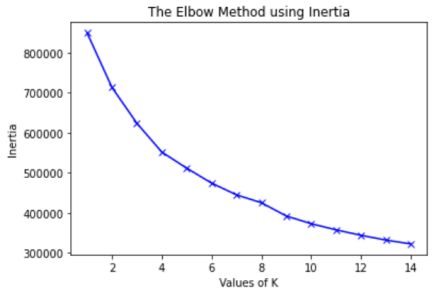
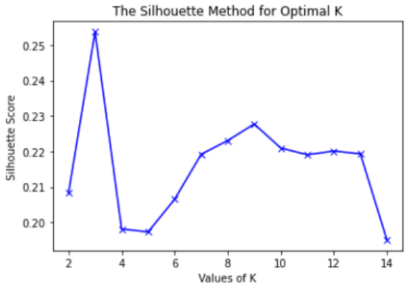
Project Report ML Perspective

Data-Driven Market Segmentation: A Machine Learning Perspective

Feature Engineering & Scaling Techniques

- **Scaling Techniques:** Standardization with StandardScaler (fit and transform).
- **Dimensionality Reduction:** Principal Component Analysis (PCA) was applied to reduce dimensions for better visualization.
- **Visualization:** Transformed dataset into 2D for effective visualization.
- **K-Value Determination:** Used the Elbow Method and Silhouette Score to determine the optimal number of clusters.





Insight: Optimal K-Value = 3

- **Statistical Test:** Conducted a One-Way ANOVA test to analyze feature significance.

```
Column: BALANCE
  F-statistic: 13546.2555
  P-value: 0
  Result: Statistically significant difference between groups

Column: BALANCE_FREQUENCY
  F-statistic: 1817.7030
  P-value: 0
  Result: Statistically significant difference between groups

Column: PURCHASES
  F-statistic: 16006.0741
  P-value: 0
  Result: Statistically significant difference between groups

Column: ONEOFF_PURCHASES
  F-statistic: 9397.5773
  P-value: 0
  Result: Statistically significant difference between groups

Column: INSTALLMENTS_PURCHASES
  F-statistic: 8709.5991
  P-value: 0
  Result: Statistically significant difference between groups

Column: CASH_ADVANCE
  F-statistic: 18280.6685
  P-value: 0
  Result: Statistically significant difference between groups

Column: PURCHASES_FREQUENCY
  F-statistic: 8448.0923
  P-value: 0
  Result: Statistically significant difference between groups

Column: ONEOFF_PURCHASES_FREQUENCY
  F-statistic: 15958.4270
  P-value: 0
  Result: Statistically significant difference between groups

Column: PURCHASES_INSTALLMENTS_FREQUENCY
  F-statistic: 5493.5721
  P-value: 0
  Result: Statistically significant difference between groups

Column: CASH_ADVANCE_FREQUENCY
  F-statistic: 27464.8890
  P-value: 0
  Result: Statistically significant difference between groups

Column: CASH_ADVANCE_TRX
  F-statistic: 16045.4287
  P-value: 0
  Result: Statistically significant difference between groups

Column: PURCHASES_TRX
  F-statistic: 20973.7154
  P-value: 0
  Result: Statistically significant difference between groups

Column: CREDIT_LIMIT
  F-statistic: 8663.4739
  P-value: 0
  Result: Statistically significant difference between groups

Column: PAYMENTS
  F-statistic: 5790.1256
  P-value: 0
  Result: Statistically significant difference between groups

Column: MINIMUM_PAYMENTS
  F-statistic: 1095.2634
  P-value: 0
  Result: Statistically significant difference between groups

Column: PRC_FULL_PAYMENT
  F-statistic: 1677.3445
  P-value: 0
  Result: Statistically significant difference between groups

Column: TENURE
  F-statistic: 383.4717
  P-value: 5.306e-166
  Result: Statistically significant difference between groups
```

Anova Insight:

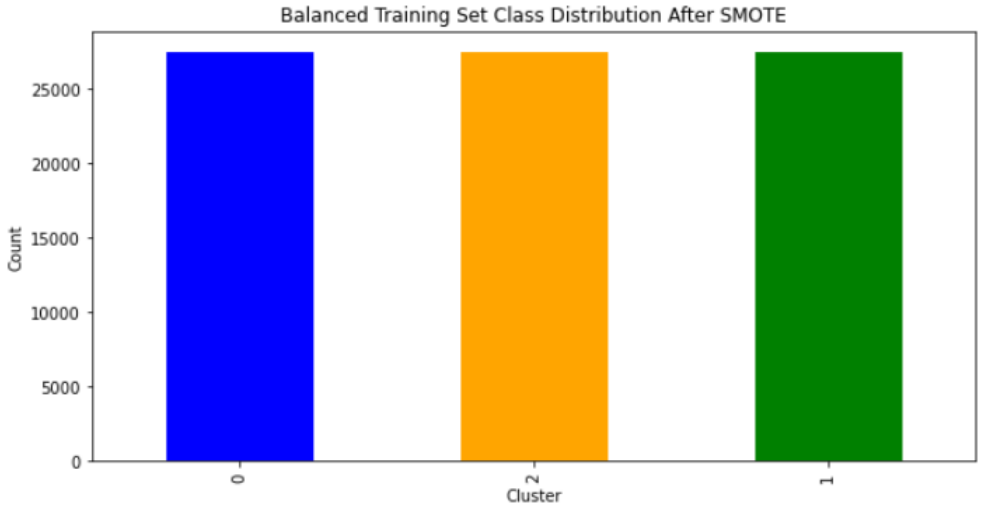
- **P-Value**: All p-values are 0, meaning the differences between clusters are statistically significant.
- **F-Statistic**: High values indicate strong differences between cluster means for each variable.
- **Conclusion**: Each cluster shows distinct customer behaviors (e.g., balances, purchases). This confirms meaningful segmentation and can guide targeted actions.

Data Balancing

SMOTE Technique (Synthetic Minority Over-sampling Technique): Utilized SMOTE to balance the dataset and address class imbalance.

Imbalance Data Set:

After Applying SMOTE Techniques:



Model Selection

Selected Models for Comparison:

a) Decision Tree
b) Random Forest
c) XG-Boost

Reason for Selection: These models are suitable for classification tasks, offer flexibility in handling various data types, and provide feature interpretation insights.

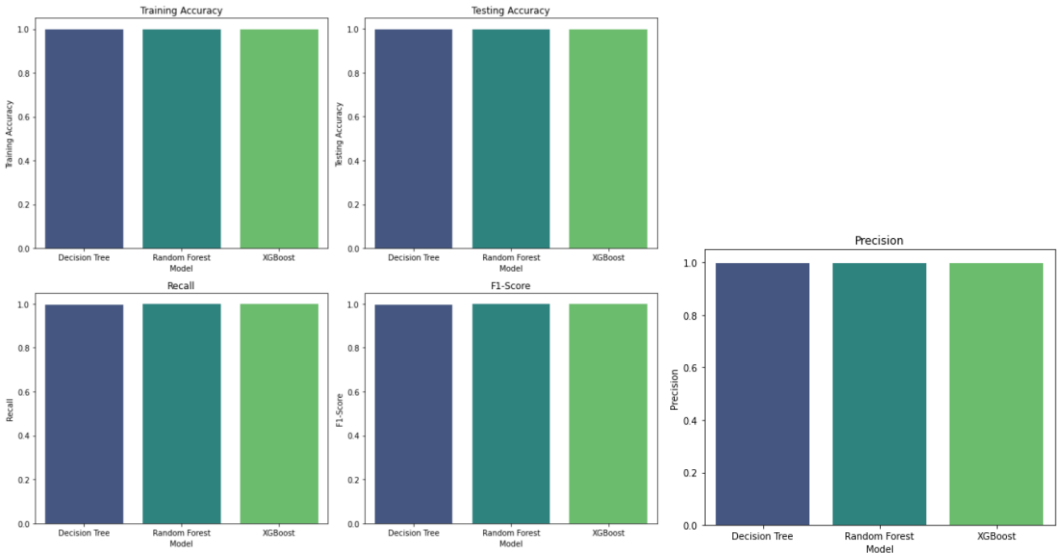
Model Training

Training Process: Each selected model was trained on the preprocessed data. (70-30)

Parameters and Settings: Model-specific parameters were set based on initial tuning efforts.

Comparative Analysis of Evaluation Metrics

Evaluation Metrics: The models were evaluated based on Accuracy (Train and Test), Precision, Recall, and F1-Score.



Results Comparison Table: Comparative results of Accuracy, Precision, Recall, and F1-Score for each model on test data were presented in a table format.

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score
Decision Tree	1.0	0.9991	0.998301	0.999078	0.998689
XGBoost	1.0	0.9994	0.998593	0.999711	0.999151
Random Forest	1.0	0.9994	0.998593	0.999711	0.999151

Summary & Conclusion

- The Decision Tree model was chosen over XGBoost and Random Forest.
- All models achieved similar high accuracy, making them effective for the task.
- XGBoost and Random Forest are ensemble methods with higher computational costs.
- Decision Tree provides comparable accuracy with lower computational requirements.
- Therefore, Decision Tree was the more efficient choice for this task.