SALES FORECASTING
A PROJECT REPORT

Submitted in the partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING IN

ARTIFICIAL INTELLIGENCE AND

MACHINE LEARNING

Submitted by:

Aryan Gupta
20BCS6656

Ankith Raj
20BCS6684

Aryan Khuswaha
20BCS6691

Sheikh Shahnawaz Hussain
20BCS6628

Under the Supervision of:

Dr.Alankrita Aggarwal

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,
PUNJAB

10 JANUARY, 2023

CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

BONAFIDE CERTIFICATE

Certified that this project report " **Sales Forecasting** " is the bonafide work of " Aryan Gupta , Ankith Raj , Aryan Kushwaha and Sheikh Shahnawaz Hussain" who carried out the project work under my/our supervision.

SIGNATURE

SUPERVISOR

HEAD OF THE DEPARTMENT

SIGNATURE

Submitted for the project viva-voce examination held on

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

I would like to thank my supervisor, Dr. Alankrita Aggarwal Mam, for her guidance and advice through each stage of making this project and giving me and my teammates Mr. Ankith Raj, Mr. Sheikh Shahnawaz Hussain and Mr. Aryan Khushwaha this opportunity to work on a minor project in which we can show our true potential, creativity, and hard work.

I would also like to thank my family and friends, who have been a constant support and have always motivated me to work hard and bring out the best in me.

This project’s success and end necessitated a great deal of direction and assistance from many people, and we are extremely fortunate to have received it all as part of the project's completion.

We owe everything we've accomplished to their oversight and help, and we'd like to express our gratitude.

ABSTRACT

Forecasting sales is a common and essential use of machine learning (ML). This paper discusses need of Sales Forecasting. Sales Forecasting is the estimate of amount of sales to be expected for a item/product or products for a future period of time. Using Sales Forecasting the management of the enterprise can take decision regarding operations planning, scheduling, production programming inventories of various types, physical distribution and operating profits on the basis of sales forecasts. It also contains some additional benefits like deciding investment proporsals like modernization, expansion of existing units etc. Sales forecasts are essential to make proper arrangement for training the man-power in its own unit or sending them to other industries in the country or abroad to meet the future needs of expertise.

Keywords: sales forecasting, expansion, planning, ML , Training

Table of Contents

List of Tables

Chapter 1: Introduction

Sales forecasting involves analyzing a company's historical sales data to make predictions about its future sales performance. It is an essential aspect of financial planning that helps companies plan for short- and long-term growth. Like all forecasting processes, there is a level of risk and uncertainty involved, so it is importantfor sales forecasting teams to acknowledge this uncertainty in their forecasts. Sales forecasting is a widely practiced corporate strategy that involves setting objectives, creating action plans, and allocating budgets and resources to achieve those objectives. Several techniques are commonly used for sales forecasting, including Linear Regression, Decision Tree, Random Forest, and XG BOOST.

In this project, we conducted multiple linear regression to predict the future sales. There were several different factors that we analyzed in our regression model starting with a full model withall the variables and then moving towards a reduced model by eliminating insignificant variables.We used several different exploratory analyses to identify the key variables for our regression equation such as correlation plots, heatmaps, histograms etc.

Few other time series forecasting models could have been used as the weekly sales is highly dependent on the past year. Moreover, ARIMA modelling techniques like exponential smoothening and holt winters could have helped us capture the seasonality in the model in a better way. Furthermore, ARIMAX model would have enabled us to have an accurate time seriesmodel based on previous weeks of data as well as factor in few important variables like holiday and department type to get an even better accuracy

For this project, we have used the dataset available from ‘Walmart Store Sales Forecasting’ project that was available on Kaggle. In this dataset, we have weekly sales data for 45 stores and99 departments for a period of 3 years. In addition, we had store and geography specific information such as store size, unemployment rate, temperature, promotional markdowns etc. Using these factors, we needed to develop a regression model that can forecast the sales and is also computationally efficient and scalable. The key issues that we have faced in this analysis is the large dataset that resulted into several computational challenges because of which we had to modify our approach in addressing the problem. We also faced significant challenges in identifying the right variables on which the analysis could be conducted.

1.1 PROBLEM DEFINITION

The main focus of this project is to use training data, specifically the data from Big Mart, to createa Sales Forecasting system. By analyzing this data and using machine learning techniques, the goal is to develop accurate models that can predict sales transactions for any company. Sales forecasts can be useful for setting benchmarks, evaluating the impact of new initiatives, and planning resources to meet expected demand. Additionally, they can be used to project future budgets.

The problem at hand is to develop an accurate sales forecasting model that can predict future sales figures for a company. The objective is to provide reliable estimates of sales quantities or revenue, enabling the company to make informed decisions related to inventory management, resource allocation, production planning, and overall business strategy.

Background: Sales forecasting plays a crucial role in any business, allowing organizations to anticipate demand, plan budgets, set targets, and optimize operations. However, accurately predicting future sales can be challenging due to various factors such as market fluctuations, changing consumer behavior, seasonal trends, competitive dynamics, and internal factors like promotional activities and pricing strategies. Therefore, it is essential to develop a robust forecasting model that takes into account these variables and provides accurate sales predictions.

Data Availability: To tackle the sales forecasting problem, historical sales data, along with relevant auxiliary data, is available. The historical sales data contains information such as date/time of sale, product or service description, quantity sold, revenue generated, customer information, and any additional features that might be relevant to sales patterns. Auxiliary data could include external factors like economic indicators, market trends, competitor data, and promotional activities.

The sales forecasting project aims to develop an accurate and reliable model to predict future sales for a company. By leveraging historical sales data and relevant variables, the project aims to provide actionable insights that will enable the company to make informed decisions regarding inventory management, resource allocation, production planning, and overall business strategy.

1.2 PROJECT OVERVIEW

- Sales Forecast helps in predicting the short-term or long-term sales performance ofthat company.
- The data for the Sales will be taken from Kaggle.
- Data wrangling (Data profiling, missing value treatment, exploratory data analysis) willbe performed.
- Prediction will be automated with Machine learning Models thus saving a lot of time.
- We will be working on Pandas, Matplotlib, Model Training, AWS,HTML, andCSS, Flask.
- By leveraging historical sales data and relevant variables.
- The project aims to provide actionable insights that will enable the company to make informed decisions regarding inventory management, resource allocation, production planning, and overall business strategy.

1.3 HARDWARE SPECIFICATIONS

1.3.1 PC

A pc is a personal computer that can be used for multiple purposes depending on its size, capabilities, and price. They are to be operated directly by the end-user. Personalcomputers are single-user systems and are portable. Our web application program willbe installed on the pc for our clients to use it. This makes it feasible for individual use.

1.4 SOFTWARE SPECIFICATIONS

1.4.1 Jupyter Notebook:

Jupyter Notebook is a web-based open-source application that is used for editing, creating running, and sharing documents that contain live codes, visualization, text, and equations. Its core supported programming languages are Julia, R, and Python. Jupyter notebook comes withan IPython kernel that allows the programmer to write programs in python. There are over 100kernels other than IPython available for use.

1.4.2 Atom Text editor

Atom is a text and source code editor which works across all operating systems. It speeds up find-and-replace operations by an order of magnitude and improves loading performance for large, single-line files It’s a desktop application built with HTML, JavaScript, CSS, and Node.jsintegration.

1.4.3 AWS

Amazon Web Services, Inc. (AWS) is a subsidiary of Amazon that provides on-demand cloud computing platforms and APIs to individuals, companies, and governments, on a metered pay- as-you-go basis. Through AWS server farms, these cloud computing web services offer software tools and distributed computer processing capability. One of these services is Amazon Elastic Compute Cloud (EC2), which enables customers to have a virtual computer cluster at

their disposal that is always accessible via the Internet. The majority of a real computer's features, such as hardware central processing units (CPUs) and graphics processing units (GPUs) for processing, local/RAM memory, hard-disk/SSD storage, a choice of operating systems, networking, and pre-loaded application software including web servers, databases, and customer relationship management, are all emulated by AWS's virtual computers (CRM).

1.4.4 FLASK

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework- related tools.

1.4.5 MS-EXCEL

Microsoft produced Microsoft Excel, a spreadsheet, for Windows, macOS, Android, and iOS. Ithas calculating or computing capabilities, graphing tools, pivot tables, and the Visual Basic for Applications macro programming language (VBA). The Microsoft Office programme package includes Excel.

1.4.6 Visual Studio Code

Microsoft created the source-code editor Visual Studio Code, generally known as VS Code, for Windows, Linux, and macOS using the Electron Framework. Debugging support, syntax highlighting, intelligent code completion, snippets, code refactoring, and integrated Git are among the features. Users may modify the theme, keyboard shortcuts, settings, and add functionality by installing extensions.

With 74.48% of respondents saying that they use it, Visual Studio Code was rated as the most popular development environment tool in the Stack Overflow 2022 development Survey.

Software Tool Used	**Description**	**Logo**
Jupyter Noebook	Jupyter Notebook is a web-based open-source application that is used for editing, creating, running, and sharing documents that contain live codes, visualisations, text, and equations. There are over 100 kernels other than IPython available for use.	jupyter
Atom Text Editor	Atom is a text and source code editor which works across all operating systems. It speeds up find-and-replace operations by an order of magnitude and improves performance of files	
Visual Sudio Code	Visual studio code is an open-source code editor built for Windows, Mac OS, Linux which can be used for various programming languages like Java, JavaScript, Python, C, C++, Node.js.	
Flask	Flask is a micro web framework written in Python. It is classified as microframework because it does not require particular tools or libraries. It has no database abstraction layer, formvalidation, or any other components where pre-existing third-party libraries provide common functions.	Flask

LITERATURE REVIEW

2.1 Existing System Summary

Year and citation	Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 3928 - 3936 Received 15 April 2021; Accepted 05 May 2021.	https://www.anaplan.com/blog/sales-forecasting-guide/ 2022	T uyls, Karl & Maes, Sam & Vanschoenwinkel, B..(2023). Machine Learning Techniques for sales forecasting
Article Title	“Intelligent Sales Prediction Using Machine Learning Techniques	“MachineLearning Models for Sales Forecasting”	“Forecast of Sales of Walmart Store Using Big Data Applications”
Purpose ofthe study	The forecast is composed of a smoothed averaged adjusted for a linear trend. Then the forecast is also adjusted for seasonality Machine learning algorithms such as Generalized Linear Model (GLM), Decision Tree (DT) and Gradient Boost Tree (GBT) are used in prediction of future sales.	Regression algorithm captures the patterns in the whole set of stores or products.The analysis includes the attributes such as mean sales value of historical data, state and school holiday flags,distance from store to competitor’s store, storeassortment type are considered in prediction. Various machine learning models such as Random Forest, Neural network, Lasso regularization, Arima model and Extra Tree model are used to analyze the data.	The plan calls for gathering vast amounts of sales-related data, which is then sent to HDFS (Hadoop's distributed file system) for map reduction. To forecast sales, the Holt Winters algorithm is employed. The algorithm exhibits seasonality, trend, and randomness. Data sets are utilised to train the algorithm, and then it is used to predict sales.
Tools/ Software used	- Jupyter Notebook	- Jupyter Notebook	- Jupyter Notebook
Comparison of techniques done	- Generalized Linear Model (GLM) - Decision Tree (DT) - Gradient Boost Tree (GBT)	- Lasso - Neural Network	- Neural Network - Decision Tree (DT)
Evaluation parameters	- Model Accuracy	- Model Accuracy	- Model Accuracy

Table 2.1: Literature review summary

2.2 Proposed System

- This project mainly focuses on developing a system that can Predict Sales of aCompany
- Data wrangling (Data profiling, missing value treatment, exploratory data analysis) willbe performed.
- We will be working on Pandas, Matplotlib, Model Training, AWS,HTML, andCSS.
- The proposed system for the sales forecasting project consists of several components and processes aimed at developing an accurate and reliable sales forecasting model.
- The system incorporates data collection, preprocessing, model development, evaluation, and implementation, along with ongoing monitoring and iteration.

3. DESIGN PROCESS/FLOW

- Sales forecasting plays a critical role in the success of any business. It provides essential information to allocate resources, hire new staff, manage costs, and increase quotas. The goal of sales forecasting is to provide accurate predictions that businesses can use to makeinformed and impactful decisions.
- Sales forecasts help businesses make informed decisions about staffing, inventory, productlines, and marketing efforts. It allows sales managers and representatives to spot potentialissues and address them before they become problems.
- Sales forecasting is a valuable tool for sales managers and leaders to set

realistic goals. Sales forecasts form the basis of your entire strategy throughout the year, and the insights lay the groundwork not just for the company’s vision, but also for the direction of sales team.

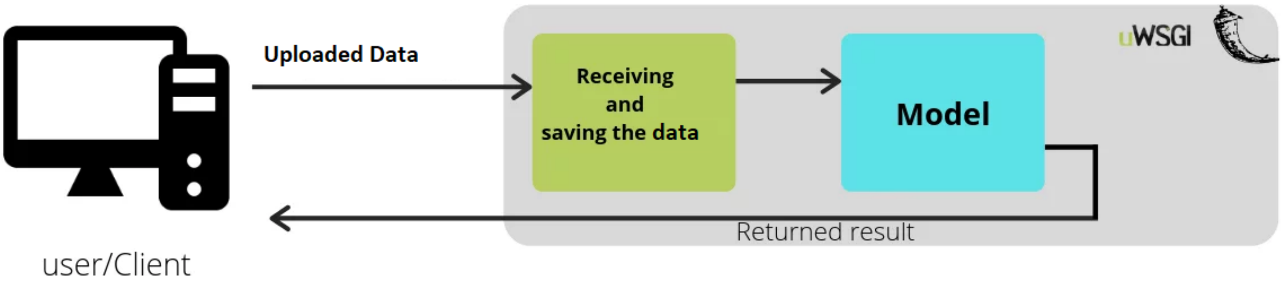
- Sales forecasting also enables sales managers and leaders to set realistic goals for their teams. It forms the basis of a company's strategy throughout the year and provides insightsthat guide the direction of the sales team.
- The design process for the sales forecasting project involves a systematic approach to developing a robust and accurate forecasting model. It includes several key stages and steps to ensure a well-structured and effective design. Here is an overview of the design process for the sales forecasting project:
- Project Planning and Scope Definition:
- Define the project objectives and scope, including the desired accuracy levels, time horizons, and specific requirements.
- Identify the stakeholders and understand their needs and expectations.
- Define the key metrics for evaluating the forecasting model's performance.
- Data Collection and Exploration:
- Gather historical sales data, including relevant features and variables.
- Explore and analyze the data to identify any patterns, trends, or seasonality.
- Identify and collect any auxiliary data that may impact sales, such as economic indicators or competitor information.
- Data Preprocessing and Feature Engineering:
- Handle missing values, outliers, and inconsistencies in the data.
- Conduct feature engineering techniques to extract relevant features and create

additional variables.

- Encode categorical variables and normalize numerical variables, if required.
- Model Selection:
- Evaluate different forecasting models based on their suitability for the project objectives and data characteristics.
- Consider time series models like ARIMA, exponential smoothing, or machine learning algorithms such as regression, random forests, or neural networks.
- Select the most appropriate model or combination of models.
- We will handle this problem in a structured way.
- Loading Packages and Data
- Data Structure and Content
- Exploratory Data Analysis
- Missing Value Treatment
- Feature Engineering
- Encoding Categorical Variables
- Label Encoding
- PreProcessing Data
- Modeling
- Linear Regression
- RandomForest Regressor
- XGBoost
- Deployment

4. METHODOLOGY

The following methodology will be followed to achieve the objectives defined for the proposed research work:



The system flowchart of our project, which is based on Sales Forecasting, is illustrated in the diagram. It is split into two parts: the backend and the frontend. The Bigmart dataset make up the backend element. The Python code is separated into preprocessing, model building, training, and prediction sections.

Flask is a micro web framework written in Python, is what we used for the front-end and connecting to Python Code. Front End is made Using HTML/CSS and consists of various input fields like Item weight, Item visibility, Item Type, Item MRP, Outlet Type, Outlet Location type. User will enter values in these fields and the values will be passed to model and the result of future outlet sales will be shown to the user in JSON Format.

Data collection and preparation, exploratory data analysis, feature engineering, model selection, training and model development, model evaluation, and outcome analysis are typical processes in the approach for sales forecasting.

1. Data Collection and Preparation:

Important elements in the sales forecasting process include data preparation and collecting. Collect historical sales data, including relevant features such as date/time of sale, product description, quantity sold, revenue generated, and customer information. You can take the following actions to guarantee that you gather and prepare the appropriate data for precise sales forecasting:
Set your forecasting timeframe: Choose the time period, such as a quarter or a year, for which you wish to forecast sales.

- Find reliable data sources: Ascertain the locations of the data sources you'll need for sales forecasting, including sales records, customer information, market research studies, and economic indicators.

- Make sure your data is correct, full, and consistent by cleaning and preparing it. This could entail clearing out duplicate records, fixing mistakes, and standardising data formats.

- Set your parameters: Determine the important factors—such as pricing, promotions, seasonality, and economic indicators—that will affect sales.

- Decide on the level of information you require for your data by categorising it by product, customer, or area.

- Review your data: Analyse your data using statistical methods and tools to spot trends and patterns. Regression analysis, time series analysis, and other techniques might be used in this.

Gather historical sales information that includes pertinent details like the date and time of the transaction, the description of the product, the number of units sold, the amount of money made, and the identity of the buyer.

Collect supplementary information, such as market trends, competitor information, economic statistics, and promotional activity data.
Handle missing numbers, outliers, and inconsistencies as part of the pre-processing of the data.

2. Exploratory Data Analysis (EDA):

In order to understand patterns, trends, and relationships that can help forecasting models, exploratory data analysis (EDA), a crucial stage in sales forecasting, involves analysing and visualising data. Investigate the distribution of the variables, look for patterns, and find seasonality or trends by conducting exploratory data analysis.
To acquire insights, visualise the data using plots, charts, and summary statistics.

The following are some crucial actions you may do for EDA in sales forecasting:

- Determine important factors: Make a list of the important factors that will affect sales, including pricing, promotions, seasonality, and economic indicators.

- Assemble data: Gather historical sales information as well as data from other relevant sources, including market research studies, customer information, and economic indicators.

- Data preparation and cleaning: Make sure your data is correct, full, and consistent. This could entail clearing out duplicate records, fixing mistakes, and standardising data formats.

- Visualise data: To examine your data and find patterns and trends, use graphs, charts, and other visualisation tools. This can assist you in discovering links between various elements, such as how pricing affects sales volume.

- Conduct statistical analysis: Examine your data and look for patterns between variables using statistical methods and tools. Regression analysis, time series analysis, and other techniques might be used in this.

- Finding outliers is important since they can have an impact on your forecasting models.

- Refine forecasting models: To make your forecasting models more accurate, use the EDA's insights.

For example, the following strategies can be utilised in EDA for sales forecasting:

- Plotting sales data across time as a time series can help you spot trends, seasonality, and other patterns.

- Scatter plots can be used to visualise relationships between two variables, such as those between pricing and sales volume.

- Heat maps: These can be used to spot patterns in consumer behaviour, such as the most popular combinations of products.

- Box plots: These are useful for finding abnormalities and outliers in sales data.

You can obtain a deeper knowledge of your data and improve the accuracy of your forecasts by incorporating EDA into your sales forecasting process.

3. Feature Engineering:

Creating additional features or variables from current data in order to enhance the effectiveness of machine learning models for sales forecasting is known as feature engineering.

Determine any characteristics that might have an effect on sales, such as time-related factors, lag factors, or product characteristics.

Extracting pertinent data from the data will enable the creation of new features.

For the purpose of modelling, normalise numerical features and encode categorical variables.

The following are some essential phases for feature engineering in sales forecasting:

- Determine the pertinent variables: The important factors that will affect sales, such as pricing, promotions, seasonality, and economic indicators, should be identified.

- Investigate connections: Use exploratory data analysis (EDA) to investigate connections between variables and find potential new features that could boost predicting precision.

- develop new variables: To develop new variables that capture crucial information from current data, use statistical techniques and domain knowledge. You may develop variables to track the effect of promotions or the time lag between changes in economic data and shifts in sales, for instance.

- Test new variables: To ascertain their effect on performance, test the new variables in your forecasting models. To hone your new variables and select the ideal mix for optimum performance, you might need to repeat this step.

Following are some particular methods that can be utilised in feature engineering for sales forecasting:

- The influence of earlier sales or economic factors on present sales volumes is captured by lag variables.
- Moving averages: These can capture patterns over time and smooth out noisy data.
- A new product launch or a shift in interest rates are examples of market events or market developments that are captured by indicator variables.
- Terms used to describe interactions between two or more variables include the effect of promotions on product sales volumes, for example.

You may increase the precision of your forecasting models and make better business judgements by incorporating feature engineering into your sales forecasting process.

Identify potential features that may impact sales, such as time-related variables, lag variables, or product attributes.
Create additional features by extracting relevant information from the data.
Normalize numerical features and encode categorical variables for modeling purposes.

4. Model Selection:

Model selection, which entails selecting the ideal machine learning model for your data and business requirements, is a crucial phase in the sales forecasting process. Following the following essential stages will help you choose a model for sales forecasting:

Evaluate various forecasting models suitable for the sales forecasting task, such as ARIMA, exponential smoothing, regression, random forests, or neural networks.

Consider the characteristics of the data, time dependencies, and specific requirements of the project:

- Describe the issue with your forecast: Decide what specific forecasting issue, such as anticipating sales volume or revenue, you are attempting to resolve.
- Determine the pertinent variables: The important factors that will affect sales, such as pricing, promotions, seasonality, and economic indicators, should be identified.
- Assemble data: Gather historical sales information as well as data from other relevant sources, including market research studies, customer information, and economic indicators.
- Data preparation To make sure your data is correct, whole, and consistent, clean and prepare it.
- Divide your data into training and testing sets to compare the effectiveness of various models.

- Choose potential models: Select a group of potential models to test, such as neural networks, decision trees, random forests, or linear regression.

- Review models: Each candidate model should be trained on the training set before having its performance assessed on the testing set. To compare the effectiveness of various models, use performance indicators like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE).

- Improve model performance by refining candidate models and exploring novel approaches like feature engineering as you iterate the model selection process.

- Ultimate model of choice: Select the model that performs the best for your company's requirements and use it for forecasting.

The ideal model for your company's requirements will depend on a number of variables, including the quantity and quality of the data available, the complexity of the forecasting problem, and the desired level of accuracy.

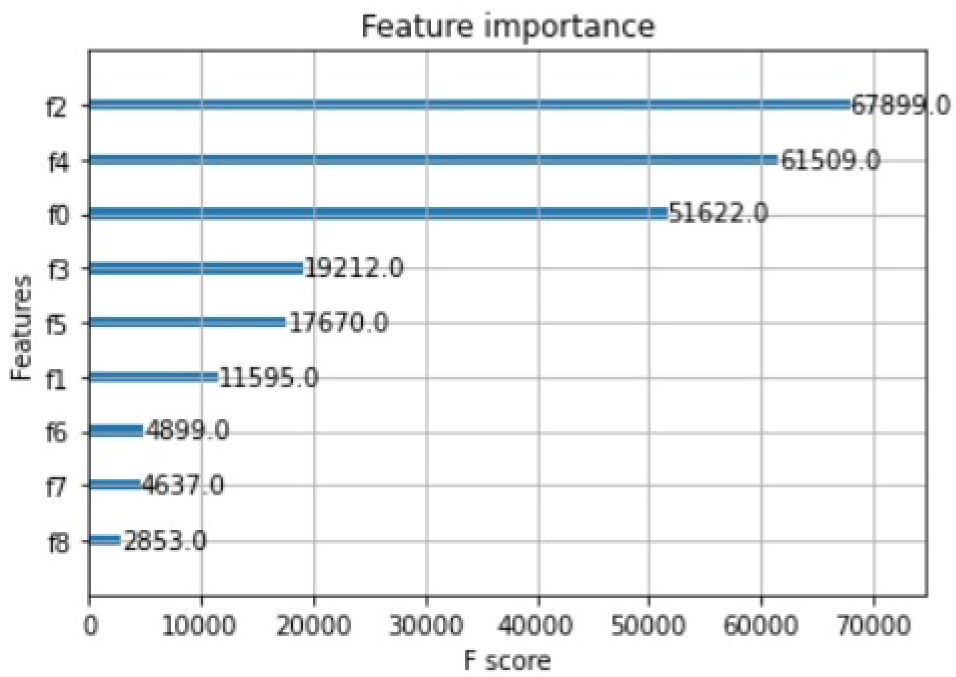
You can increase the accuracy of your sales forecasting and improve the quality of your business decisions by following these steps and continuously improving your model selection methodology.

5. Models Used:

XGBoost

Extreme Gradient Boosting, often known as XGBoost, is a well-liked machine learning method that excels at handling a wide range of supervised learning issues. It is a member of the gradient boosting family of algorithms and is often employed in both professional applications and data science contests.

The model shows the relative importance of each feature with their feature score



The XGBoost library, which provides an effective implementation of the XGBoost algorithm, may be used with Python.

You must first use pip to install the library before you can use XGBoost:
pip install xgboost

The XGBoost library may be imported into your Python script or notebook after installation using the snippet:
import xgboost as xgb

Let's now examine a few of XGBoost's most important ideas and characteristics:

Gradient Boosting: The ensemble learning approach known as gradient boosting is the foundation of XGBoost. In order to generate a powerful predictive model, it integrates many weak prediction models (usually decision trees). Each succeeding model fixes the errors created by the earlier models as it creates the models in a sequential fashion.

Decision Trees: As base learners, XGBoost uses decisions trees. A specific loss function, such as mean squared error (for regression) or log loss (for classification), is minimised through decision trees that are constructed repeatedly. Maximum depth, minimum child weight, and splitting criteria are just a few of the decision tree features that may be customised using XGBoost.

Regularization: To avoid overfitting, XGBoost uses regularization methods. It regulates the model's complexity by applying L1 (LASSO) and L2 (Ridge) regularization terms to the objective function.

XGBoost has a built-in technique to determine the scores for each feature's relevance. These scores measure the relative relevance of every characteristic in the dataset and aid in determining which elements have the most influence.

In order to maximise efficiency and scalability, XGBoost is built to take use of parallel processing capabilities. The building of trees may be done in concurrently, which expedites the training process.

Let's talk about some typical XGBoost applications now:

Classification: Binary and multiple-class classification tasks are both compatible with XGBoost. It has obtained state-of-the-art outcomes in many categorization competitions. It is capable of diagnosing diseases, detecting fraud, and handling datasets with imbalances successfully.

Regression: XGBoost is frequently used for issues involving regression. It is ideal for jobs like forecasting home prices, stock market trends, and customer lifetime value since it can anticipate continuous values.

XGBoost may be used for learning to rank tasks, where the objective is to arrange a group of things according to how relevant they are to a question. It has been used in prediction of ad click-through rates, recommendation systems, and search engines.

Dataset outliers or anomalies can be found using XGBoost's anomaly detection feature. It can identify departures from such patterns by learning the typical patterns from labelled data, which helps with fraud detection, network intrusion detection, or system monitoring.

The feature significance ratings provided by XGBoost may be used to conduct feature selection. These scores may be used to determine which elements are most important and to eliminate those that are unnecessary or redundant, simplifying the model and making it easier to understand.

These are only a few examples of the uses for XGBoost. It is an important tool in many machine learning applications thanks to its performance and adaptability.

Random Forest

The widely used ensemble learning method Random Forest is utilised for both classification and regression problems. To provide a more reliable and accurate prediction model, it mixes numerous decision trees. The scikit-learn package, which offers a complete set of machine learning capabilities in Python, may be used to implement the Random Forest method.

If you haven't previously, you must first install scikit-learn in order to utilise Random Forest in Python:

```
pip install scikit-learn
```

After installation, you may import the Random Forest classifier or regressor from scikit-learn's ensemble module:

```
from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor
```

Let's now explore the main ideas and characteristics of Random Forest:

Random Forest is a technique for ensemble learning that combines the predictions of many decision trees. It seeks to reduce overfitting and enhance generalisation by averaging the findings of separate trees.

Decision Trees: Decision trees serve as the foundational learners in Random Forest. Recursively dividing the feature space into subgroups based on various characteristics and splitting criteria, decision trees are created. A random subset of the training data is used to train each decision tree in the Random Forest.

Random Subspace: Random Forest adds further randomization by building each decision tree with a random subset of characteristics at each node. The term "feature bagging" or "random subspace method" refers to this procedure. It promotes variation within the ensemble and helps to decorrelate the trees.

Bootstrap Aggregating (Bagging): Random Forest makes use of a method known as bootstrap aggregating. By randomly selecting with replacement, it divides the training data into several subgroups. After then, each subset is utilised to train a different decision tree in the forest. The use of bags enhances the model's overall stability by lowering variation.

Voting and Prediction: Random Forest uses majority voting to aggregate different trees' predictions while performing categorization tasks. Every tree makes a vote for a certain class, and the class receiving the most votes is the one whose forecast is ultimately used. In regression tasks, the final prediction is calculated by averaging the predictions of each individual tree.

Let's now cover a few typical Random Forest applications:

Classification: For classification problems, Random Forest is frequently employed. High-dimensional datasets, noisy data, and unbalanced classes are all handled with good performance. Healthcare, banking, and image identification are just a few of the industries where it has been used.

Regression: Random Forest is also useful for problems involving regression. It is capable of dealing with both linear and nonlinear interactions between features and can predict continuous values. Housing pricing, stock market patterns, and demand predictions have all benefited from its use.

The built-in feature importance measure offered by Random Forest quantifies the relative importance of each feature in the dataset. The most informative characteristics for the job at hand can be chosen using this information for feature selection.

Random Forest can be used for anomaly detection, where the objective is to find observations that significantly deviate from the norm. It can identify unexpected or abnormal instances by learning the patterns of typical occurrences, which helps with fraud detection, network monitoring, and quality control.

Ensemble Learning Comparison: Random Forest may serve as a baseline or benchmark model to assess the performance of various ensemble learning methods. It can be used to assess the potency of novel methods and determine whether they perform better than the Random Forest method.

These are only a few examples of the uses for Random Forest. It is a commonly used method in machine learning applications because to its adaptability, robustness, and interpretability.

Linear Regression

A common statistical modelling method for determining the connection between a dependent variable and one or more independent variables is linear regression. The dependent variable can be predicted as a linear combination of the independent variables on the assumption that there is a linear relationship between the variables. Finding the best-fitting line that reduces the discrepancy between the observed and predicted values is the goal of linear regression.

You may conduct linear regression in Python by implementing the technique from scratch or by utilising a variety of libraries, such as scikit-learn and statsmodels. Let's concentrate on scikit-learn, a well-liked Python machine learning library:

If you haven't previously, you must install the library before using linear regression in scikit-learn:

```
pip install scikit-learn
```

Once set up, you can import the LinearRegression class from the ‘linear_model’ module:

```
from sklearn.linear_model import LinearRegression
```

Let's now examine the fundamental ideas and characteristics of linear regression:

Simple Linear Regression: In simple linear regression, the dependent variable (goal) is predicted using just one independent variable (feature). A straight line with an intercept and a slope is used to represent the relationship between the variables.

Multiple Linear Regression: Multiple independent variables are utilised to predict the dependent variable in multiple linear regression. In a higher-dimensional space, the relationship between the variables is represented by a hyperplane.

In a linear regression, the line equation is written as y = b0 + b1x1 + b2x2 +... + bn*xn, where y is the dependent variable, b0 is the intercept, and b1 to bn are the coefficients (slopes) linked to the independent variables x1 to xn.

Estimation of Coefficients: Using an approach like conventional least squares, linear regression calculates the coefficients by minimising the sum of squared residuals (difference between observed and predicted values). Each independent variable's influence and direction on the dependent variable are shown by the calculated coefficients.

Model Evaluation: A number of measures, including mean squared error (MSE), root mean squared error (RMSE), coefficient of determination (R-squared), and others, can be used to assess the accuracy of a linear regression model. These metrics may be utilised for model comparison and model selection as they show how well the model fits the data.

Some common applications of linear regression are as follows:

For problems involving predictive modelling, linear regression is frequently utilised. Because it can anticipate continuous values, it is appropriate for tasks like demand forecasting, customer lifetime value estimation, and sales forecasting.

Data trends and patterns may be analysed using linear regression. It can shed light on the direction and size of change over time by fitting a line to historical data. This is helpful for forecasting the stock market, the economy, and financial analysis.

Evaluation of Relationships: The strength and direction of a link between variables may be assessed using linear regression. It allows for the quantification and identification of factors that significantly affect the dependent variable. Studies in the social sciences, marketing research, and medicine can all benefit from this.

Important Feature: In multiple linear regression, the independent variables' coefficients show the significance of and effect on the dependent variable. Positive coefficients imply a favourable association, whilst negative coefficients imply an unfavourable relationship. Using this knowledge, one may choose features and comprehend a phenomenon's primary causes.

Analysis of Residuals: The residuals, or discrepancies between observed and anticipated values, are a feature of linear regression that may be used to analyse residuals. The identification of outliers, heteroscedasticity (unequal variance), and assumption violations using residual analysis can assist provide light on how to enhance the model.

These are just a few instances of the many fields in which linear regression may be used. It is an effective tool for outcome analysis and prediction across a wide range of domains due to its clarity, interpretability, and capacity to capture linear correlations.

6. Training and Model Development:

In order to generate machine learning models that can reliably forecast future sales volumes or income, training and model development are essential processes in the sales forecasting process. Split the pre-processed data into training and validation sets, considering the desired time period for validation.

You can use the following important procedures for training and model development in sales forecasting:

- Describe the issue with your forecast: Decide what specific forecasting issue, such as anticipating sales volume or revenue, you are attempting to resolve.

- Determine the pertinent variables: The important factors that will affect sales, such as pricing, promotions, seasonality, and economic indicators, should be identified.

- Assemble data: Gather historical sales information as well as data from other relevant sources, including market research studies, customer information, and economic indicators.

- Data preparation To make sure your data is correct, whole, and consistent, clean and prepare it.

- Divide your data into training and testing sets to compare the effectiveness of various models.

- Choose potential models: Select a group of potential models to test, such as neural networks, decision trees, random forests, or linear regression.

- Model training: Using methods such as gradient descent, random search, or Bayesian optimisation, train each candidate model on the training data.

- Model evaluation: Use performance metrics like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) to assess each model's performance on the testing set.

- Model refinement: To increase model performance, iterate the model building process, fine-tuning potential models and investigating novel approaches like feature engineering or hyperparameter tweaking.

- Model validation: To make sure the final model works well with new data, validate its performance using methods like cross-validation or out-of-sample testing.

- Deploy the finished forecasting model, and then continuously assess its effectiveness, modifying and upgrading it as required.

It's crucial to keep in mind that the accuracy of your models and their capacity to capture the intricate correlations between variables that affect sales will depend on the quality and quantity of your data, as well as the quality and quantity of your data.

Train the selected model(s) on the training data, incorporating the identified features. Optimize the model's hyperparameters using techniques like grid search, random search, or Bayesian optimization.

You may increase the accuracy of your sales forecasting and make better business decisions by following these steps and consistently improving your training and model building methodology.

7. Model Evaluation:

The evaluation of your machine learning models' accuracy and efficiency in estimating future sales volumes or revenue is a crucial phase in the sales forecasting process. Evaluate the trained model(s) on the validation set using appropriate evaluation metrics such as mean absolute error (MAE), root mean squared error (RMSE), or mean absolute percentage error (MAPE).

You can use the following essential methods to evaluate models used in sales forecasting:

- Set your evaluation criteria: Select one or more evaluation metrics, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or Mean Absolute Percentage Error (MAPE), that are in line with your company's goals.

- Check the model's performance on the training data to see if overfitting, which happens when the model is too complicated and fits the training data too closely, has occurred.

- Analyse model performance on test data to judge how well it can forecast future sales volumes or revenue. Analyse model performance on test data.

- Compare model performance: To discover which model performs the best, compare the performance of various models using the evaluation measures you have selected.

- Model refinement: To increase model performance, iterate the model building process, fine-tuning potential models and investigating novel approaches like feature engineering or hyperparameter tweaking.

- Ensure that the resulting model works well on new data by validating its performance using methods like cross-validation or out-of-sample testing.

- Make informed business decisions based on the predictions of the model by interpreting the evaluation's results.

Compare the model's performance against established benchmarks or business requirements.
Assess the model's accuracy, reliability, and ability to capture sales patterns.

The ideal model for your company's requirements will depend on a number of variables, including the quantity and quality of the data available, the complexity of the forecasting problem, and the desired level of accuracy. You can increase the accuracy of your sales forecasting and improve the quality of your business decisions by following these steps and continually improving your model evaluation methodology.

In order to obtain understanding of the variables influencing sales performance and guide future business decisions, the outputs of the sales forecasting model are finally analysed and interpreted.

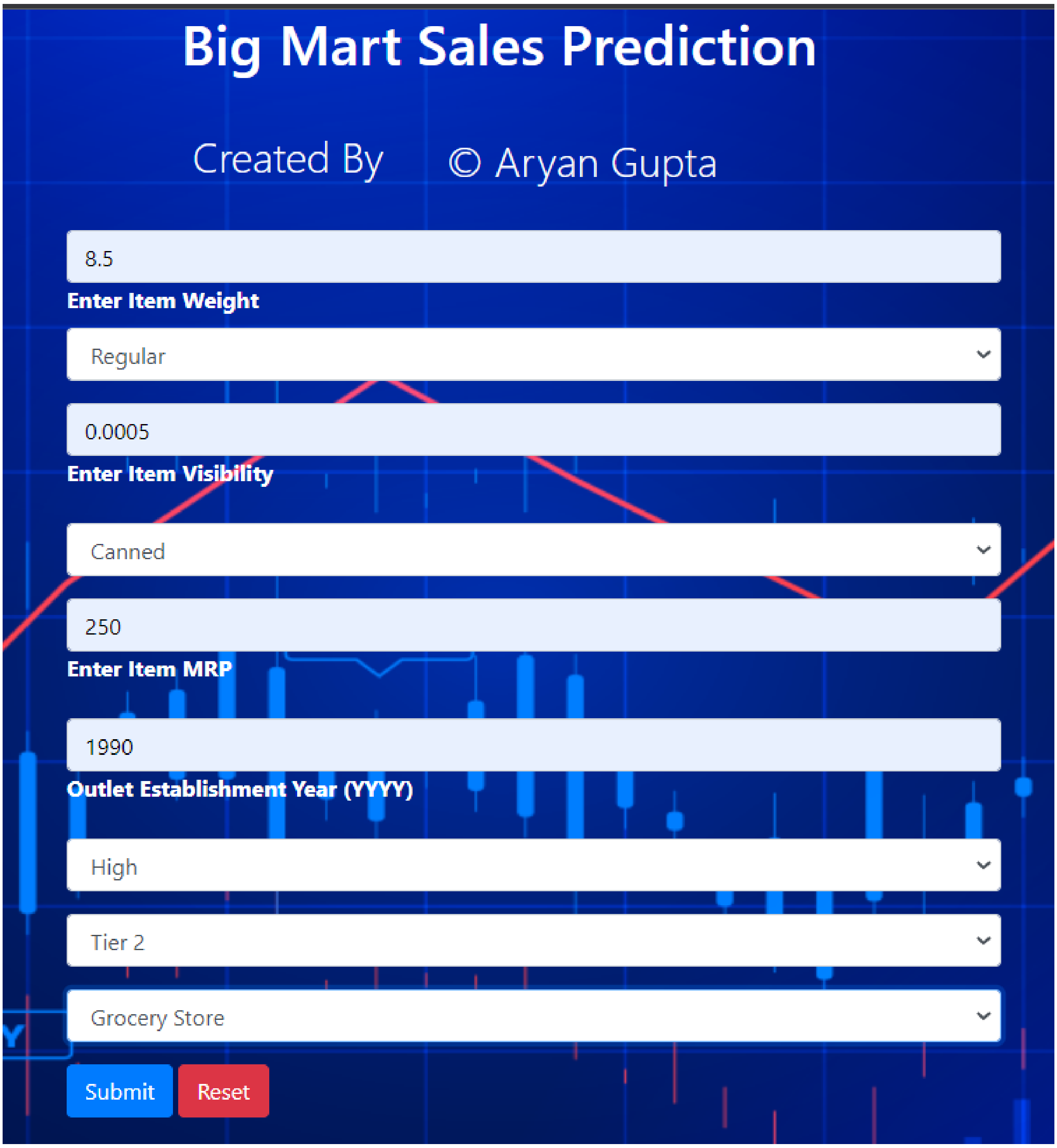
6. RESULT ANALYSIS AND VALIDATION

The sales forecasting process includes a result and analysis stage where you understand the output of your machine learning models and use it to guide strategic business decisions.

The sales of several Big Mart outlets have been predicted using machine learning methods like Linear Regression, K-Nearest Neighbors, XGBoost, and Random Forest. For each of the four methods, many metrics that affect the accuracy of results are tabulated, including Root Mean Squared Error (RMSE), Variance Score, Training and Testing Accuracy, and more.

Business Validation: Include stakeholders and subject-matter experts in the process of validation. Compare the predicted sales figures to your customers' expectations, market trends, or industry standards to confirm the accuracy. Refine and enhance the forecasting model by taking into account their comments and ideas.

Sales forecasting is a process that is improved iteratively. Revisit the forecasting model after analysing the findings and identifying potential areas for improvement. To improve forecasting accuracy over time, make continuous model improvements based on user feedback, data updates, and changes in the business environment.



Big Mart Sales Prediction

Created By © Aryan Gupta

8.5
Enter Item Weight

Regular

0.0005
Enter Item Visibility

Canned

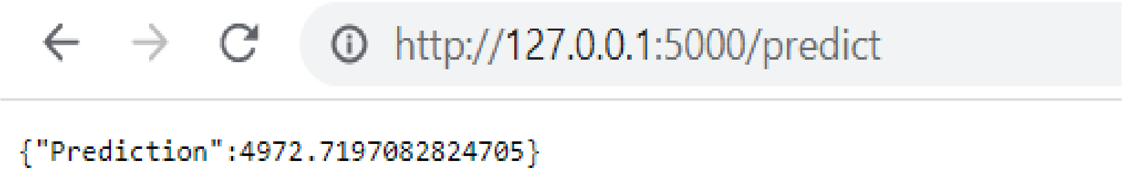
250
Enter Item MRP

1990
Outlet Establishment Year (YYYY)

High

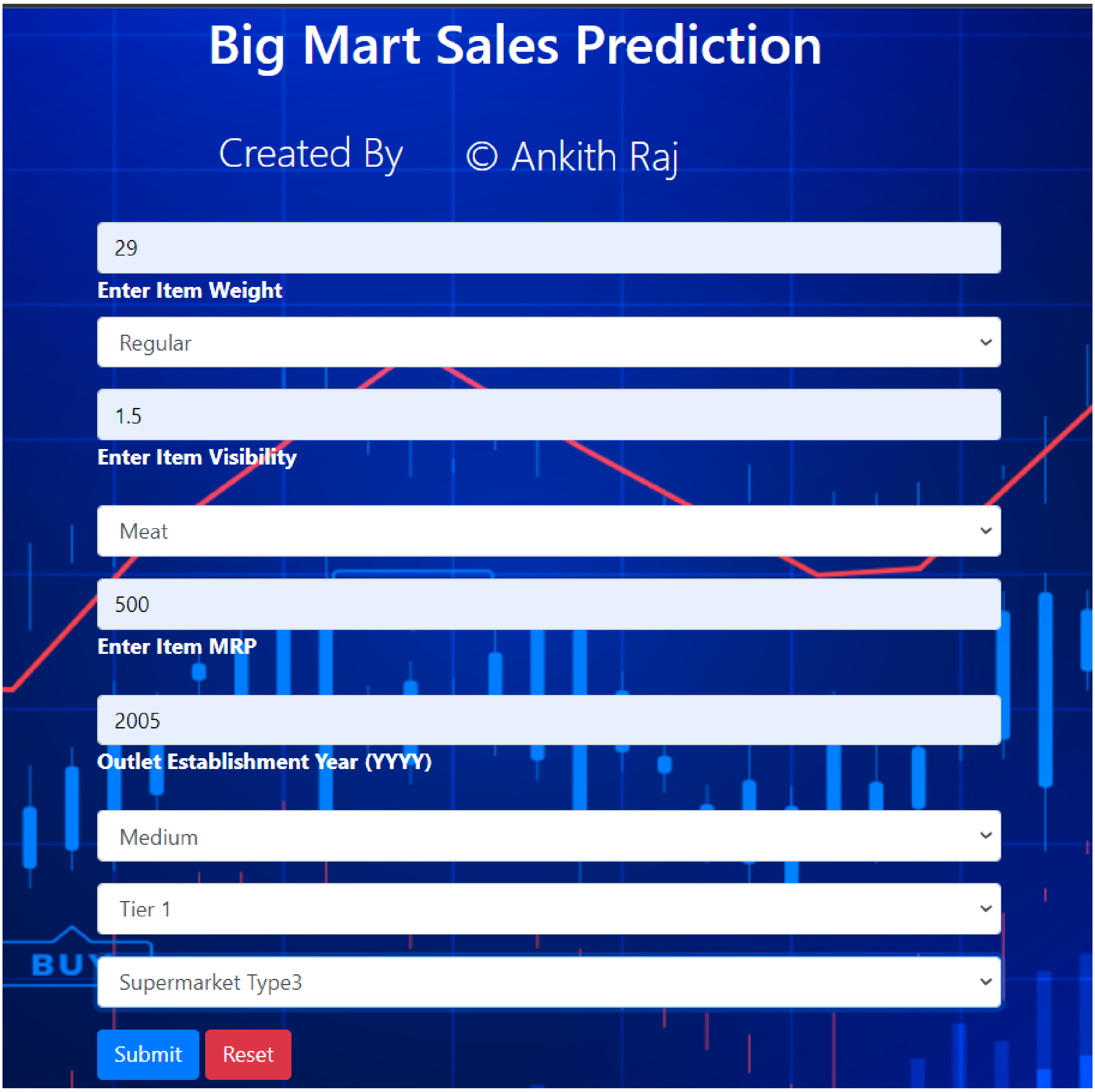
Tier 2

Grocery Store



http://127.0.0.1:5000/predict

```
{"Prediction":4972.7197082824705}
```



Big Mart Sales Prediction

Created By © Ankith Raj

29
Enter Item Weight

Regular

1.5
Enter Item Visibility

Meat

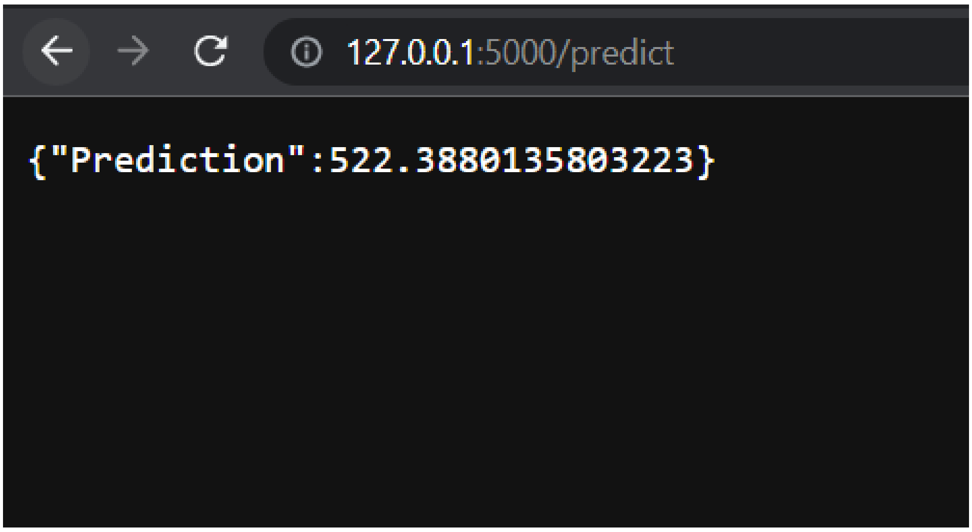
500
Enter Item MRP

2005
Outlet Establishment Year (YYYY)

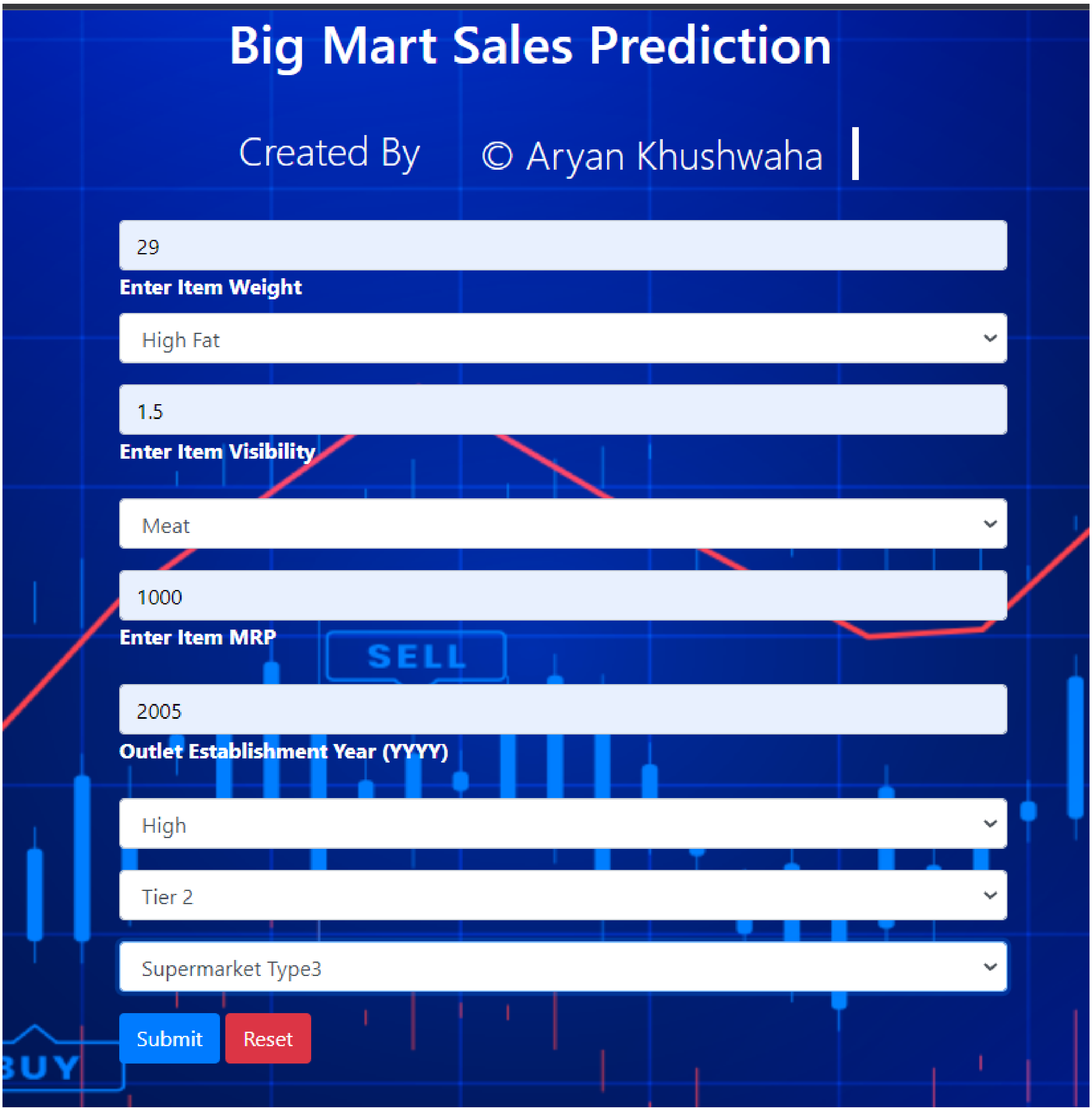
Medium

Tier 1

Supermarket Type3



```
{"Prediction":522.3880135803223}
```



Big Mart Sales Prediction

Created By © Aryan Khushwaha

29
Enter Item Weight

High Fat

1.5
Enter Item Visibility

Meat

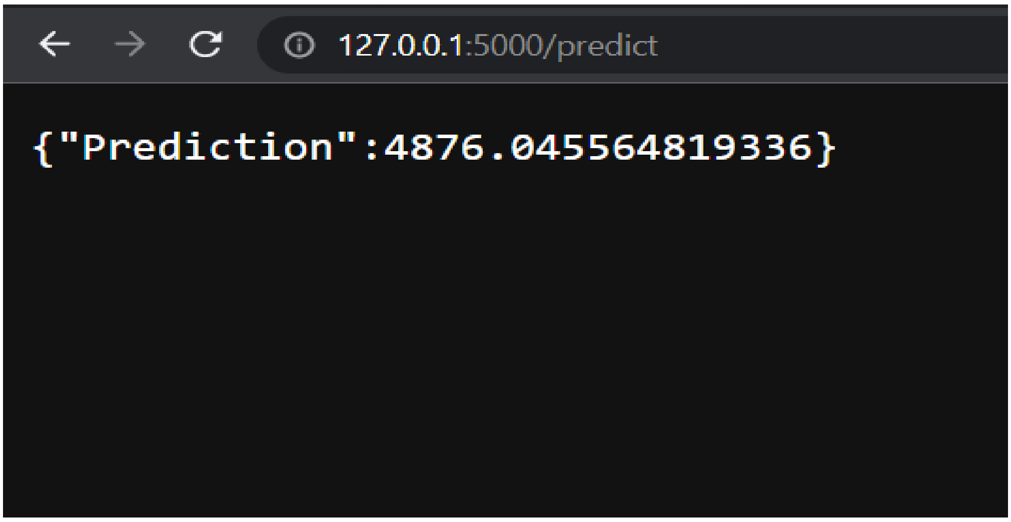
1000
Enter Item MRP

2005
Outlet Establishment Year (YYYY)

High

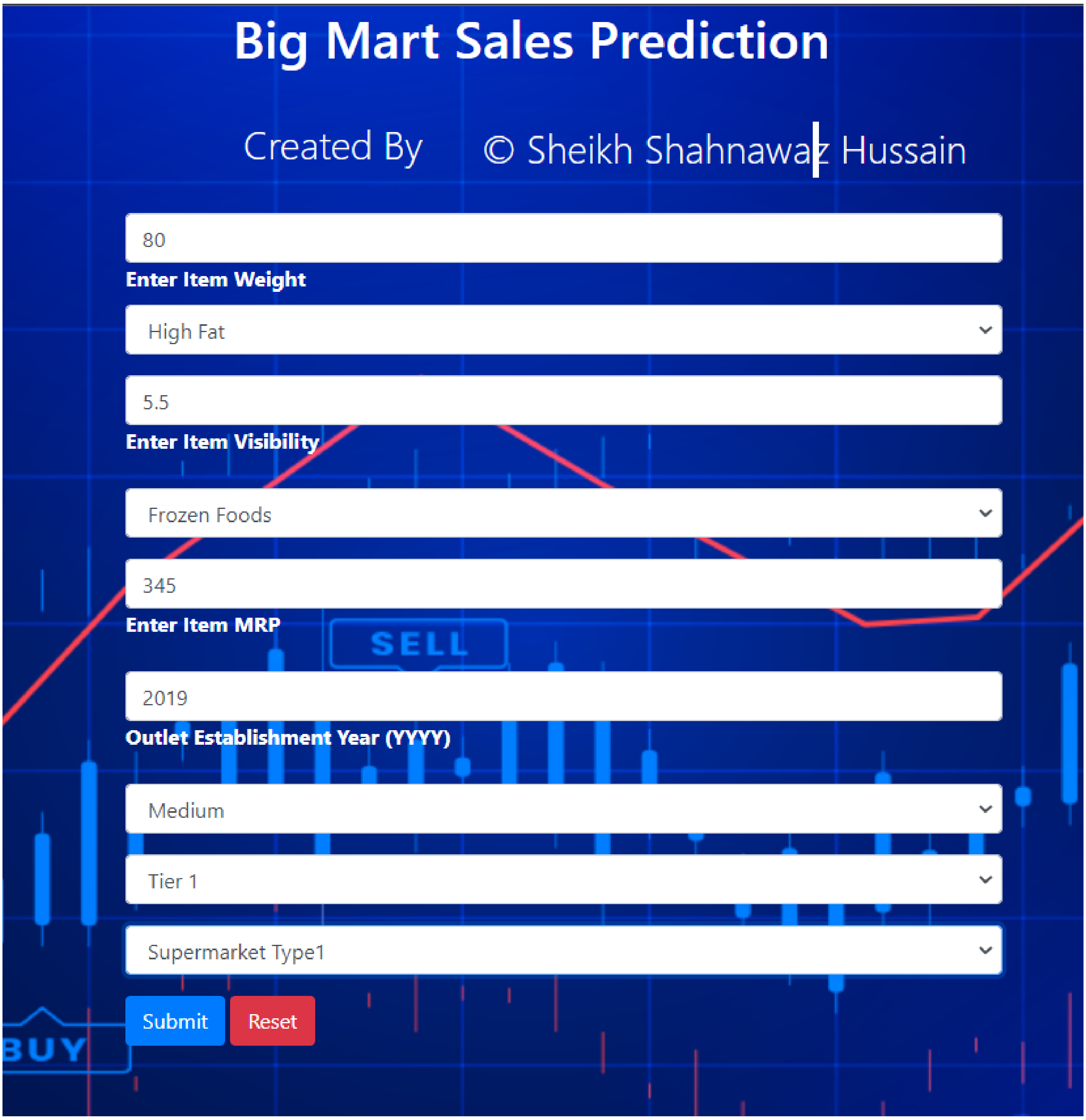
Tier 2

Supermarket Type3



127.0.0.1:5000/predict

```
{"Prediction":4876.045564819336}
```



Big Mart Sales Prediction

Created By © Sheikh Shahnawaz Hussain

80
Enter Item Weight

High Fat

5.5
Enter Item Visibility

Frozen Foods

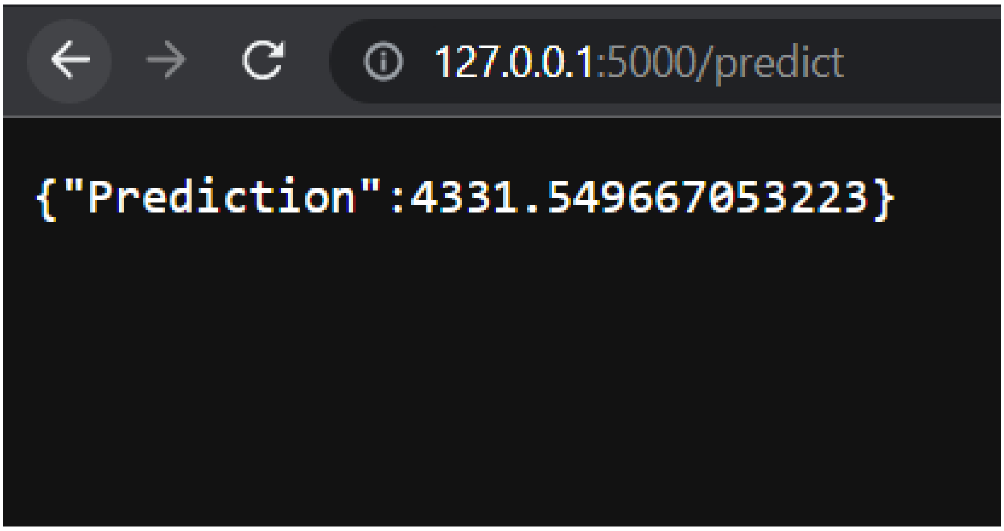
345
Enter Item MRP

2019
Outlet Establishment Year (YYYY)

Medium

Tier 1

Supermarket Type1



127.0.0.1:5000/predict

```
{"Prediction":4331.549667053223}
```

6. CONCLUSION AND FUTURE WORK

The usage of machine learning approaches proves to be a crucial feature for designing business plans while taking into consideration consumer buying habits, as traditional methods are not very helpful to commercial organisations in revenue growth. Businesses can adopt effective tactics for expanding sales and stepping unafraid into the competitive world by using sales predictions based on a variety of factors, including past sales. There is a certain degree of predictability when it comes to sales patterns, according to the analysis and modelling done for the sales forecasting project. It is feasible to estimate future sales success by examining previous data and applying the right forecasting tools. However, it's crucial to remember that sales projections are not always accurate and can be impacted by a range of outside factors, including adjustments to the economy, changes in consumer behaviour, and unanticipated events. Overall, sales forecasting is a useful tool for organisations to plan and make defensible judgements about their operations, but it should only be used as a guide. Businesses can more successfully traverse the complexity of the market and accomplish their objectives by fusing the insights received through sales forecasting with other pertinent information and knowledge.

In conclusion, sales forecasting plays a critical role in firms for successful planning, resource allocation, and decision-making. Future sales patterns may be effectively forecasted using advanced techniques like linear regression. The following are some crucial points and potential areas for additional sales forecasting research:

- Data amount and Quality: The quality and amount of data that is accessible has a significant impact on how accurately sales projections are made. In the future, efforts might be directed on enhancing data gathering procedures, guaranteeing data consistency, and investigating other sources of pertinent data. The forecasting models can be improved by include external elements like economic statistics, market movements, or weather patterns.

- Feature engineering is the process of finding and developing accurate predictors that represent the underlying dynamics and patterns of the sales data. The precision of sales projections can be increased by investigating new features or changing current ones. The most pertinent predictors can be found by using feature selection approaches.

- Time-Series Analysis: Seasonality and time-dependent trends are frequently seen in sales data. To capture the innate time-dependent patterns and boost forecasting accuracy, future studies may utilise time-series analytic methods as ARIMA (AutoRegressive Integrated Moving Average), exponential smoothing approaches, or state-space models.

- Even though linear regression is a popular method for sales forecasting, looking into more advanced machine learning models can produce better outcomes. Non-linear correlations, interactions, and complex patterns in the data can be captured by models like decision trees, random forests, gradient boosting, and neural networks.

- Ensemble Techniques: Ensemble techniques integrate many forecasting models to provide forecasts that are more precise. Future research can concentrate on creating ensemble models that take advantage of the advantages of various methodologies, such as fusing predictions from other models or integrating linear regression with other machine learning models.

- Evaluation and model selection: Choosing the best strategy requires carefully assessing the effectiveness of forecasting models. The comparison of various models using appropriate assessment criteria and the execution of rigorous statistical tests to ascertain the significance of performance differences might be the subject of future research.

- Real-Time Forecasting: In the fast-paced corporate world of today, real-time sales forecasting is becoming more and more crucial. Future research should focus on creating models that can deliver precise projections in a timely manner, incorporate the most recent data, and adjust to shifting market conditions.

- Continuous Model Improvement: As new data become available, sales forecasting models should be regularly reviewed and updated. Over time, the precision and applicability of predictions may be increased by integrating feedback loops and methods for model retraining.

FUTURE WORK:

Future research in a number of areas related to sales forecasting can help close some research gaps and raise the precision and dependability of predictive models. Future studies will focus on a number of crucial topics, including:

1) Real-time prediction**:** The ability to predict sales in real-time can be a game-changer for businesses, allowing them to quickly adjust their strategies based on market trends and customer behaviour. User can Upload their CSV Files of any organization and get the future sales.

2) We intend to expand this research to a larger scale in the future so that different embedding models can be considered on a wider range of datasets.

3) To create an android application based on this machine learning model.

4) To integrate this model to an offline based system so that it runs completely without an internet connection.

5) To Deploy the Project on Platform like AWS so that anyone with the Link can have access to it.

Researchers may contribute to increasing the precision and dependability of sales forecasting models, allowing businesses to deploy their resources more effectively and make better business decisions, by focusing on these upcoming study areas.

REFERENCES

[1] https://www.kaggle.com/datasets/shivan118/big-mart-sales-prediction-datasets

[2] https://www.saleshacker.com/sales-forecasting-101/

[3] https://hbr.org/1971/07/how-to-choose-the-right-forecasting-technique

[4] https://www.liveplan.com/blog/the-best-way-to-forecast-sales-and-revenue/

[4] https://repository.upenn.edu/cgi/viewcontent.cgi?article=1186&context=marketing_papers

[5] https://portal.abuad.edu.ng/lecturer/documents/1588001295ICH_254_NOTE_2.pdf

[6] Gudmundsson, K., & Helgason, H. (2020). Machine learning in sales forecasting. Available at: https://www.arnia.com/blog/machine-learning-in-sales-forecasting

[7] Hu, Y., Yang, H., & Chao, C. (2018). Sales Forecasting with a Hybrid Method of ARIMA and XGBoost. In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 2813-2818). IEEE.

[8] Sarkar, S., Rana, S., & Kumar, S. (2021). Sales forecasting using machine learning algorithms: A comprehensive review. Journal of Retailing and Consumer Services, 59, 102376.

[9] https://www.yesware.com/blog/sales-forecast/

[10] https://blog.hubspot.com/sales/accurate-sales-forecasting-model-tips

[11] https://whatfix.com/blog/sales-forecasting/

[12] https://www.scoro.com/blog/sales-forecasting/

[13] https://www.freshworks.com/crm/sales-forecasting/

[14] https://www.klipfolio.com/resources/articles/sales-forecasting

[15] https://www.insightsquared.com/sales-forecasting/