**FINAL REPORT**

**THE UNIVERSITY OF TEXAS AT ARLINGTON**

**ANALYSIS OF SUPERVISED LEARNING MODELS ON RETAIL SALES DATA**

**Professor:** Ramakrishna Koganti

**Course**: CSE-5301-001

**PROJECT GROUP -3**

**Team members:**

* L R Harshita Pasumarthy - 1002133962
* Karishma Thama - 1002113365
* Mahin Roddur - 1001727433
* Sameer Khan Pathan - 1002102704

**Introduction:**

Accurate sales forecasting is crucial for businesses in the retail industry to make informed decisions about production, inventory management, and resource allocation. However, inaccurate forecasting can lead to ineffective decision-making, resulting in significant financial losses. This project aims to compare three supervised learning models, Decision tree algorithm, Regression algorithm, and Gaussian Naive Bayes classifier, for sales forecasting in the retail industry. The goal is to determine which algorithm provides the most accurate and reliable sales forecasts.

**Problem Definition:**

The problem addressed in this project is inaccurate sales forecasting in the retail industry, which can lead to inefficient decision-making. This can result in overproduction, excess inventory, and suboptimal allocation of resources, leading to financial losses for businesses.

Before moving onto selecting a model, we needed to choose a dataset to work with. This is the dataset that was used for analysis: [Store Item Demand Forecasting (Kaggle).](https://www.kaggle.com/competitions/demand-forecasting-kernels-only/overview) While it does not directly depict sales monetarily, demand is generally proportional to sales revenue, and it also models consumer behavior better than raw sales figures would.

**Predictive Model Selection:**

Three algorithms, Decision tree Regression, Linear Regression and Gaussian Naive Bayes classifier, were selected for this project based on their common usage in the retail industry for sales forecasting.

Decision tree regression creates a decision tree by iteratively segmenting the data into subsets according to the most significant variable. Due to its interpretability and capacity for handling non-linear relationships between features and target variables.

Linear regression describes the connection between two variables. Due to its ease of use and capacity to identify linear correlations between characteristics and target variables, it is a well-liked option for sales forecasting.

Gaussian Naïve Bayes is a probabilistic algorithm. It makes the supposition that the characteristics have a Gaussian distribution and are independent. It is a well-liked option for sales forecasting because of how easy it is to use and how many elements it can manage.

These algorithms were chosen for this research based on their widespread use in the retail industry for sales forecasting, in addition to the appropriateness for the features and target variables in the dataset.

**Analysis:**

The analysis of the three models was performed using retail sales data, and the accuracy and reliability of each model were evaluated. The results showed that the decision tree regressor provided the most accurate and reliable sales forecasts, followed by the Gaussian Naive Bayes classifier and the Linear regression algorithm.

Here’s a summary of the error metrics of each classifier:

|  |  |  |
| --- | --- | --- |
| Algorithm | Root Mean Square Error (RMSE) | Mean Absolute Error (MAE) |
| Linear Regression | 263.402 | 216.859 |
| Decision Tree Regressor | 95.911 | 62.697 |
| Gaussian Naive Bayes Classifier | 263.402 | 216.859 |

**Conclusion:**

The decision tree regressor is the most accurate and reliable method for sales forecasting in the retail industry, followed by the Gaussian Naive Bayes classifier and the Linear Regression algorithm. Accurate sales forecasting can help businesses maintain better inventory, plan, and allocate resources in an efficient way, and improve revenue.

**Recommendations:**

Businesses in the retail industry should consider using the decision tree model for sales forecasting to make informed decisions about production, inventory management, and resource allocation. They should also ensure that the sales data used for forecasting is accurate and up to date to improve the accuracy of the forecasts. Further research can be done to evaluate the performance of other supervised learning models for sales forecasting in the retail industry.

**References:**

1. https://www.datacamp.com/tutorial/decision-tree-classification-python

2. <https://towardsdatascience.com/predicting-sales-611cb5a252de>

3.https://www.kaggle.com/code/sherpalakpa18/sales-forecasting-using-lr-rf-and-xgboost

4.https://www.kaggle.com/competitions/demand-forecasting-kernels-only/overview

**Appendix:**

Python language is used to implement the project. This uses scikit-learn library to train a decision tree for demand forecasting. The model is trained on historical sales data from the “train.csv” file and then used to make predictions on the “test.csv” file.

**Overview of core functions in code:**

|  |  |  |
| --- | --- | --- |
| **Sl No** | **Method** | **Functionality** |
| 1 | pd.read\_csv() | Reads a CSV file and returns a pandas DataFrame object |
| 2 | pd.to\_datetime() | Converts a column of a pandas DataFrame object to a datetime format |
| 3 | groupby() | Groups the rows of a pandas DataFrame object based on one or more columns |
| 4 | reset\_index() | Resets the index of a pandas DataFrame object |
| 5 | DecisionTreeRegressor() | Creates a Decision Tree Regressor model object |
| 6 | train\_test\_split() | Splits a dataset into training and validation sets for machine learning |
| 7 | fit() | Trains a machine learning model using the input training data |
| 8 | predict() | Uses a trained machine learning model to make predictions on new data |
| 9 | mean\_absolute\_error() | Calculates the mean absolute error between the predicted and actual values |
| 10 | mean\_squared\_error() | Calculates the mean squared error between the predicted and actual values, and can also calculate RMSE |
| 11 | accuracy\_score() | Calculates the accuracy of a machine learning model for classification tasks |
| 12 | precision\_score() | Calculates the precision of a machine learning model for classification tasks |
| 13 | confusion\_matrix() | Calculates the confusion matrix of a machine learning model for classification tasks |
| 14 | plot\_tree() | Plots a decision tree visualization for a Decision Tree model object |
| 15 | heatmap() | Creates a heatmap visualization of a pivot table in a pandas DataFrame object |
| 16 | pivot\_table() | Creates a pivot table in a pandas DataFrame object with specified rows, columns, and aggregation functions |
| 17 | LinearRegression() | Creates a Linear Regression model object |

**Code Snippets:**

**Note:** These are not complete code snippets, the execution file is shared in the submission.

**A picture containing text

Description automatically generated**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

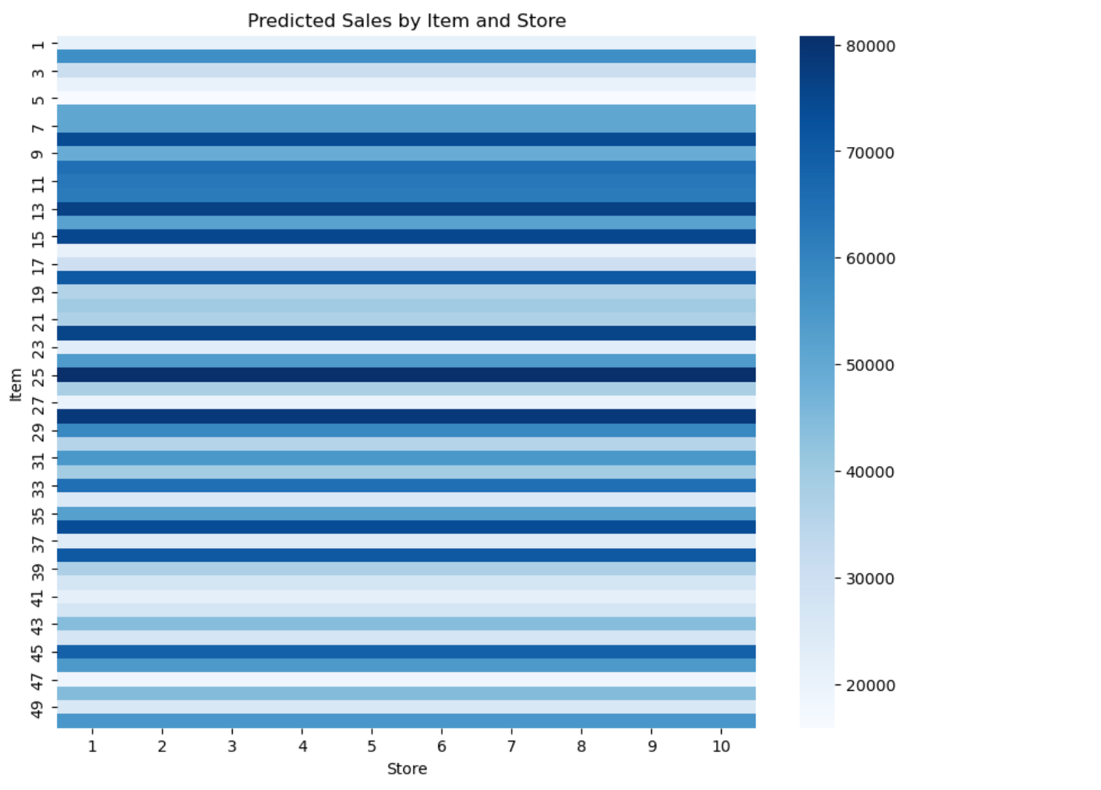
Description automatically generated

Chart

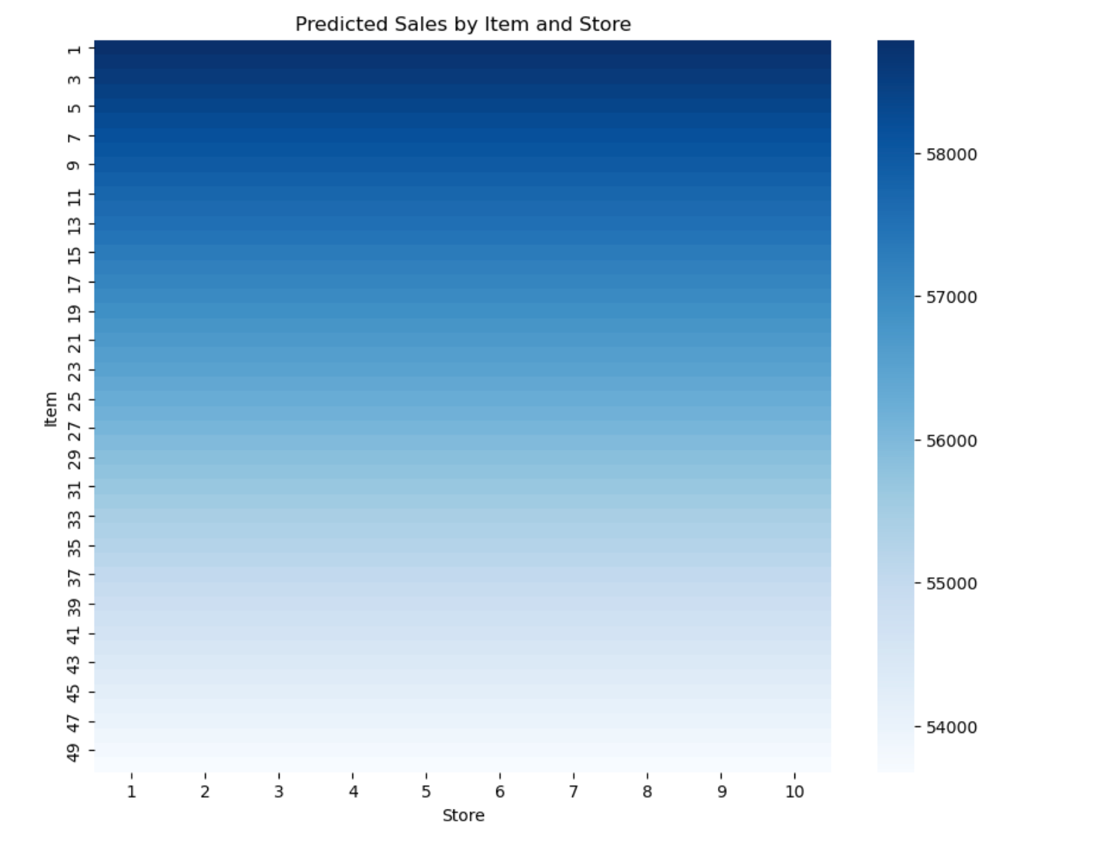
Description automatically generated with medium confidence

Predictions based on Store and Item:

Decision Tree Regressor:



Linear Regression:



Guassian Naïve Bayes Classifier:

