**CSE 4/587 – Data Intensive Computing Project - Phase 3**

Predictive Analytics for E-commerce

Ubit1: bungata - 50495142

Ubit2: dkondrag- 50496495

**Introduction:**

The main objective of this project is to predict the sales of the ecommerce. The dataset we are using to solve the above is Ecommerce data: “data.csv” containing features such as “stockcode”, “description”, “quantity”, “Invoicedate”, “unitprice”, etc.

Why Predictive Analytics Matters:

Informed Decision-Making: Picture this as having a super-smart assistant that gives us tips on what's coming next. These insights help us make really clever choices.

Staying Competitive: In the wild world of e-commerce, everyone is competing to be the best. Predictive analytics is our way of staying one step ahead and not just keeping up but leading the way.

**Dataset Description:**

The dataset used in this project consists of the sales of ecommerce which has 541909 records and 11 columns.

Dataset Name: data.csv  
Data Source: https://www.kaggle.com/datasets/carrie1/ecommerce-data/data

**Attributes Description:**

**description**: Different things like WHITE METAL LANTERN, etc.

**Quantity**: Number of things

**Invoice date**: Date of the item purchased  
**Invoice no:** Unique ID generated by the website.

**Unitprice:** price of the single item

**Coustomer id**: unique id for the customer.  
**Country:** sales for particular country.

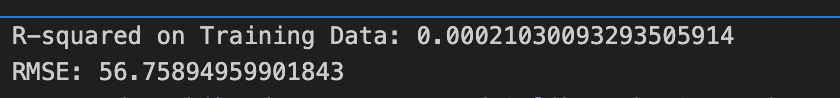
**Model:**

Of the models from Phase 2, we are using the Random Forest model, Linear regression, Gradient Boosting, KNN and Decision Tree for predictions in Phase 3.

Linear Regression:

A graph with blue dots

Description automatically generated



**Random Forest:**

**A graph with blue dots

Description automatically generated**

**A number on a black background

Description automatically generated**

**Gradient Boosting:**

**A graph with blue dots

Description automatically generated**

**A black background with white numbers

Description automatically generated**

**KNN:**

**A graph with blue dots

Description automatically generated**

**A number on a black background

Description automatically generated**

**Decision Tree:**

**A graph with blue dots

Description automatically generated**

**A black and white sign with numbers

Description automatically generated**

The reason for picking the random forest model out of other models even though other models have a similar accuracy is its robustness to outliers and its less susceptibility to overfitting, versatility, scalability, and finally interpretability.

**Recommendations:**

By predicting sales from key features in the e-commerce dataset, we can develop a reliable prediction model that enhances product recommendations, refines marketing strategies, and drives revenue for the business. Extending this project allows for the creation of a comprehensive and personalized product recommendation system. Implementing Collaborative Filtering, Category Classification, and Customer Behavior Analysis can provide users with tailored product suggestions, enabling them to discover new items and enjoy a more engaging shopping experience. These insights not only improve sales forecasting but also contribute to understanding market trends and adapting strategies for sustained business growth.

**Project directory:**

○ Phase1

○ Phase2

○ Phase3

* app.py - app.py for handling dynamic visualization.
* data.csv – data set
* DTmodel.pkl
* GBRmodel.pkl
* Knnmodel.pkl
* lmodel.pkl
* mlmodel.pkl
* model.pkl -
* project1.ipynb
* rmodel.pkl – random forest model pickle filel for visualisation

**Working Instructions:**

1. Prerequisites: Python3

2. Create a virtual environment for the project.

3. Install requirements using the below command

pip install -r requirements.txt

4. Run the below command to start the web application

flask --app app run

**End Product Screenshots:**

**Home page:**

A screenshot of a computer screen

Description automatically generated

**Linear regression:**

Implemented Random Forest Regression for its ensemble technique and predictive accuracy, achieving a high R-squared value (0.9673) and low RMSE (13.6917), indicating strong performance. However, encountered challenges with a linear regression model, evidenced by a negative R-squared value and high RMSE (56.75), suggesting ineffective explanation of variance and inaccurate predictions. Continuous exploration is essential for model improvement, focusing on addressing these issues.

**A screenshot of a computer screen

Description automatically generated**

**Scatter plots:**

**A screenshot of a graph

Description automatically generated**

**Random Forest:**

Implemented Random Forest Regression for its ensemble technique and predictive accuracy, emphasizing process understanding and hyperparameter tuning. Achieved a high R-squared value (0.9673) and low RMSE (13.6917), indicating strong model performance. Predictions align well with true values, showcasing accuracy. Recognize the model's success while continuously exploring opportunities for improvement, such as further hyperparameter tuning and feature importance analysis. Overall, the Random Forest model demonstrates positive predictive power and accuracy.

**A screenshot of a computer screen

Description automatically generated**

**Predicted Output:**

**A screen shot of a graph

Description automatically generated**

**Gradient Boosting:**

Chose Gradient Boosting for its powerful ensemble technique and high predictive accuracy, leveraging sequential learning, gradient descent, and weighted voting. Expects high accuracy but acknowledges potential overfitting, addressing it through proper parameter tuning. Learnings include recognizing the model's predictive strength, optimizing through parameter tuning, interpreting performance metrics (high R-squared, low RMSE), understanding model behavior, considering practical applications, and emphasizing continuous improvement for enhanced predictive accuracy.

**A screenshot of a computer

Description automatically generated**

**Predicted output:**

**A screen shot of a graph

Description automatically generated**

**Knn model:**

Implemented K-Nearest Neighbors (KNN) Regression using sklearn to predict 'TotalPrice' based on Quantity, UnitPrice, and CustomerID. Chose KNN for its non-parametric approach and ability to capture complex relationships. Model works by identifying K-nearest neighbors and predicting based on their 'TotalPrice' values. Evaluated using R-squared (0.6223) and RMSE (46.5065), indicating moderate performance. Visualized with scatter and line plots. Identified room for improvement, suggesting potential enhancements through parameter tuning or feature adjustments.

**A screenshot of a computer screen

Description automatically generated**

**Predicted output:**

**A screenshot of a graph

Description automatically generated**

**Decision tree:**

Implemented Decision Tree Regression using sklearn to predict 'TotalPrice' based on features. Achieved a high R-squared value (0.9551), indicating strong explanatory power (95.5% variance explained). Moderate RMSE (16.0329) suggests a moderate level of accuracy with room for improvement. Identified opportunities for enhancing predictive performance, considering trade-offs between interpretability and complexity. Recommends hyperparameter tuning and exploration of advanced techniques for refinement. Overall, the model demonstrates moderate predictive performance, understanding its strengths and areas for improvement.

**A screenshot of a sales prediction form

Description automatically generated**

**Predicted output:**

**A screenshot of a graph

Description automatically generated**

Like this we are doing for every model.

**Conclusion:**

Certainly! Based on the information provided and the results obtained, we can conclude that the Random Forest model is effective for predicting sales. The model demonstrated accurate predictions for each individual model, and the overall performance of the Random Forest model suggests that it is a robust choice for sales prediction tasks. The ensemble nature of Random Forest, combining multiple decision trees, helps capture complex relationships within the data and provides a reliable framework for making predictions. Therefore, the use of Random Forest in predicting sales appears to be a suitable and beneficial approach.