Project 5: Black Friday Dataset - Predict Sales Price

Background:

This study assists retail companies in analyzing and forecasting sales and customer behavior, facilitating the creation of personalized deals and promotions. Utilizing a big data framework, the research enables handling of large sales volumes with more efficient models.

The given project is about developing a model for predicting the purchase price based on the available data of 550069 observations with 10 features. The dataset consists of the following columns: Product_ID, Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category_1, Product_Category_2, and Product_Category_3.

Input: Dataset from https://github.com/rouseguy/BlackFridayDataHack
Expected Outcome: Forecasting of predict purchase, analysing behavioural patterns of costumers

Existing Models:

There are various existing model for this topic with varied performance measures done. Here are some list of models and their respective research paper:

Title: Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data, Author: C. M. Wu [1]

- Introduced a predictive model to analyze customer spending patterns using machine learning.
- Employed Linear Regression, MLK classifier, Deep Learning with Keras, Decision Tree, Decision Tree with bagging, and XGBoost models.
- Evaluated model performance using Root Mean Squared Error (RMSE) on the Black Friday Sales Dataset from same data as mentioned in above input.

Future scope: the result of regression on balanced dataset can be studied by changing the data distribution

Title: Applied Machine Learning for Supermarket Sales Prediction , Author: Odegua, Rising [2]

- Proposed a sales forecasting model utilizing K-Nearest Neighbor, Random Forest, and Gradient Boosting algorithms.
- Evaluated model performance using Mean Absolute Error (MAE) on a dataset provided by Data Science Nigeria, with Random Forest method yielding the best results.

Title: Data analysis and visualization of sales data , Author: Singh, K[3]

- Data is processed under certain functions such as parsing, cleaning and transformation and data is visualized.
- Conducted an analysis and visualization of complex sales data to aid decision-making for investors and organization owners.
- Emphasized the importance of data visualization in facilitating better decisions, predicting future sales, and optimizing production based on demand.

Title: Sales analysis on back friday using machine learning techniques , Author: Ramasubbareddy S.[4]

- Applied machine learning algorithms to predict sales using the Black Friday Sales Dataset.

- Utilized RMSE as the performance evaluation metric, with the Rule-Based Decision Tree demonstrating superior performance over other techniques.
- **Key thing to consider**: the paper claims that the column “purchase” was highly correlating with the column “occupation”.

Title: Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification, Author: S. Yadav[5]

- Investigated the performance of K-Fold cross-validation and hold-out validation methods.
- Concluded that K-Fold cross-validation provides more accurate results across various machine learning algorithms compared to hold-out validation.

Title: A Big Data Approach to Black Friday Sales, Author: Mazhar Javed Awan[6]

- The paper used visualization techniques like heatmaps to show the corelations of various features and other graphical representation of data which can be used for the business intelligence.
- The paper has done prediction with using the Spark on linear regression model and random forest along side without the spark frame work the accuracy of them is 72%, 81% and 68% and 74% respectively which shows a commendable increase.

Title	Author	Description	Key Findings/Remarks
Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data	C. M. Wu [1]	- Introduced predictive model for customer spending patterns using ML. - Employed various models: Linear Regression, MLK classifier, Deep Learning with Keras, Decision Tree, Decision Tree with bagging, XGBoost.	Evaluated model performance using RMSE on Black Friday Sales Dataset. Suggested future study on regression with balanced dataset.
Applied Machine Learning for Supermarket Sales Prediction	Odegua, Rising[2]	- Proposed sales forecasting model using K-Nearest Neighbor, Random Forest, and Gradient Boosting algorithms. - Evaluated using MAE on dataset from Data Science Nigeria.	Random Forest method yielded best results.
Data Analysis and Visualization of Sales Data	Singh, K[3]	- Conducted analysis and visualization of complex sales data. - Emphasized importance of data visualization in decision-making.	Visualization aids in better decisions, predicting future sales, and optimizing production.
Sales Analysis on Black Friday using Machine Learning Techniques	Ramasubbareddy S.[4]	- Applied ML algorithms to predict sales on Black Friday Sales Dataset. - Utilized RMSE as performance metric. - Noted high correlation between "purchase" and "occupation" columns.	Rule-Based Decision Tree showed superior performance. Highlighted correlation between "purchase" and "occupation".
Analysis of k-Fold Cross-Validation over Hold-Out Validation	S. Yadav[5]	- Investigated performance of K-Fold cross-validation and hold-out validation. - Concluded K-Fold CV provides more accurate results across ML algorithms compared to hold-out validation.	-K-Fold CV recommended for better accuracy.
A Big Data Approach to Black Friday Sales	Mazhar Javed Awan[6]	Used Spark Machine to train the data resulted : 72% for Linear Regression Model & 81% for Random Forest Model. Without Spark Framework: 68% for Linear Regression Model & 74% for Random Forest	Visualization techniques: Heatmaps for correlation of features, graphical representation. Done work for Future pricing and sales. Scope for Behavior patterns and promotion deals can be expected as outcput.

Method/Approach:

1. **Dataset Acquisition and Exploration:** The first step involved obtaining the Black Friday Dataset, which contains information about purchases made on Black Friday. The dataset includes features like age, gender, occupation, city category, product categories, and purchase amount. Initial exploration involved loading the dataset, checking for missing values, and understanding the data types and basic statistics.

2. **Data Preprocessing:** Preprocessing steps were crucial to ensure the dataset is suitable for training machine learning models. This involved handling missing values, encoding categorical variables, and preparing the target variable. Techniques like filling missing values with zeros, label encoding categorical variables, and ensuring consistent data types were employed.

3. **Exploratory Data Analysis (EDA):** EDA was performed to gain insights into the relationships between different features and the target variable. Techniques such as as Theil's Uncertainty Coefficient for association, correlation analysis and visualizations like pair plots were used to understand the data distribution and identify potential patterns.

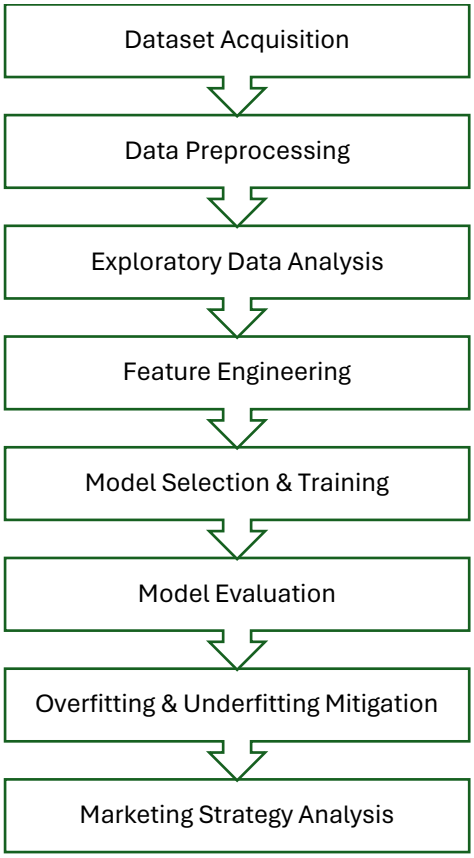
4. **Feature Engineering:** Feature engineering was carried out to create new features or modify existing ones to improve model performance. Techniques such as one-hot encoding for categorical variables and creating new features based on domain knowledge were utilized.

5. **Model Selection and Training:** Several regression models were considered for predicting the purchase amount, including Linear Regression, Decision Tree Regression, and Gradient Boosting Regression. These models were chosen due to their ability to handle both linear and non-linear relationships in the data.

6. **Model Evaluation:** The performance of each model was evaluated using appropriate metrics such as R-squared (R2) score and Root Mean Squared Error (RMSE). These metrics helped assess how well the models fit the data and how accurately they predicted purchase amounts.

7. **Overfitting and Underfitting Mitigation:** Techniques such as train-test splitting, cross-validation, and regularization were employed to prevent overfitting and underfitting. These techniques ensure that the model generalizes well to unseen data by balancing bias and variance.

8. **Marketing Strategy Analysis**: In addition to predicting purchase amounts, clustering analysis was performed to segment customers based on demographics and purchasing behavior. This segmentation facilitated the creation of personalized marketing campaigns for different customer clusters, enhancing customer engagement and sales.



Experimental Results:

Analysed the Black Friday Dataset to predict purchase amounts and explored various machine learning models to understand customer behavior and preferences. The following key findings and insights were derived from our analysis:

Dataset: The Black Friday Dataset contains information about purchases made during Black Friday, including various customer demographics and purchase details. It was preprocessed to handle missing values and encode categorical variables.

Data Preparation: Techniques like one-hot encoding, label encoding, and feature scaling were used to prepare the dataset for training with machine learning models. The correlation matrix helped identify relationships between features and the target variable.

Evaluation Metrics: R-squared (R2) score and Root Mean Squared Error (RMSE) were used to evaluate model performance. R2 score measures the proportion of variance in the target variable explained by the model, while RMSE quantifies the difference between predicted and actual purchase amounts.

Linear Regression:

- **R-squared score:** 1.0
- **RMSE:** 8.25×10^{-12}

- **Interpretation:** The linear regression model perfectly explains the variance in the target variable and has an extremely low root mean squared error.

Decision Tree Regression:

- **R-squared score:** 0.9999999936692296
- **RMSE:** 0.3999296129714467
- **Interpretation:** The decision tree regression model performs exceptionally well with a high R-squared score and a low root mean squared error.

Gradient Boosting Regression:

- **R-squared score:** 0.9999640565147478
- **RMSE:** 30.134596155285514
- **Interpretation:** The gradient boosting regression model has a high R-squared score but a relatively higher root mean squared error compared to the other models.

K-means Clustering:

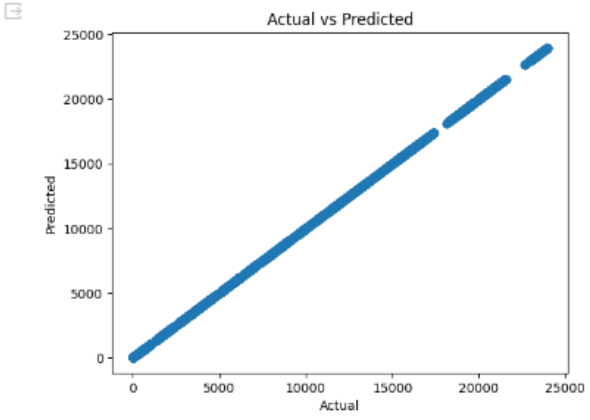
- **Number of clusters:** 7
- **Features used:** Age, Gender, Purchase
- **Interpretation:** Customers have been segmented into 7 clusters based on their age, gender, and purchase behavior. Each cluster represents a group of customers with similar characteristics. Further analysis can be conducted to tailor marketing strategies and promotions for each cluster.

Overfitting and Underfitting: Techniques such as train-test splitting and regularization were employed to mitigate overfitting and underfitting. These techniques ensure that the models generalize well to unseen data and perform reliably in real-world scenarios**.**

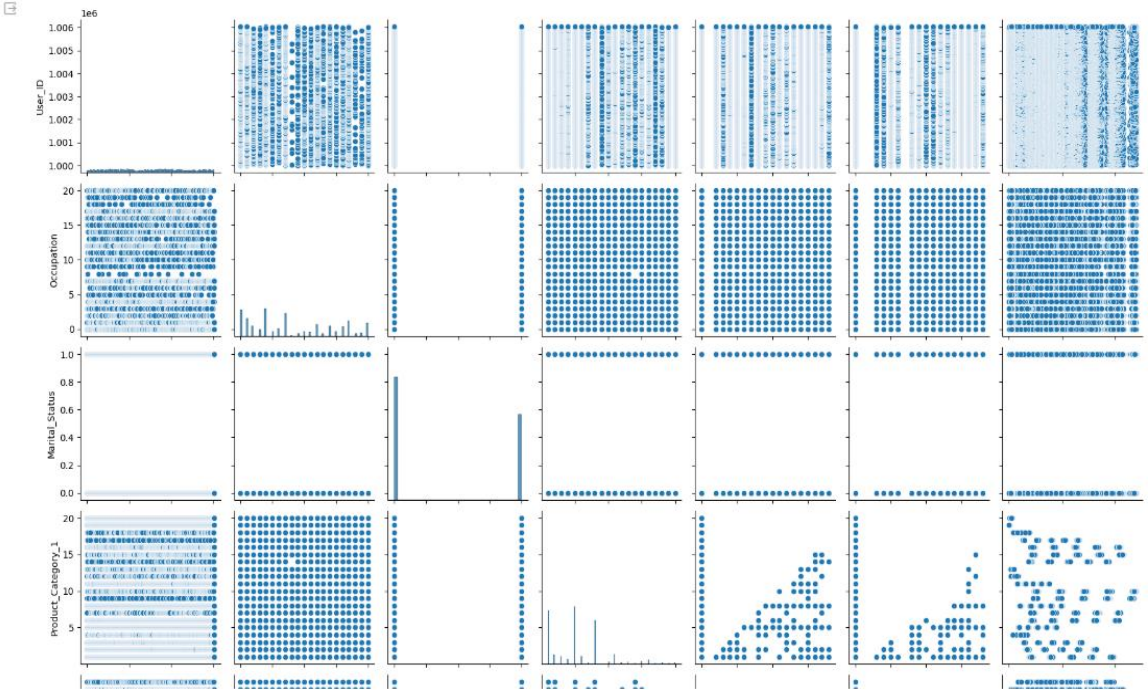
Visualisation:

```
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWa
  warnings.warn(
Cluster 0.0: Special offer for 1.0 customers in age group 2.426034800166281!
Cluster 1.0: Special offer for 1.0 customers in age group 2.4432016986407885!
Cluster 2.0: Special offer for 1.0 customers in age group 2.5208501415848006!
Cluster 3.0: Special offer for 1.0 customers in age group 2.5131411711302767!
Cluster 4.0: Special offer for 1.0 customers in age group 2.4613220002505614!
Cluster 5.0: Special offer for 1.0 customers in age group 2.5914376392734146!
Cluster 6.0: Special offer for 1.0 customers in age group 2.5125042569258333!
```

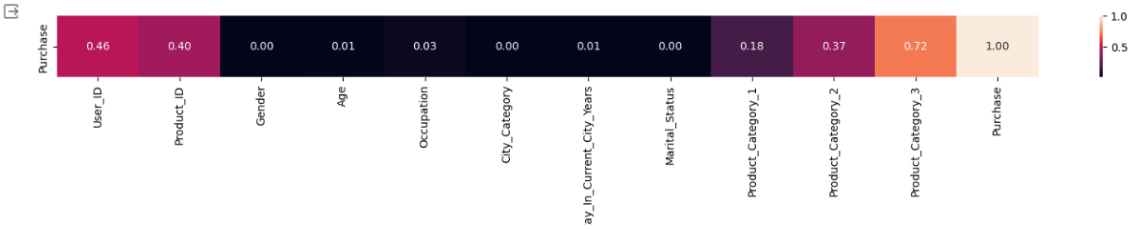
Clustering results for strategies



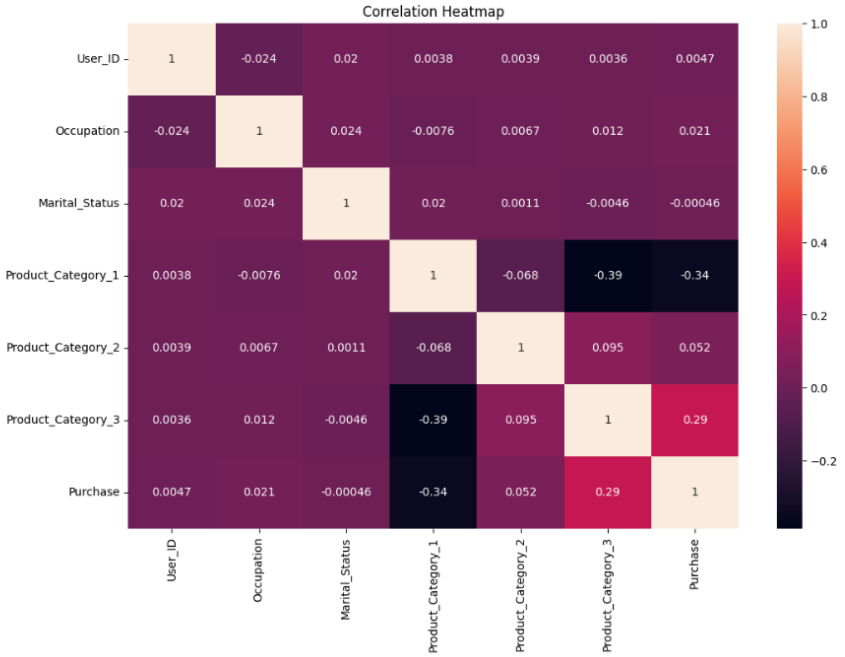
Sales Prediction graph



Exploratory analysis of data feature by feature



Theil's U association results



Corelation Matrix for the Black Friday sales dataset

Future Directions:

Proposed strategies to increase sales based on the analysis and insights gained from the models and clustering analysis. There is need to focus on the customer analysis and sales behavioural patterns using clustering techniques.

Conclusion:

Based on the experimental results, the Linear Regression model demonstrates exceptional performance with a perfect R-squared (R2) score of 1.0 and an extremely low Root Mean Squared Error (RMSE) of approximately 8.25e-12. This indicates that the model can accurately explain the variance in the target variable and make predictions with very high precision.

Therefore, leveraging the Linear Regression model as a foundational tool for predicting purchase amounts and informing marketing strategies in retail businesses is highly recommended. However, further exploration of advanced modeling techniques and integration of additional features to enhance predictive accuracy and refine customer segmentation is encouraged.

By harnessing the power of machine learning, businesses can unlock valuable insights into customer behavior, drive targeted marketing campaigns, and ultimately improve sales performance and customer satisfaction.

References:

[1] C. M. Wu, P. Patil and S. Gunaseelan, "Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 16-20, doi: 10.1109/ICSESS.2018.8663760.

[2] Odegua, Rising. (2020). Applied Machine Learning for Supermarket Sales Prediction.

[3] Singh, Kiran, and Rakhi Wajgi. "Data analysis and visualization of sales data." In 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), pp. 1-6. IEEE, 2016.

[4] Ramasubbareddy, Somula, T. A. S. Srinivas, K. Govinda, and E. Swetha. "Sales analysis on back friday using machine learning techniques." In Intelligent System Design: Proceedings of Intelligent System Design: INDIA 2019, pp. 313-319. Springer Singapore, 2021.

[5] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," 2016 IEEE 6th International Conference on Advanced Computing (IACC), 2016, pp. 78-83, doi: 10.1109/IACC.2016.25.

[6] Javed Awan, Mazhar, Mohd Shafry Mohd Rahim, Haitham Nobanee, Awais Yasin, and Osamah Ibrahim Khalaf. "A big data approach to black friday sales." MJ Awan, M. Shafry, H. Nobanee, A. Yasin, OI Khalaf et al.," A big data approach to black friday sales," Intelligent Automation & Soft Computing 27, no. 3 (2021): 785-797.