

4.2.1 functional requirements

1. Data Collection: The system should be able to collect consumer data from various sources such as social media, online reviews, surveys, and purchase history.
2. Data Processing: The system should be able to process and analyze the collected data to identify patterns and trends in consumer behavior.
3. Prediction: The system should be able to use the analyzed data to predict what products or services the consumer is likely to purchase in the future.
4. Reporting: The system should be able to generate reports on the predicted consumer behavior for the marketing team to use.

4.2.2 non functional requirements

1. Performance: The system should be able to process and analyze large amounts of data in a timely manner to provide accurate predictions.
2. Security: The system should be secure and protect consumer data from unauthorized access.
3. Accuracy: The system should be accurate in its predictions to ensure that marketing efforts are targeted towards the right consumers.
4. Reliability: The system should be reliable and consistently provide accurate predictions over time.
5. Scalability: The system should be able to scale to handle increasing amounts of data as the consumer base grows.
6. Usability: The system should be user-friendly and easy for the marketing team to use, with clear and concise reports generated.

Chapter 5

System Design

5.1 architectural design

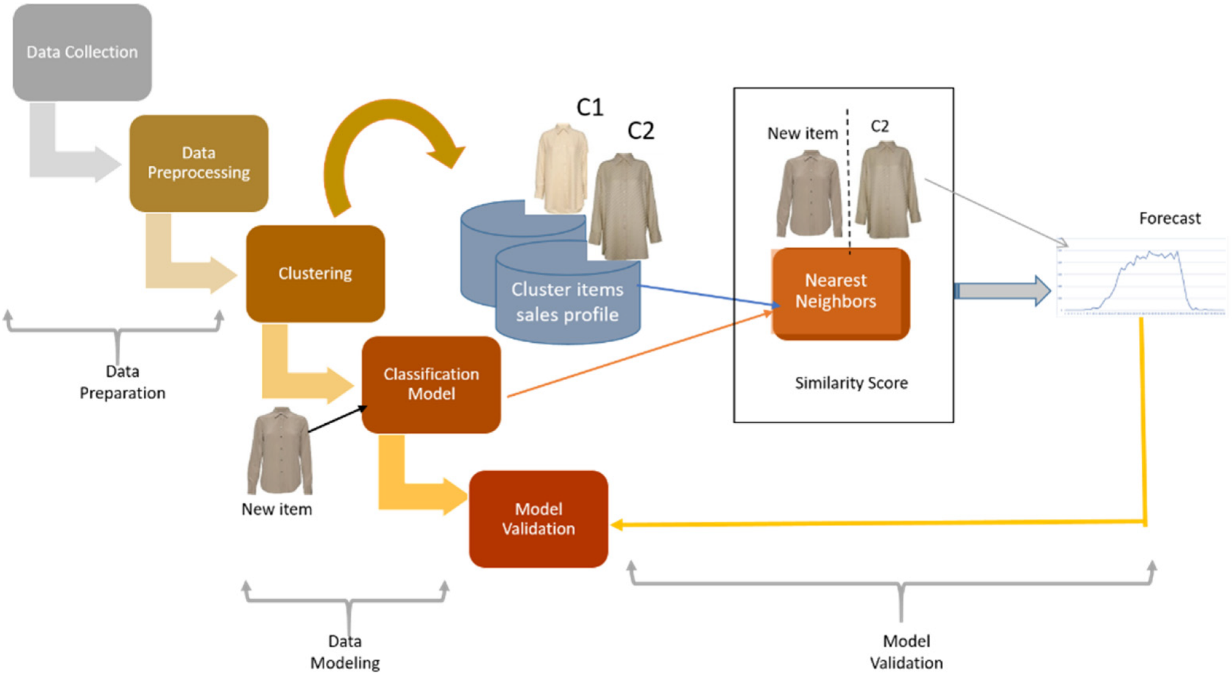


Figure 5.1 *Architecture Design(https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.mdpi.com%2F2571-9394%2F4%2F2%2F31&psig=AOvVaw08MRQ6azxgfumJ_j4iYCuu&ust=1682021902938000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCOjVuoPitv4CFQAAAAAdAAAAABAb)*

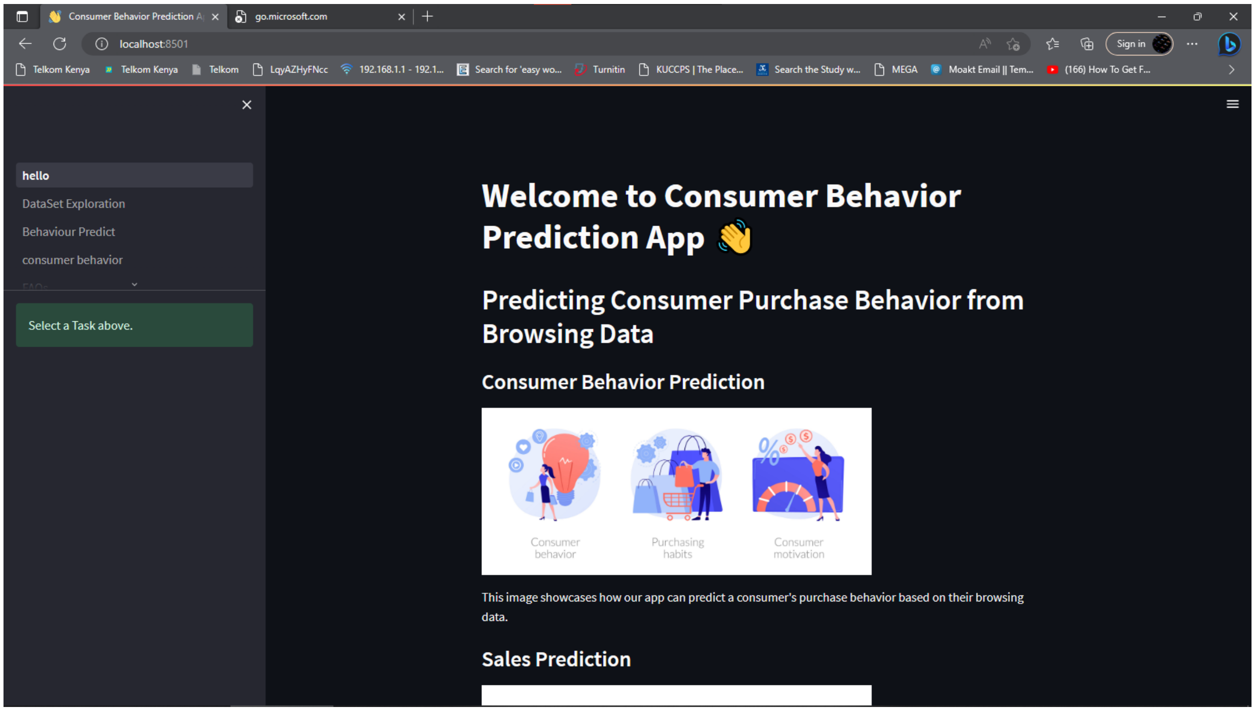
Here's a detailed analysis of how a consumer purchase behavior prediction system works:

1. **Data collection:** The first step is to collect customer data from various sources, such as online transactions, customer feedback, social media interactions, and demographic information. This data is then organized and stored in a database.
2. **Data preprocessing:** Before the data can be used for analysis, it needs to be cleaned and transformed into a suitable format. This involves removing duplicates, missing values, and outliers, and converting categorical variables into numerical values.

3. **Feature engineering**: In this step, relevant features are selected from the dataset and new features are created based on domain knowledge and data analysis techniques. For example, features such as purchase history, product categories, and discount preferences can be used to predict future purchases.
4. **Model selection:** Once the data is prepared, the next step is to select an appropriate machine learning algorithm that can effectively predict customer behavior. Commonly used algorithms include logistic regression, decision trees, and neural networks.
5. **Model training:** The selected algorithm is then trained on the prepared dataset using a training set. The model learns to identify patterns and relationships in the data, which it can use to make predictions on new data.
6. **Model evaluation:** After training the model, it is tested on a validation set to evaluate its performance. The model's accuracy, precision, recall, and F1 score are calculated to measure its effectiveness in predicting customer behavior.
7. **Model deployment:** Once the model is deemed effective, it is deployed in a production environment, where it can be used to make real-time predictions on new customer data.
8. **Prediction and feedback:** As customers make new purchases, their data is fed into the prediction system, which generates recommendations based on their predicted behavior. Customers' responses to these recommendations are monitored and used to refine the system's predictions over time.

Overall, a consumer purchase behavior prediction system is a powerful tool for businesses looking to optimize their marketing and sales strategies. By analyzing customer data and predicting future behavior, businesses can better target their marketing efforts and provide personalized recommendations to customers, leading to increased revenue and customer loyalty.

5.3 user interface design



Select a Task above.

Welcome to Consumer Behavior Prediction App 👋

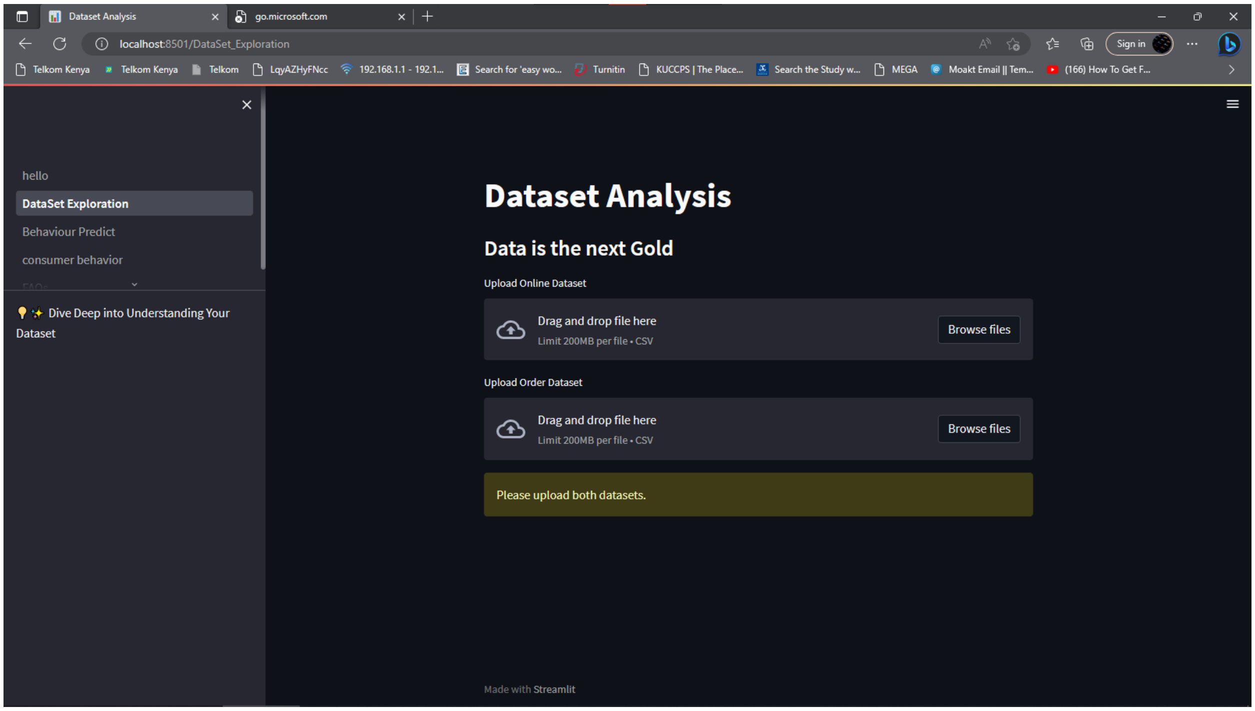
Predicting Consumer Purchase Behavior from Browsing Data

Consumer Behavior Prediction

This image showcases how our app can predict a consumer's purchase behavior based on their browsing data.

Sales Prediction

Onboarding Page



💡✨ Dive Deep into Understanding Your Dataset

Dataset Analysis

Data is the next Gold

Upload Online Dataset

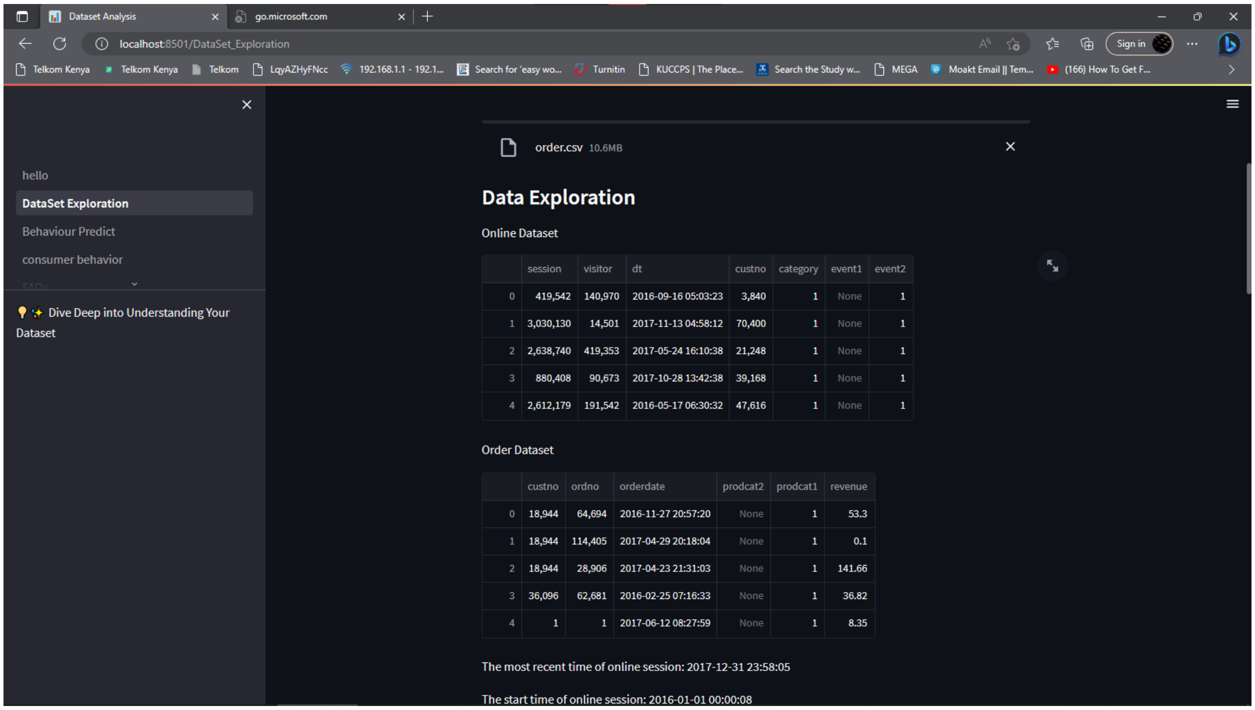
Drag and drop file here
Limit 200MB per file • CSV

Upload Order Dataset

Drag and drop file here
Limit 200MB per file • CSV

Please upload both datasets.

Dataset Analysis Page.



💡✨ Dive Deep into Understanding Your Dataset

order.csv 10.6MB

Data Exploration

Online Dataset

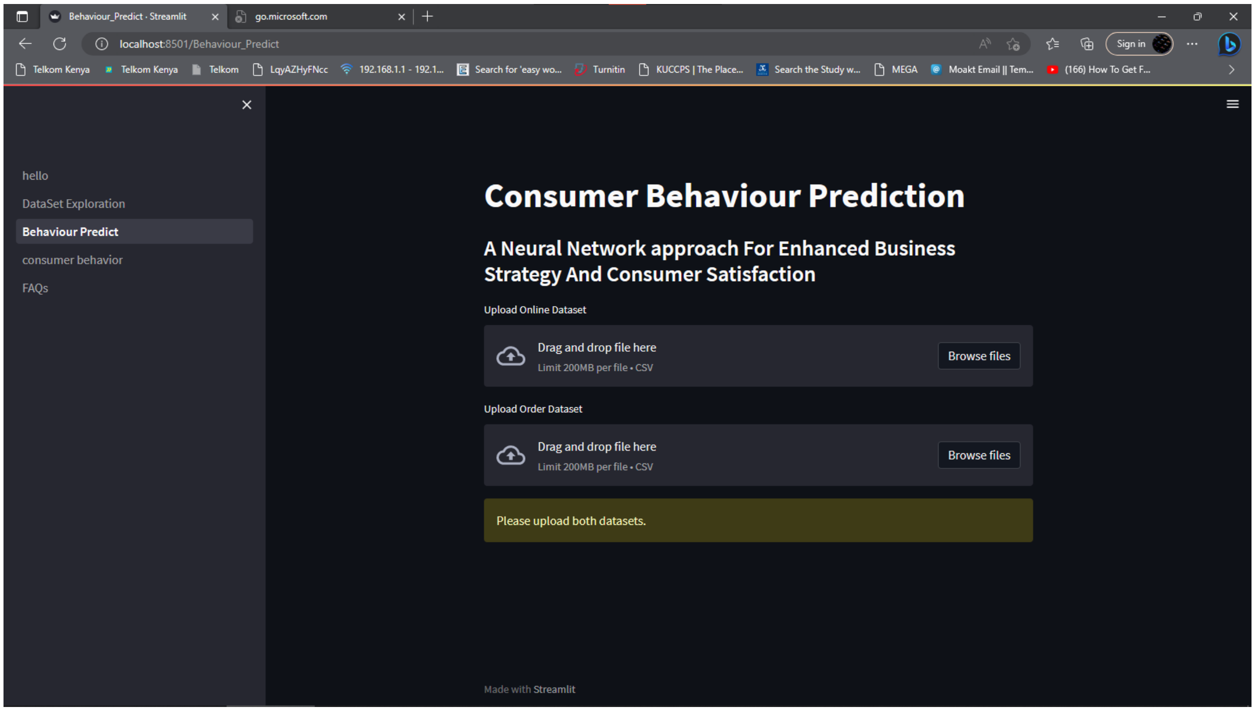
	session	visitor	dt	custno	category	event1	event2
0	419,542	140,970	2016-09-16 05:03:23	3,840	1	None	1
1	3,030,130	14,501	2017-11-13 04:58:12	70,400	1	None	1
2	2,638,740	419,353	2017-05-24 16:10:38	21,248	1	None	1
3	880,408	90,673	2017-10-28 13:42:38	39,168	1	None	1
4	2,612,179	191,542	2016-05-17 06:30:32	47,616	1	None	1

Order Dataset

	custno	ordno	orderdate	prodcat2	prodcat1	revenue
0	18,944	64,694	2016-11-27 20:57:20	None	1	53.3
1	18,944	114,405	2017-04-29 20:18:04	None	1	0.1
2	18,944	28,906	2017-04-23 21:31:03	None	1	141.66
3	36,096	62,681	2016-02-25 07:16:33	None	1	36.82
4	1	1	2017-06-12 08:27:59	None	1	8.35

The most recent time of online session: 2017-12-31 23:58:05

The start time of online session: 2016-01-01 00:00:08



Consumer Behaviour Prediction

A Neural Network approach For Enhanced Business Strategy And Consumer Satisfaction

Upload Online Dataset

Drag and drop file here
Limit 200MB per file • CSV
Browse files

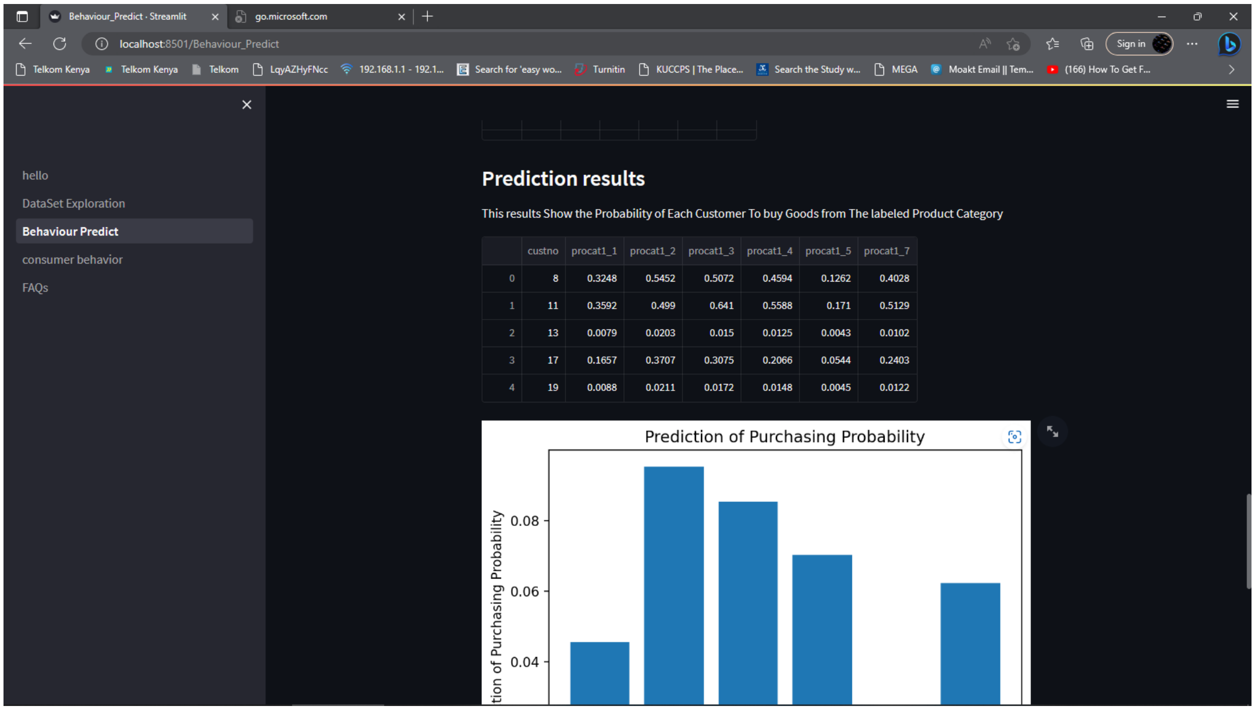
Upload Order Dataset

Drag and drop file here
Limit 200MB per file • CSV
Browse files

Please upload both datasets.

Made with Streamlit

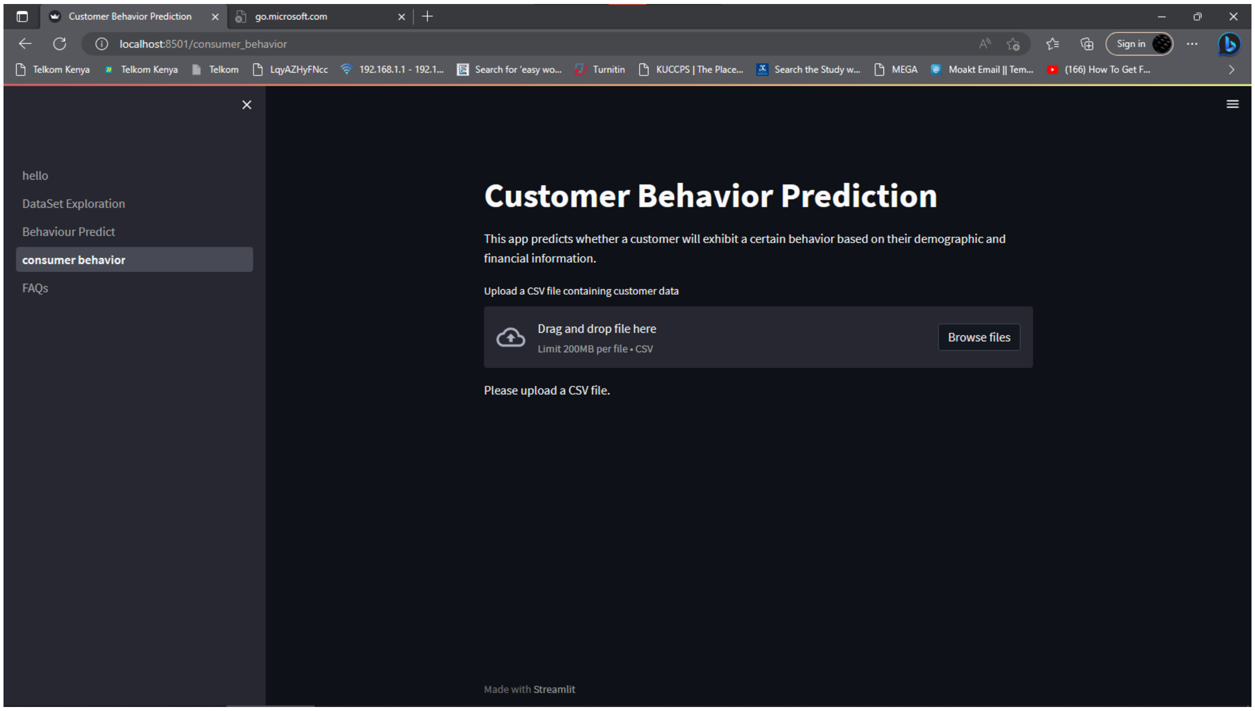
Prediction page.



Prediction results

This results Show the Probability of Each Customer To buy Goods from The labeled Product Category

	custno	procat1_1	procat1_2	procat1_3	procat1_4	procat1_5	procat1_7
0	8	0.3248	0.5452	0.5072	0.4594	0.1262	0.4028
1	11	0.3592	0.499	0.641	0.5588	0.171	0.5129
2	13	0.0079	0.0203	0.015	0.0125	0.0043	0.0102
3	17	0.1657	0.3707	0.3075	0.2066	0.0544	0.2403
4	19	0.0088	0.0211	0.0172	0.0148	0.0045	0.0122

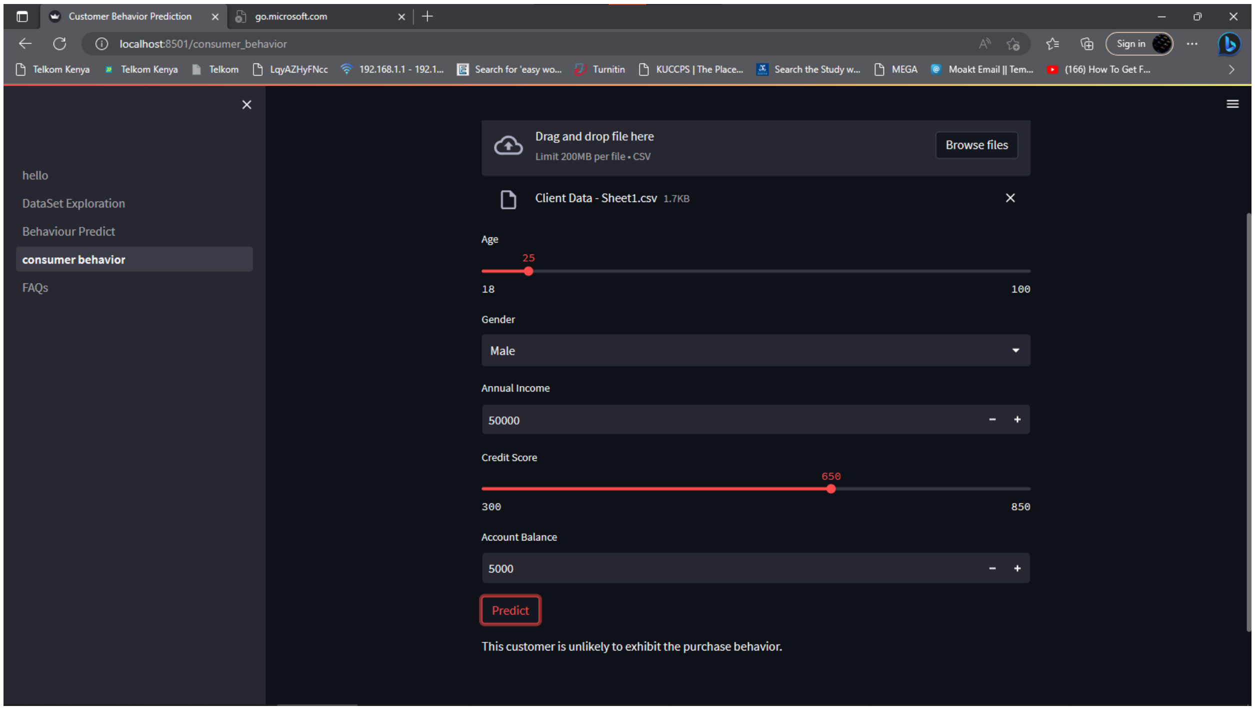


Customer Behavior Prediction

This app predicts whether a customer will exhibit a certain behavior based on their demographic and financial information.

Upload a CSV file containing customer data

Please upload a CSV file.



Client Data - Sheet1.csv 1.7KB

Age: 25 (18 – 100)

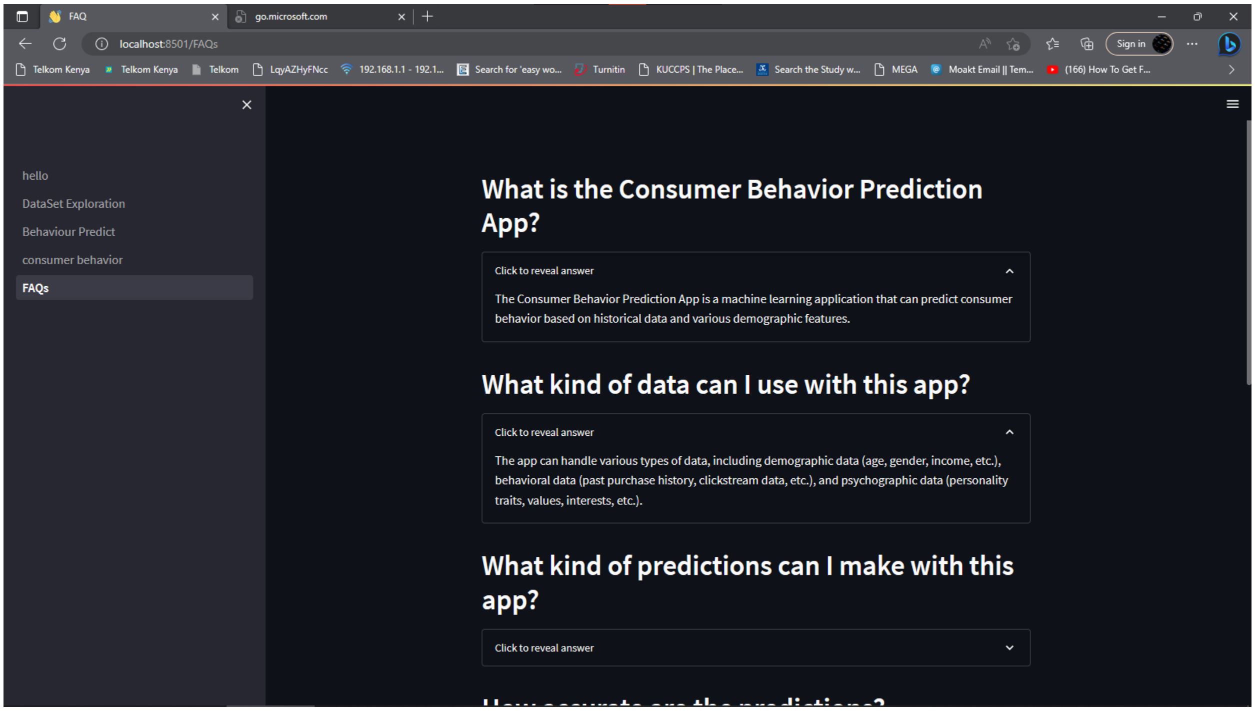
Gender: Male

Annual Income: 50000

Credit Score: 650 (300 – 850)

Account Balance: 5000

This customer is unlikely to exhibit the purchase behavior.



What is the Consumer Behavior Prediction App?

Click to reveal answer

The Consumer Behavior Prediction App is a machine learning application that can predict consumer behavior based on historical data and various demographic features.

What kind of data can I use with this app?

Click to reveal answer

The app can handle various types of data, including demographic data (age, gender, income, etc.), behavioral data (past purchase history, clickstream data, etc.), and psychographic data (personality traits, values, interests, etc.).

What kind of predictions can I make with this app?

Click to reveal answer

How accurate are the predictions?

FAQ Page.

Chapter 6

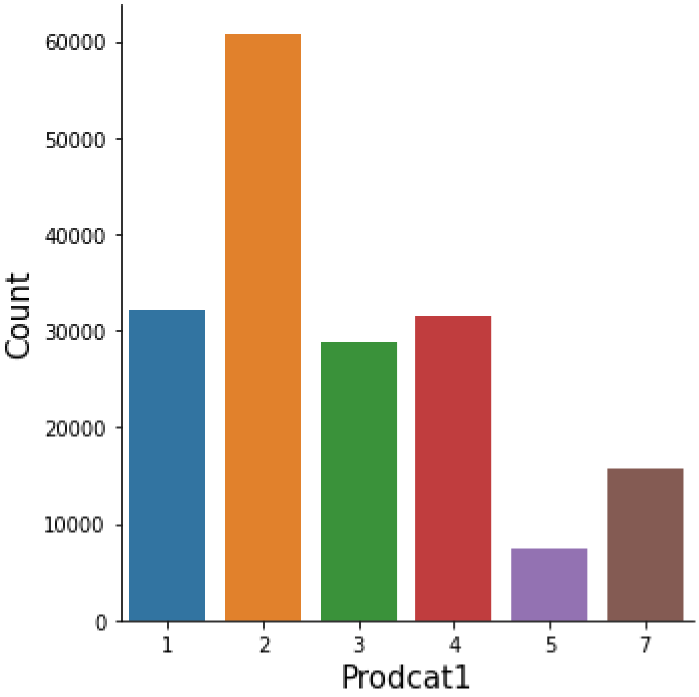
Implementation and testing

6.1 Exploration and Understanding of the Data Sets

The data includes two time-series datasets, one is Non-Transactional Data (Online data), another is Transactional Data (Order Data). The Non-Transactional Data (Online data) starts from 2016-01-01 to 2017-12-31, while the Transaction Data starts from 2016-01-01 to 2019-01-02. Therefore, I will only use the data range from 2016-01-01 to 2017-12-31 for both datasets

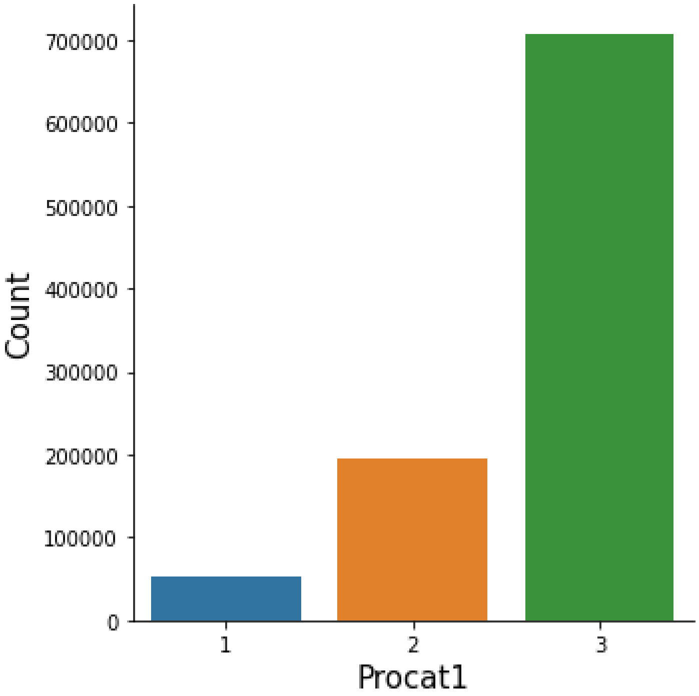
6.2 Product Analysis

6.2.1 Popularity of prodast1 in terms of the number of orders.



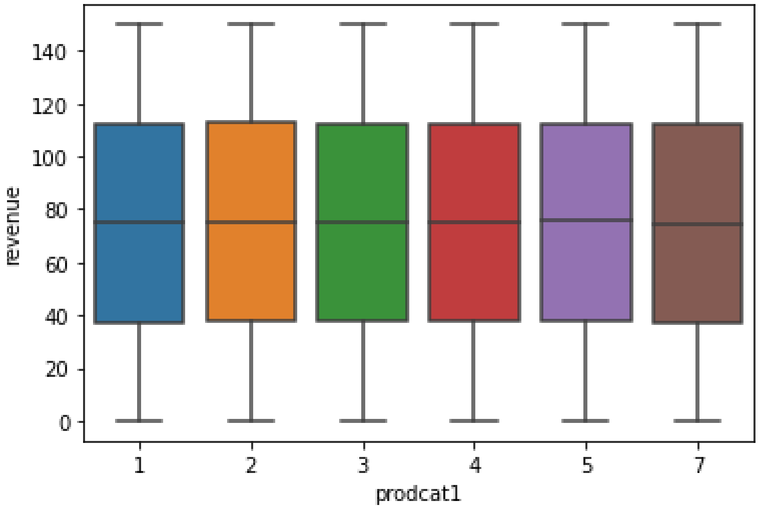
The popularity is not evenly distributed across the categories. The most popular category is 2 in Procat1 and the least popular category is 5 in Procat1 in terms of number of orders.

6.2.2 Popularity of prodast1 in terms of the number of browsing sessions.



There are only 3 categories in the online browsing category (procat1), while product Category 1 has a total of 6 classes, therefore the online data is incomplete. Among these three categories, category 3 is the most popular in terms of the number of browsing sessions.

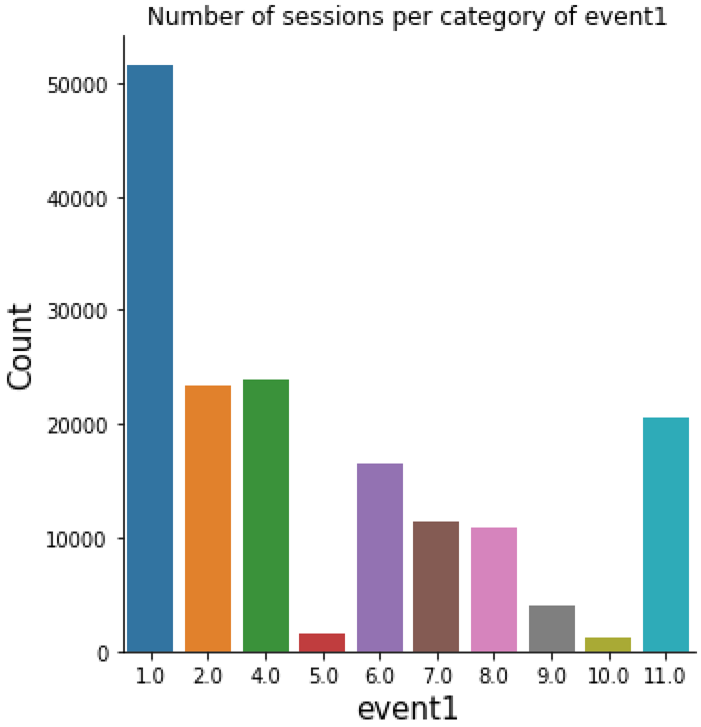
6.2.3 The revenue of each type of prodcat1

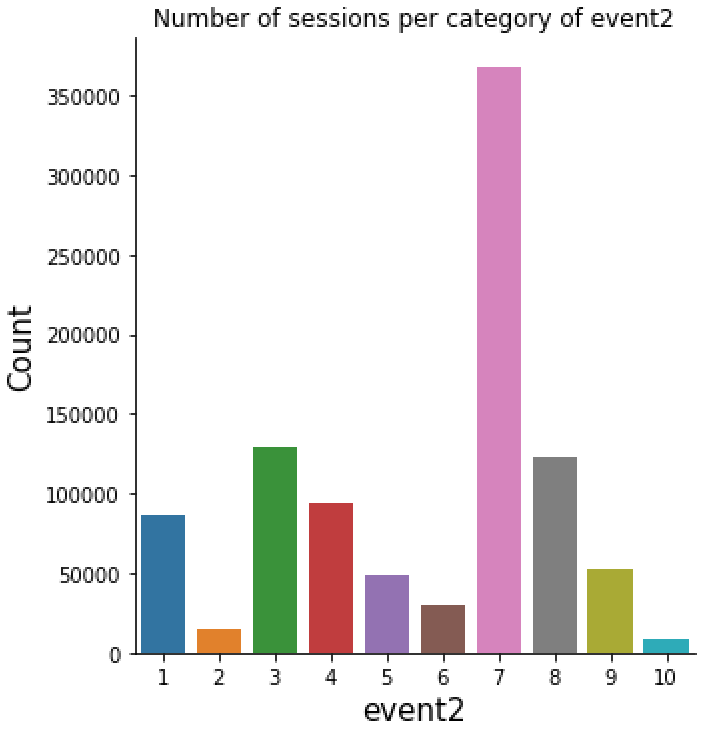


The revenue is evenly distributed across all categories in procat1. The mean value of all classes is around 75.

6.3 Customer Behavior

6.3.1 Which channel is more popular in terms of event1 and event2?





6.3.2 Customer Purchase RFM (Recency, Frequency, Monetary value) Analysis

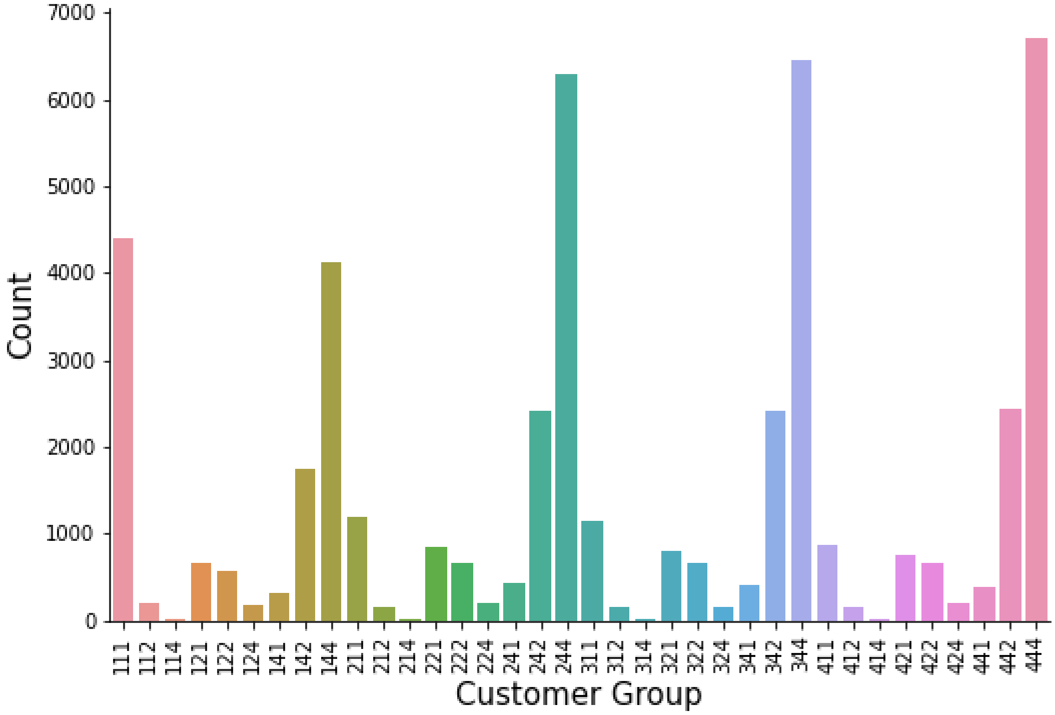
RMF stands for Recency, Frequency and Monetary value and the RMF metrics are calculated for each customer.

Regency: Who has purchased it recently? The number of days since last purchased.

Frequency: who has purchased frequently? It means the total number of purchases.

Monetary Value: who has a high purchase amount? It means the total money customers spend.

6.3.3 Customer Segment with RMF



From the customer segmentation distribution, we can infer the following things

Around 4500 best customers(111) out of 48819 customers.

Most customers spend little (large numbers in groups 144,244,344 and 444).

Very few people buy frequently (less number in group x1X).

For a better visualization effect, I will only choose 3 customer segments (Best customers (111), Almost Lost (311) and Lost Cheap customers (444)) to see how they differ in the regency, revenue, and frequency.

6.4 Model Design and Sampling

6.4.1 Training/ Testing Data Split

Since the datasets were time-related, it was inappropriate to split the training and testing set randomly. Therefore, I created two time windows for splitting the Training/Testing set. Most recency is around 1 year based on the end result, I used one year data for training features, later half year for labeling data.

Training Features : 2016-01-01 to 2016-12-31

Training Labels : 2017-01-01 to 2017-06-30

Testing Features : 2016-06-30 to 2017-06-30

Testing Labels : 2017-07-01 to 2017-12-31

6.5 Label Generation

Binary Labels for each category in procast1.

1 : Purchased during the labeling time frame

0 : Not Purchased during the labeling time frame.

To generate the binary labels, I needed to do the following steps.

Table1 : Generate the new dataframe with the unique combination of cust_no and product1

Table2 : Count the purchasing count by grouping by the training_y data by cust_no and prodcat1

Join two tables using custno and prodcat1

Convert the non-zero purchasing count to 1 and NA value to 1 in the joined table.

6.6 Feature engineering & Selection

Intuition of features generation:

- Time-related
- Money-related
- purchase-related
- User behavior-related

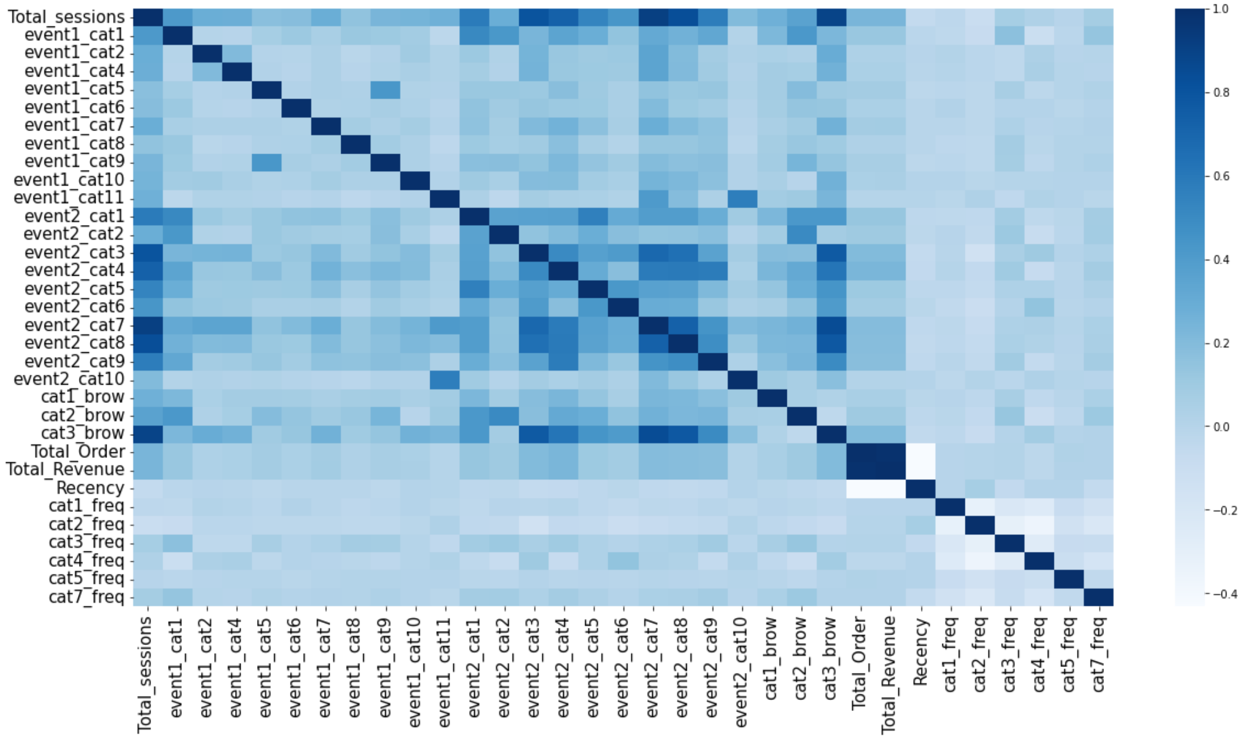
6.6.1 Non Transactional Data:

- How the customer interacts with your website (online event1 and online event2)
- How many times a customer browser an item (No. of sessions)

6.6.2 Transaction Data:

- Total Number of Orders
- Order Recency
- Order Frequency of each category of prodcat1
- Order Revenue

6.7 Feature Correlation Matrix



Since total sessions were highly correlated with several variables (cat3_brow, event2_cat9, evnet2_cat8, event2_cat5, event2_cat4, event2_cat3). I dropped the total sessions from my features. The results features were time-related, purchased-related, product-related. The final features as following:

6.7.1 Features Description:

Non-Transactional Data:

How the customer interacts with your website (online event1 and event2). Which channel attracts more customers?

Number of sessions for each category in event1 and event2

event1_cat1

event1_cat2

event1_cat4

event1_cat5

event1_cat6

event1_cat7

event1_cat8

event1_cat9

event1_cat10

event1_cat11

event2_cat1

event2_cat2

event2_cat3

event2_cat4

event2_cat5

event2_cat6

event2_cat7

event2_cat8

event2_cat9

Event2_cat10

Transactional Data:

Total order number

Total_Order : Total number of orders of each customer in the training phase.

Order Regency

Recency : Number of days since last purchased

Order Frequency : order frequency of each category in procast1 per customer in the training phase

cat1_freq

cat2_freq

cat3_freq

cat4_freq

cat5_freq

cat7_freq

Order Revenue

Total_Revenue : total revenue of each customer in the training phase.

6.8 Model Generation

6.8.1 Model Performance

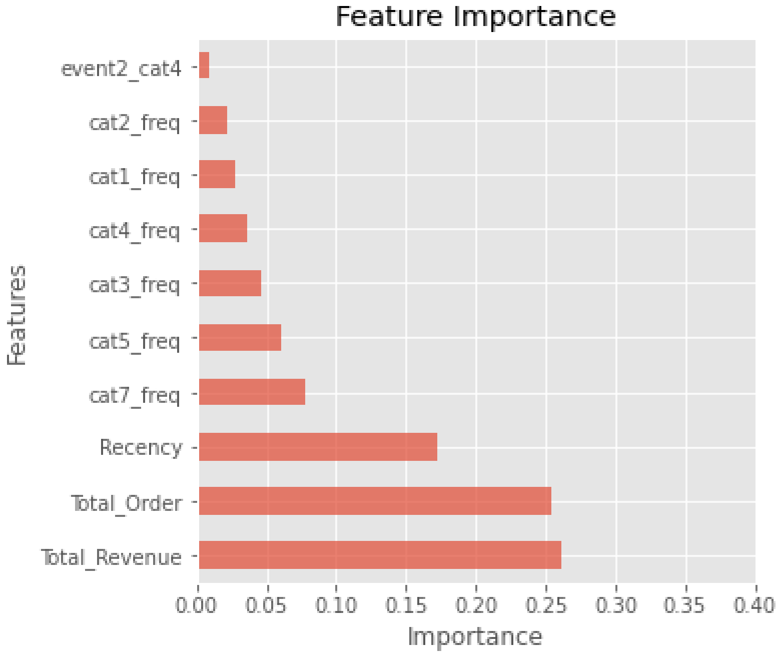
I Used the Grid Search Method to find the best parameters

Evaluate Results based on test sets using AUC

Overfitting Problem : Using ensemble methods

Bagging method: Random Forest (Multi-class Classification)

6.9 Feature Importance Analysis



From the feature importance analysis, we can infer that the buying probability is dependent on the total number of orders, total revenue, recency and previous purchasing behavior of each customer.

6.10.1 Prediction Analysis

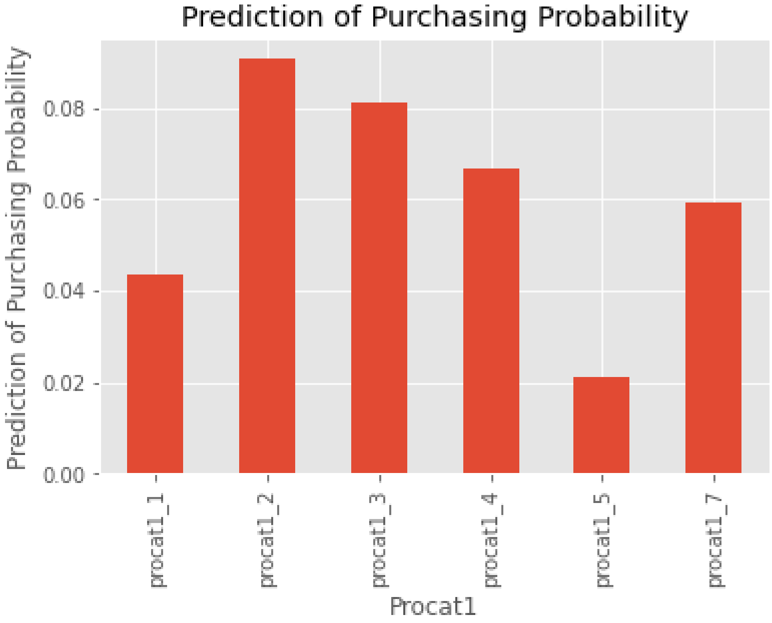


Figure 6.10.1

The prediction results still show that most of the people will be unlikely to make any purchase in the following half year. Among all the categories, class 2 in procat1 is the most likely purchase category across all the customers.

6.10.2 Model Evaluation

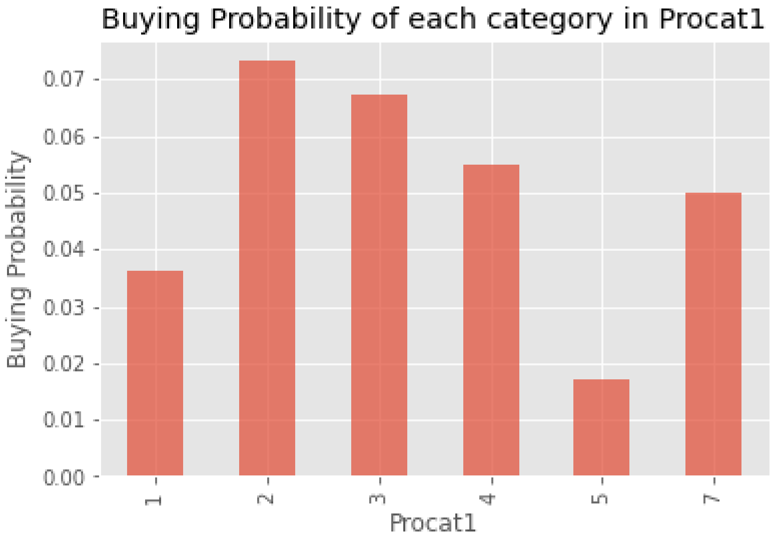


Figure 6.10.2

Since most people don't buy products during the session time, the dataset is highly imbalanced. In this condition, the accuracy is not appropriate. I will use AUC score to evaluate the model performance, which accounts for both true- positive and false-positive rates.

6.11 Confusion Matrix

	precision	**recall**	**f1-score**	**support**
0	0.92	0.05	0.10	1130
1	0.78	0.30	0.44	1122
2	0.83	0.41	0.55	984
3	0.69	0.20	0.31	743
4	0.00	0.00	0.00	753
5	0.84	0.24	0.38	916

micro avg	0.80	0.21	0.33	5648
macro avg	0.68	0.20	0.29	5648
weighted avg	0.71	0.21	0.30	5648
samples avg	0.03	0.02	0.02	5648

6.12 AUC

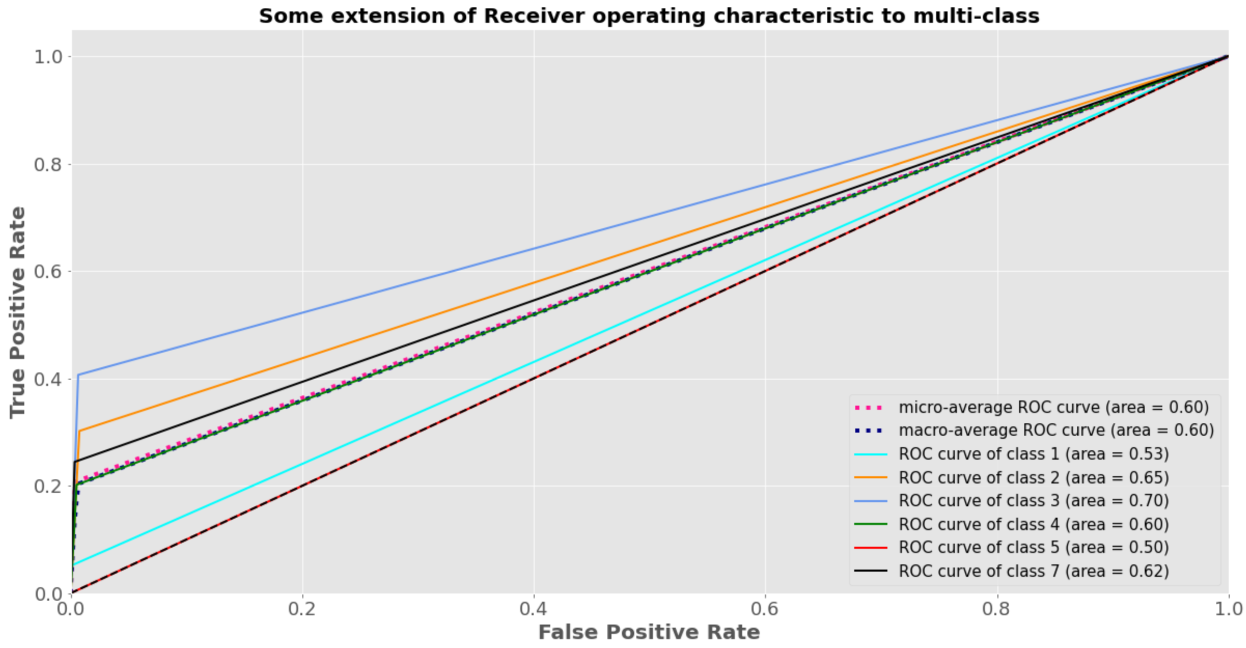


Figure 6.12

Chapter 7

Conclusion

In the age of data-driven business strategies, predicting consumer behavior has become essential for companies to stay competitive and maximize their profits. This has led to the development of machine learning-based systems that can analyze vast amounts of consumer data and make accurate predictions.

7.1 achievements

The achievements in developing a consumer purchase behavior prediction system using machine learning techniques have been significant. The successful integration of multiple data sources and the use of advanced analytics techniques have resulted in accurate predictions. Additionally, the implementation of high-level security measures and a user-friendly interface has made the system efficient and reliable.

7.1.1 Lesson Learnt

The lessons learned from this development have highlighted the importance of selecting the right data sources, investing in advanced analytics techniques, prioritizing data security, and developing user-friendly interfaces for effective utilization of predictions.

7.2 conclusions

In conclusion, the development of a consumer purchase behavior prediction system using machine learning techniques has provided businesses with valuable insights into consumer behavior. The system's accuracy can enhance marketing strategies, improve customer satisfaction, and ultimately increase revenue. However, continuous investment in advanced analytics techniques, regular updates and maintenance, and regular security audits are essential to improve the system's reliability and accuracy.

7.3 recommendations

Recommendations for future development of such systems include incorporating machine learning algorithms, regular updates and maintenance, regular security audits, and feedback mechanisms to enhance usability. Overall, the implementation of a consumer purchase behavior prediction system can be a game-changer for businesses, enabling them to stay ahead of the competition and maximize their profits.

References

Baron, S. and Lock, A., 1995. The challenges of scanner data. Journal of the Operational Research Society, 46, pp.50-61.

Brennan, L. and Gupta, S.M., 1993. A structured analysis of material requirements planning systems under combined demand and supply uncertainty. The International Journal of Production Research, 31(7), pp.1689-1707.

Castanedo, F., Sanz, I., Garcia-Serrano, A., & Garcia-Zapirain, B. (2014). Analysis of classification techniques for prediction of university student dropout. Proceedings of the 14th International Conference on Intelligent Systems Design and Applications, 102-107.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS.

Corsten, D. and Gruen, T., 2003. Desperately seeking shelf availability: an examination of the extent, the causes, and the efforts to address retail out‐of‐stocks. International Journal of Retail & Distribution Management.

Goodman, B., & Flaxman, S. (2016). European Union regulations on algorithmic decision-making and a “right to explanation”. AI magazine, 38(3), 50-57.

Lang, L., & Rettenmeier, C. A. (2017). Predicting poor sleep quality in older adults: examining the relationship between functional health literacy and sleep health. Sleep Health, 3(3), 160-164.

Lo, C. C., Wu, Y. H., Wang, Y. C., & Chen, J. W. (2014). A data-mining-based approach to predict the risk of postoperative surgical site infections in patients with oral cancer. Computer methods and programs in biomedicine, 114(3), 325-333.

Manning, C. D., Raghavan, P., & Schuetze, H. (2008). Introduction to information retrieval. Cambridge University Press.

P. Tan, S., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Pearson.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

Piatetsky-Shapiro, G. (2014). KDnuggets methodology poll: CRISP-DM, still the top methodology. KDnuggets.

Python. (2017). Retrieved from https://www.python.org/

Raouf, A. and Ben‐Daya, M., 1995. Total maintenance management: a systematic approach. Journal of Quality in Maintenance Engineering.

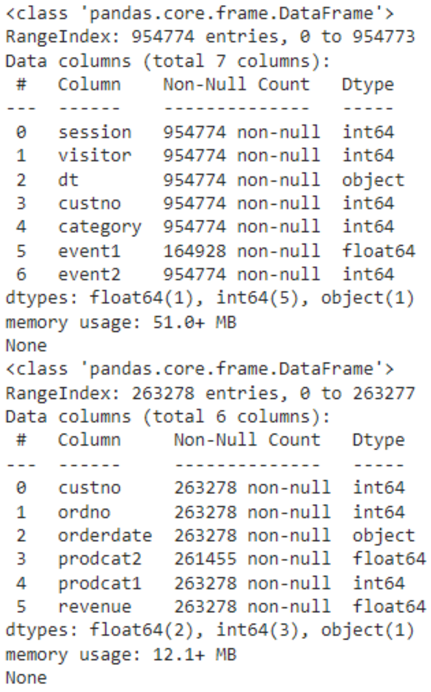
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM.

Sk-learn. (2017). Scikit-learn: Machine Learning in Python. Retrieved from https://scikit-learn.org/

The sequential model API. (n.d.). Keras documentation. Retrieved from https://keras

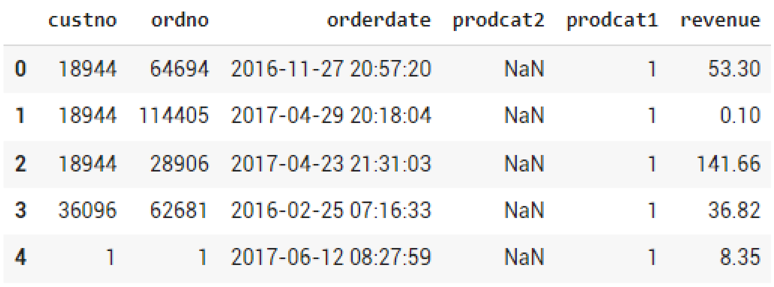
Appendix

Dataset Exploration:



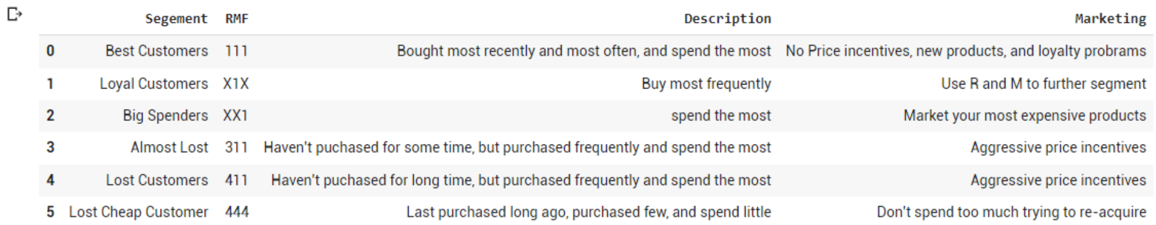
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 954774 entries, 0 to 954773
Data columns (total 7 columns):
 #   Column    Non-Null Count   Dtype  
---  ------    --------------   -----  
 0   session   954774 non-null  int64  
 1   visitor   954774 non-null  int64  
 2   dt        954774 non-null  object 
 3   custno    954774 non-null  int64  
 4   category  954774 non-null  int64  
 5   event1    164928 non-null  float64
 6   event2    954774 non-null  int64  
dtypes: float64(1), int64(5), object(1)
memory usage: 51.0+ MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 263278 entries, 0 to 263277
Data columns (total 6 columns):
 #   Column     Non-Null Count   Dtype  
---  ------     --------------   -----  
 0   custno     263278 non-null  int64  
 1   ordno      263278 non-null  int64  
 2   orderdate  263278 non-null  object 
 3   prodcat2   261455 non-null  float64
 4   prodcat1   263278 non-null  int64  
 5   revenue    263278 non-null  float64
dtypes: float64(2), int64(3), object(1)
memory usage: 12.1+ MB
None
```

Dataset Head:



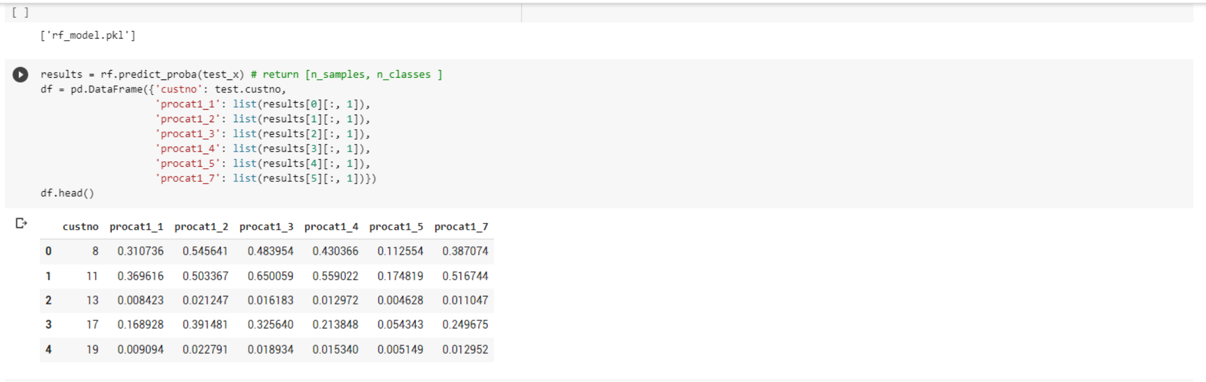
	custno	ordno	orderdate	prodcat2	prodcat1	revenue
0	18944	64694	2016-11-27 20:57:20	NaN	1	53.30
1	18944	114405	2017-04-29 20:18:04	NaN	1	0.10
2	18944	28906	2017-04-23 21:31:03	NaN	1	141.66
3	36096	62681	2016-02-25 07:16:33	NaN	1	36.82
4	1	1	2017-06-12 08:27:59	NaN	1	8.35

Customer Segmentation with RMF



	Segement	RMF	Description	Marketing
0	Best Customers	111	Bought most recently and most often, and spend the most	No Price incentives, new products, and loyalty probrams
1	Loyal Customers	X1X	Buy most frequently	Use R and M to further segment
2	Big Spenders	XX1	spend the most	Market your most expensive products
3	Almost Lost	311	Haven't puchased for some time, but purchased frequently and spend the most	Aggressive price incentives
4	Lost Customers	411	Haven't puchased for long time, but purchased frequently and spend the most	Aggressive price incentives
5	Lost Cheap Customer	444	Last purchased long ago, purchased few, and spend little	Don't spend too much trying to re-acquire

Prediction:



```
[ ]
['rf_model.pkl']
```

```
results = rf.predict_proba(test_x) # return [n_samples, n_classes ]
df = pd.DataFrame({'custno': test.custno,
                   'procat1_1': list(results[0][:, 1]),
                   'procat1_2': list(results[1][:, 1]),
                   'procat1_3': list(results[2][:, 1]),
                   'procat1_4': list(results[3][:, 1]),
                   'procat1_5': list(results[4][:, 1]),
                   'procat1_7': list(results[5][:, 1])})
df.head()
```

	custno	procat1_1	procat1_2	procat1_3	procat1_4	procat1_5	procat1_7
0	8	0.310736	0.545641	0.483954	0.430366	0.112554	0.387074
1	11	0.369616	0.503367	0.650059	0.559022	0.174819	0.516744
2	13	0.008423	0.021247	0.016183	0.012972	0.004628	0.011047
3	17	0.168928	0.391481	0.325640	0.213848	0.054343	0.249675
4	19	0.009094	0.022791	0.018934	0.015340	0.005149	0.012952