**Forecasting with Machine Learning**

*Full name:*

*Class:*

Abstract:

This thesis investigates the use of Machine Learning (ML) techniques in predicting. It focuses on the most important parts of data collecting, the many types of data used in forecasting, and provides a practical example of project data and code. The purpose is to offer a thorough grasp of data collecting and to explain how ML algorithms may be used to make reliable predictions.

The thesis begins by highlighting the importance of data collecting in predicting and then investigates various data sources and methodologies. Data quality, volume, and velocity issues are examined, as well as solutions for preprocessing, cleaning, and feature engineering to enhance forecasting models.

1. **Introduction**

Businesses across all sectors have the continual challenge of producing accurate forecasts about future events and trends in today's fast-paced and data-driven environment. Precision forecasting has become an important aspect in making informed decisions, maximizing resource allocation, and obtaining a competitive advantage. Traditional forecasting approaches frequently rely on statistical models and expert opinion, which may have limitations in dealing with complicated patterns and large-scale data. However, the introduction of Machine Learning (ML) has revolutionized forecas

providing powerful tools and methodologies for extracting valuable insights from massive volumes of data.

Forecasting is critical in a wide number of business disciplines. Accurate projections of stock prices, market trends, and portfolio performance, for example, can drive investment decisions and risk management methods in finance. Sales forecasting aids in the optimization of inventory management, supply chain logistics, and pricing strategies in retail. Accurate demand forecasting in healthcare guarantees enough resources and staffing levels in hospitals and clinics. Forecasting is very important in industries such as energy, transportation, marketing, and many others.

As a subset of artificial intelligence, machine learning has shown to be extremely successful in dealing with difficult predicting challenges. ML algorithms can detect hidden patterns and correlations in data, adapt to changing settings, and make accurate predictions using historical and real-time data. Businesses may improve the accuracy, efficiency, and dependability of their forecasting operations by using the potential of ML.

ML provides a varied set of algorithms appropriate for a variety of predicting problems. ML provides a toolbox capable of processing varied data sources and modeling complicated connections, ranging from classic approaches such as linear regression and decision trees to sophisticated techniques like as neural networks, ensemble models, and deep learning architectures. These algorithms can handle time series data, cross-sectional data, and even unstructured data like text or photos, allowing organizations to get important insights and make educated decisions.

1. **About Data**

Building reliable forecasting models requires the collection of relevant and high-quality data. With technological breakthroughs and the growth of digital platforms, data availability and accessibility have substantially risen. Here are some popular data collecting sources and methods for forecasting:

*2.1* Internal Data: Companies can use internal data sources such transactional records, customer databases, sales reports, and operational analytics. Internal data is a rich source of information particular to the operations of the organization, allowing for the construction of personalized forecasting models.

*2.2* Web Scraping: Web scraping is the process of obtaining information from webpages or other internet sources. It can be a good way to collect specific data points, such as product pricing, customer feedback, or competition information. When doing online scraping operations, it is critical to follow website terms of service and regulatory requirements.

*2.3* Surveys and Questionnaires: Surveys and questionnaires enable firms to collect information directly from their target audience or relevant stakeholders. This method is very beneficial for gathering subjective information or recording opinions and preferences. Online platforms, email, and even in-person interviews can be used to conduct surveys.

*2.4* IoT (Internet of Things) Devices: Businesses may now collect real-time data from sensors, equipment, and linked devices thanks to the rising ubiquity of IoT devices. These sensors offer continuous data streams, allowing for quick reactions to changing conditions and real-time forecasts.

Collaboration with external organizations, research institutes, or industry groups might offer access to specialist datasets or exclusive data sources. These collaborations have the potential to increase the diversity and depth of data available to forecasting algorithms.

*2.5* Data markets: Data markets have evolved as platforms via which organizations may obtain ready-to-use datasets from a variety of vendors. These systems provide curated and pre-processed datasets, reducing the time and effort required for data gathering and processing.

While data availability has risen, data collecting for forecasting still presents a number of issues that must be addressed:

*2.6* Data Quality: Maintaining high data quality is essential for successful forecasting. Errors, inconsistencies, missing numbers, and outliers may exist in data. To reduce the influence of poor data quality on forecasting models, robust data validation and cleaning techniques are required.

*2.7* Data Volume and Velocity: Today's data volume and velocity might be daunting. Furthermore, some sectors, such as financial markets or social media, generate data at rapid rates, necessitating effective data storage, processing, and analysis strategies to deal with the amount and velocity of data.

1. **Forecasting ratings with ML.**

The goal of this project is to create a forecasting model that uses Machine Learning techniques to predict the ratings of items found in supermarkets. Ratings are important in influencing customer decisions and may have a considerable influence on product sales and brand reputation. Businesses may acquire insights into consumer preferences, discover areas for development, and make educated decisions about product selection, marketing tactics, and customer happiness by properly forecasting ratings.

To do this, the initiative will use past data from grocery purchases as well as customer input. The dataset will comprise information such as product characteristics, purchase history, consumer demographics, and customer-provided product ratings.

Collaborating with participating supermarkets to gain access to their transactional databases will be part of the data collecting process. It will also entail establishing procedures for collecting consumer feedback via surveys or online platforms, as well as capturing their ratings and comments on purchased items.

Following data collection, preparation measures will be done to clean and convert the data into a format appropriate for modeling. Handling missing numbers, reducing outliers, encoding categorical variables, and undertaking feature engineering to extract useful information are all examples of this.

The next stage is to use Machine Learning techniques to create a prediction model. ML approaches may be investigated using regression models such as linear regression or support vector regression, as well as more complex algorithms like as random forests or gradient boosting.

The model will be trained on historical data using ratings as the goal variable. Product qualities, client demographics, and purchasing history will be used to properly anticipate the ratings. The model's parameters will be optimized during the training phase to get the greatest possible performance.

A piece of the dataset will be set aside as a test set to evaluate the model's efficacy. The trained model will then be used to predict test set ratings, and its performance will be evaluated using appropriate evaluation metrics such as mean squared error, root mean squared error, or mean absolute error.

Finally, the project will offer the forecasting model's outcomes and insights. This involves identifying important features that generate favorable or negative ratings, assessing the factors that affect product ratings, and delivering practical recommendations to enhance product quality and consumer happiness in supermarkets.

Businesses in the grocery industry may acquire significant insights into customer preferences and make data-driven decisions to improve their product offers and customer experience by effectively executing this project.

* 1. *Data set for the project.*

This dataset contains useful information for estimating ratings based on parameters such as product line, consumer type, gender, and other characteristics. With this data, we may preprocess and evaluate it before developing a Machine Learning model for rating predicting.

**Invoice ID**: Unique identifier for each transaction.

**Branch:** The branch or location of the supermarket where the transaction took place (A, B, or C).

**City:** The city where the branch is located (Yangon, Naypyitaw, or Mandalay).

**Customer type:** Whether the customer is a member or a regular customer.

**Gender:** The gender of the customer.

Product line: The category or type of product being purchased.

**Unit price:** The price of each unit of the product.

**Quantity:** The number of units purchased.

**Tax 5%:** The tax amount calculated as 5% of the total.

**Total:** The total amount including the tax.

**Date:** The date of the transaction.

**Time:** The time of the transaction.

**Payment:** The payment method used

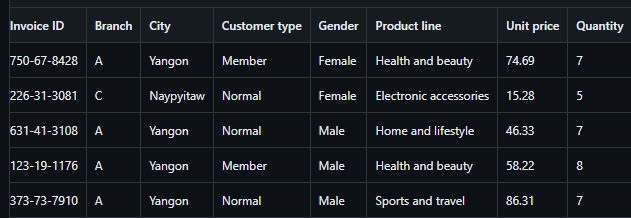
(Ewallet, Cash, Credit card).

**COGS (Cost of Goods Sold):** The cost of goods sold.

**Gross margin percentage:** The gross margin percentage for the transaction.

**Gross income:** The gross income earned from the transaction.

**Rating:** The rating provided by the customer for the purchased product on a scale from 1 to 10.





* 1. *About used the algorithms.*

Various Machine Learning techniques were used in this research to predict ratings based on the available dataset. Here are the algorithms that were used:

Linear Regression is a well-known regression technique. It presupposes that the input characteristics and the target variable have a linear relationship. It finds a line that minimizes the sum of squared errors and fits it to the data. In this project, the Linear Regression model from scikit-learn's LinearRegression() class was employed.

Decision Tree is a non-parametric method that predicts outcomes based on a tree-like model of decisions and their potential effects. It divides the data into nodes and leaves depending on characteristics, allowing it to capture non-linear interactions. For this approach, the scikit-learn DecisionTreeRegressor() class was used.

Random Forest: Random Forest is an ensemble method that makes predictions by combining numerous decision trees. It constructs a set of decision trees and averages their predictions to increase accuracy and resilience. The Random Forest model from scikit-learn's RandomForestRegressor() class was utilized.

The regression extension of Support Vector Machines is Support Vector Regression. It transforms the data into a high-dimensional feature space and searches for a hyperplane that maximizes the margin while reducing mistakes. This method makes use of the scikit-learn SVR() class.

KNN (Kindest Neighbors):

1. Nearest Neighbors is a basic yet effective method that predicts the target variable based on the training data's k closest neighbors. It computes the average of the k nearest neighbors or gives the majority label based on distance metrics. For KNN regression, the scikit-learn KNeighborsRegressor() class is used.
   1. *About the loss.*

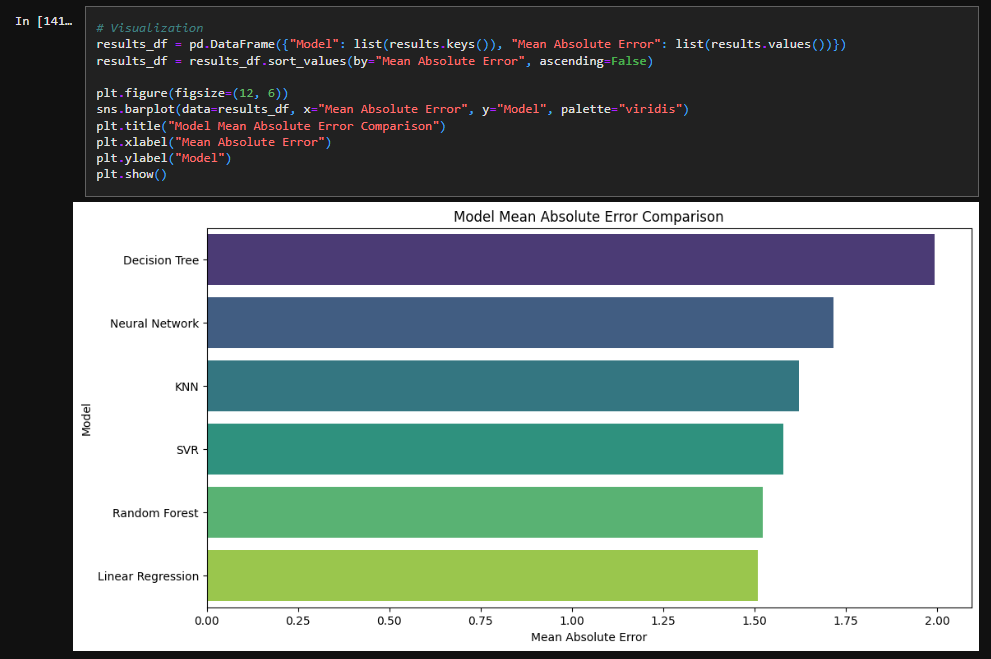
The mean absolute error (MAE) is a popular assessment statistic for regression tasks such as rating prediction. It computes the average absolute difference between anticipated and actual values.

The MAE is a simple measure of the model's accuracy since it indicates the average size of the model's mistakes. It assigns equal weight to all mistakes regardless of their direction (positive or negative).

To compute the MAE, compute the absolute differences between the predicted and actual values for each data point, and then take the average of these absolute differences. A lower MAEsuggests better performance, with values near 0 suggesting more accuracy.

The MAE can assist assess the accuracy of models in predicting consumer ratings in the context of rating prediction for supermarket items. You may identify whether model works better in terms of minimizing the average absolute difference between anticipated and actual ratings by comparing the MAE values produced from different methods.

It's worth noting that, while MAE is a useful measure of accuracy, it doesn't capture the direction or pattern of mistakes. Other metrics, such as mean squared error (MSE) or R-squared, can be used in conjunction with MAE to examine the direction of mistakes.



1. **Conclusion.**

Finally, anticipating product evaluations in supermarkets using Machine Learning approaches provides huge benefits to businesses. Organizations may obtain significant insights into client preferences, make educated decisions regarding product selection and marketing tactics, and improve customer happiness by properly forecasting ratings.

Internal data, external data, online scraping, surveys, IoT devices, collaborative partnerships, and data markets are all used to acquire data for forecasting. Each of these sources offers distinct chances to collect relevant and diverse data for the development of effective forecasting models.

This thesis discusses a project that seeks to predict ratings for grocery items. The dataset contains information such as product specifications, purchase history, customer demographics, and customer ratings.

Linear Regression, Decision Tree, Random Forest, SVR, KNN, and Neural Network (MLP) are among the algorithms used in the project. These algorithms have varying strengths and capacities, allowing for a thorough examination of their effectiveness in predicting ratings.

Mean absolute error (MAE), which assesses the average absolute difference between predicted and actual scores, is used to evaluate the models. The lower the MAE number, the better the model's accuracy in rating prediction.

This study seeks to determine the best effective model for rating predicting in the supermarket environment by using the dataset and applying multiple algorithms. The models' insights can help firms better understand client preferences, optimize product offers, and increase overall customer happiness.

In conclusion, anticipating ratings using Machine Learning approaches delivers significant information for grocery operations. This initiative helps to improve decision-making processes and consumer experiences in supermarkets by using data gathering methods, deploying various algorithms, and utilizing evaluation measures such as MAE.