Part A: Case Study Report

Part A.1 Machine Learning Solution

Report Title: Machine Learning Solution for GOBank Sales Prediction

Unit Code and Name: [Insert Unit Code and Name]

Student Name and Student ID: [Insert Student Name and ID]

Contents

Figures:

Executive Summary

This report presents a machine learning solution for predicting sales outcomes for GOBank. By employing both supervised and unsupervised learning techniques, we aim to enhance the bank's sales strategies and customer segmentation. Key insights from data analysis and model evaluations are discussed, leading to actionable recommendations for the business.

1. INTRODUCTION

In today's highly competitive financial services sector, customer acquisition and retention are paramount to the success of any banking institution. GOBank, a prominent player in this industry, faces the ongoing challenge of predicting sales outcomes and optimizing marketing efforts to enhance customer engagement and drive revenue growth. Leveraging the power of data analytics and machine learning, this project aims to build robust predictive models that can accurately forecast the likelihood of a sale. This, in turn, will empower GOBank to refine its marketing strategies, personalize customer interactions, and ultimately improve overall business performance.

Objective

The primary objective of this project is to revolutionize GOBank's sales prediction and marketing strategies through the development and deployment of advanced machine learning models. By leveraging customer demographics, behavioral data, and economic indicators, these models aim to accurately predict the likelihood of a sale. This initiative seeks to enhance GOBank's marketing

effectiveness by efficiently allocating resources to target the right customers with personalized offers, thereby optimizing conversion rates and reducing marketing costs.

To achieve this objective, the project encompasses several key tasks, including data preprocessing and exploratory data analysis (EDA), feature engineering and selection, model building and evaluation, hyperparameter tuning and optimization, and clustering analysis. These tasks are designed to ensure the reliability and effectiveness of the predictive models while providing valuable insights into customer behavior and preferences.

The project offers a compelling value proposition by delivering robust predictive models that can transform GOBank's sales processes and marketing strategies. By accurately predicting sales outcomes, GOBank can enhance marketing effectiveness, improve customer engagement, drive revenue growth, and make data-driven decisions. This data-driven approach not only fosters stronger customer relationships but also provides a competitive advantage in the banking sector.

In conclusion, through the integration of data analytics and machine learning, this project aims to empower GOBank with actionable insights that can drive sustainable financial success and secure a leading position in the financial services industry.

2. Approach

Overview of Machine Learning Approach:

- **Types and Problems**: This project involves both supervised learning (for predictive modeling) and unsupervised learning (for customer segmentation).
- **Prediction Targets**: The target variable is the 'Sale Outcome' which indicates whether a sale was made or not.

3. Data Preparation and Exploratory Data Analysis (EDA)

Data Sources, Size, Types, and Quality:

- The dataset comprises 22941 entries with 19 features include target Variable.
- It includes demographic details, contact information, and previous campaign outcomes.
- Initial quality assessment identified missing values and categorical variables requiring preprocessing.

Data Cleaning and Preprocessing:

- Missing values were handled using appropriate imputation methods.
- Categorical variables were encoded using Label Encoding.
- Numerical features were standardized for uniformity.

```python
import numpy as np
from sklearn.preprocessing import LabelEncoder, StandardScaler

# Fill missing values
data['Qualification'].fillna(data['Qualification'].mode()[0], inplace=True)
data['Last Contact Direction'].fillna(data['Last Contact Direction'].mode()[0], inplace=True)
data['Last Contact Duration'].fillna(data['Last Contact Duration'].mean(), inplace=True)
data['Number of Previous Campaign Calls'].fillna(data['Number of Previous Campaign Calls'].mean(), inplace=True)
data['Previous Campaign Outcome'].fillna(data['Previous Campaign Outcome'].mode()[0], inplace=True)

# Encode categorical variables
label_encoders = {}
categorical_columns = ['Qualification', 'Occupation', 'Marital Status', 'Home Mortgage', 'Personal Loan',
                       'Has Other Bank Account', 'Last Contact Direction', 'Last Contact Month', 'Last Contact Weekday',
                       'Previous Campaign Outcome', 'Sale Outcome']

for col in categorical_columns:
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col])
    label_encoders[col] = le

# Feature scaling
scaler = StandardScaler()
numerical_columns = ['Age', 'Last Contact Duration', 'Number of Current Campaign Calls', 'Number of Previous Campaign Calls',
                     'RBA Cash Rate', 'Employment Variation Rate', 'Consumer Confidence Index']
data[numerical_columns] = scaler.fit_transform(data[numerical_columns])

# Display the first few rows of the processed dataset
print(data.head())
```

Figure 1 Data Cleaning and Pre-processing

Exploratory Data Analysis (EDA):

- **Statistical Analysis and Visualization**: Various plots such as correlation heatmaps, count plots, and box plots were generated to understand feature relationships and distributions.
- **Key Insights**:
    - Age and qualification levels significantly influence sales outcomes.
    - Previous campaign outcomes and contact methods are strong predictors of current sales success.
    - Economic indicators like RBA Cash Rate and Employment Variation Rate also impact sales outcomes.

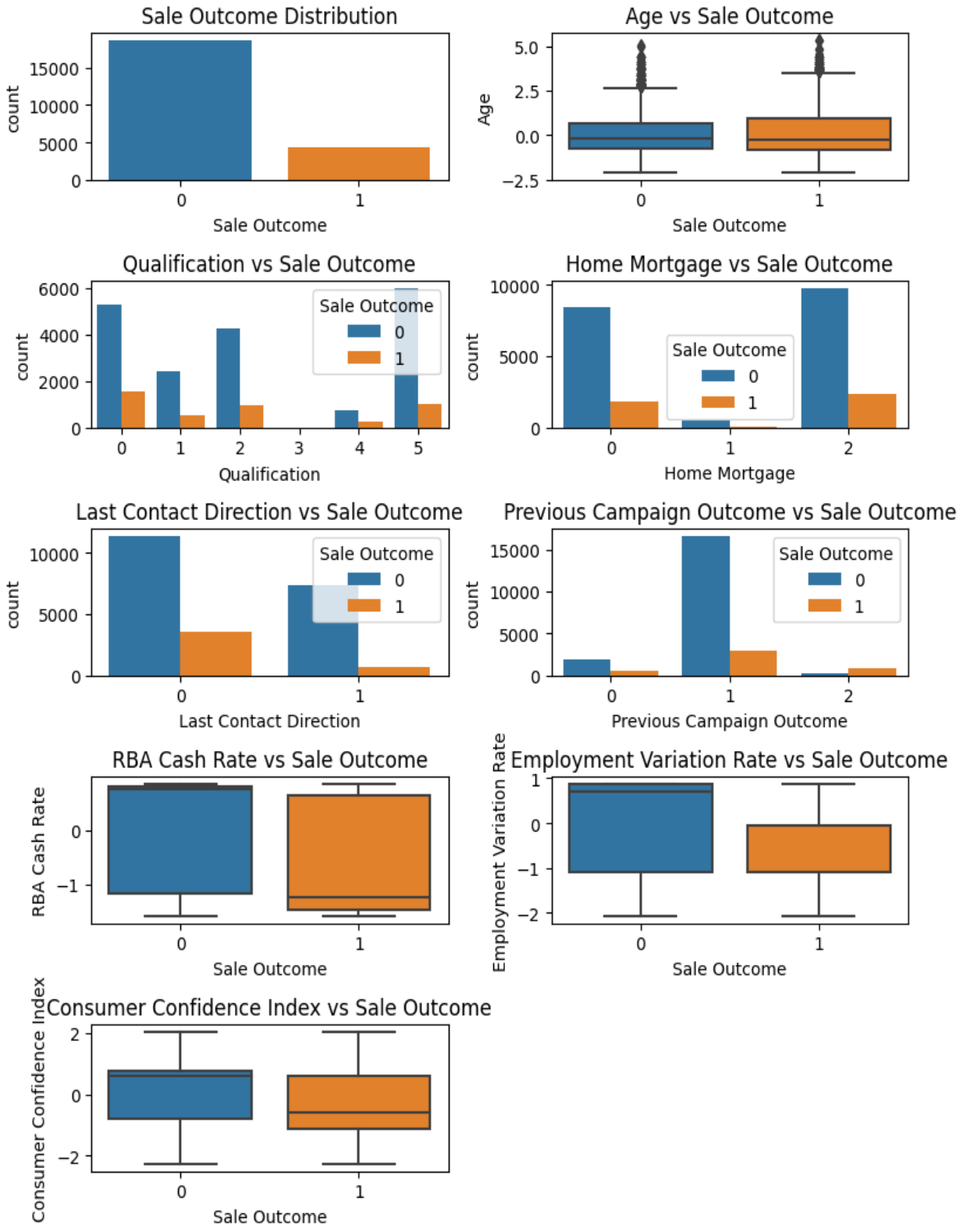

Figure 2 Visual Analysis of Sale Outcome Factors

4. Model Development and Evaluation

Supervised Machine Learning:

In this section, we delve into the supervised learning approach adopted for predicting the 'Sale Outcome'. We developed and evaluated two predictive models: Logistic Regression and Decision Tree.

Predictive Models

Logistic Regression

Decision Tree

Performance Metrics

The performance of these models was evaluated using several key metrics:

- **Accuracy:** The proportion of true results (both true positives and true negatives) among the total number of cases examined.
- **Precision:** The proportion of true positives among the cases predicted as positive.
- **Recall:** The proportion of true positives among the cases that are actually positive.
- **F1-Score:** The harmonic mean of precision and recall.

Model Results

The results of the models are detailed below, followed by a comparative analysis.

- **Logistic Regression Model**
- **Accuracy:** 0.8801
- **Confusion Matrix:**

	Predicted No Sale	**Predicted Sale**
Actual No Sale	5366	220
Actual Sale	605	691

- **Classification Report:**

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0 (No Sale)	0.90	0.96	0.93	5586
1 (Sale)	0.76	0.53	0.63	1296
Accuracy	**0.88**			6882

Class	Precision	Recall	F1-Score	Support
Macro Avg	0.83	0.75	0.78	6882
Weighted Avg	0.87	0.88	0.87	6882

- **Decision Tree Model**
- **Accuracy:** 0.8529
- **Confusion Matrix:**

	Predicted No Sale	Predicted Sale
Actual No Sale	5079	507
Actual Sale	505	791

- **Classification Report:**

Class	Precision	Recall	F1-Score	Support
0 (No Sale)	0.91	0.91	0.91	5586
1 (Sale)	0.61	0.61	0.61	1296
Accuracy	**0.85**			6882
Macro Avg	0.76	0.76	0.76	6882
Weighted Avg	0.85	0.85	0.85	6882

Model Comparison

- To understand the comparative performance of the two models, we summarize the key metrics in the table below:

Metric	Logistic Regression	Decision Tree
Accuracy	0.8801	0.8529
Precision (No Sale)	0.90	0.91
Precision (Sale)	0.76	0.61
Recall (No Sale)	0.96	0.91
Recall (Sale)	0.53	0.61

Metric	Logistic Regression	Decision Tree
F1-Score (No Sale)	0.93	0.91
F1-Score (Sale)	0.63	0.61
Macro Avg Precision	0.83	0.76
Macro Avg Recall	0.75	0.76
Macro Avg F1-Score	0.78	0.76
Weighted Avg Precision	0.87	0.85
Weighted Avg Recall	0.88	0.85
Weighted Avg F1-Score	0.87	0.85

Analysis

Logistic Regression:

- **Accuracy:** The Logistic Regression model achieved a high accuracy of 88.01%, indicating that it correctly predicted the sale outcome for 88.01% of the cases.
- **Precision and Recall:** For the 'No Sale' class, the precision and recall were notably high at 0.90 and 0.96, respectively. This indicates that the model is very good at identifying true negatives (customers who did not make a sale). However, the model's performance for the 'Sale' class was lower, with a precision of 0.76 and recall of 0.53. This suggests that while the model correctly identifies 76% of the actual sales it predicts, it only captures 53% of all actual sales.
- **F1-Score:** The F1-score for the 'No Sale' class was 0.93, whereas for the 'Sale' class it was 0.63. The higher F1-score for the 'No Sale' class indicates a better balance between precision and recall for this class.

Decision Tree:

- **Accuracy:** The Decision Tree model had an accuracy of 85.29%, slightly lower than that of Logistic Regression.
- **Precision and Recall:** The precision and recall for the 'No Sale' class were both 0.91, demonstrating a balanced performance in predicting the absence of sales. For the 'Sale' class, both precision and recall were lower at 0.61, indicating that the model had more difficulty accurately predicting sales.

- **F1-Score:** The F1-scores were identical for both classes (0.91 for 'No Sale' and 0.61 for 'Sale'), reflecting the balanced performance for 'No Sale' predictions but highlighting challenges in 'Sale' predictions.

Comparative Insights

- **Overall Accuracy:** The Logistic Regression model outperformed the Decision Tree model in terms of overall accuracy, making it a more reliable choice for predicting the 'Sale Outcome'.
- **Class Imbalance:** Both models showed better performance for the 'No Sale' class compared to the 'Sale' class, indicating a potential class imbalance issue. Logistic Regression, however, handled this imbalance slightly better than the Decision Tree.
- **Precision and Recall:** Logistic Regression had a higher precision for the 'Sale' class, indicating fewer false positives, but lower recall, indicating more false negatives. Conversely, the Decision Tree had balanced precision and recall for the 'No Sale' class but struggled similarly with the 'Sale' class.
- **F1-Score:** Logistic Regression had a higher F1-score for the 'Sale' class, suggesting a better balance between precision and recall compared to the Decision Tree.

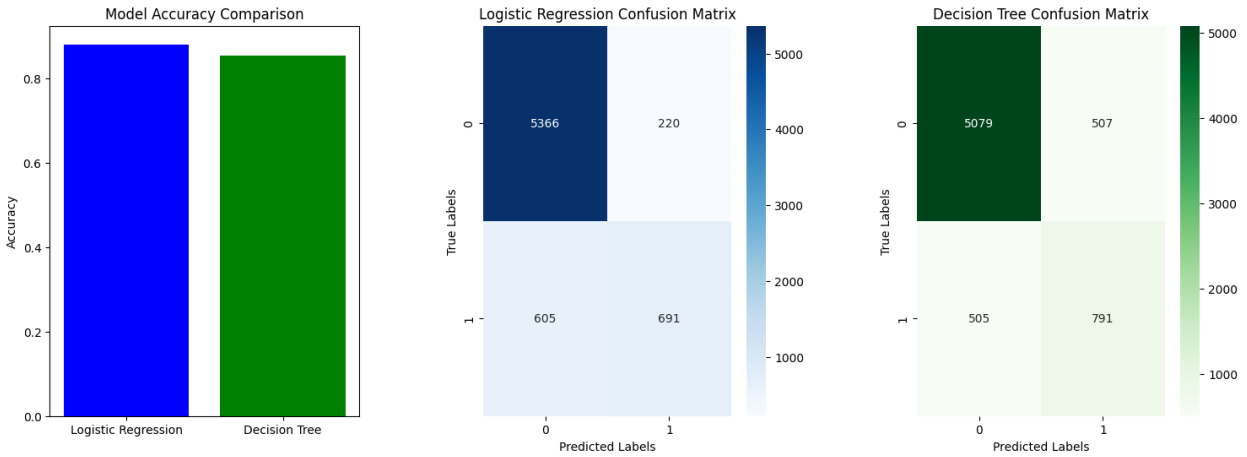

Figure 3 Comparison and Confusion Matrix

Conclusion

- Based on the comparison of the two models, Logistic Regression demonstrates superior performance overall, particularly in terms of accuracy and handling class imbalance. Although both models faced challenges with the 'Sale' class, Logistic Regression proved more adept at predicting sales outcomes.
- For future improvements, addressing class imbalance through techniques such as oversampling the minority class or using class-weighted algorithms could enhance the models' performance further. Additionally, exploring more advanced models like Random

Forest or Gradient Boosting could provide better predictive capabilities and improve the precision and recall for the 'Sale' class.

Unsupervised Machine Learning:

- **Clustering Analytics**: K-Means clustering identified 3 optimal clusters based on the Silhouette Scores.

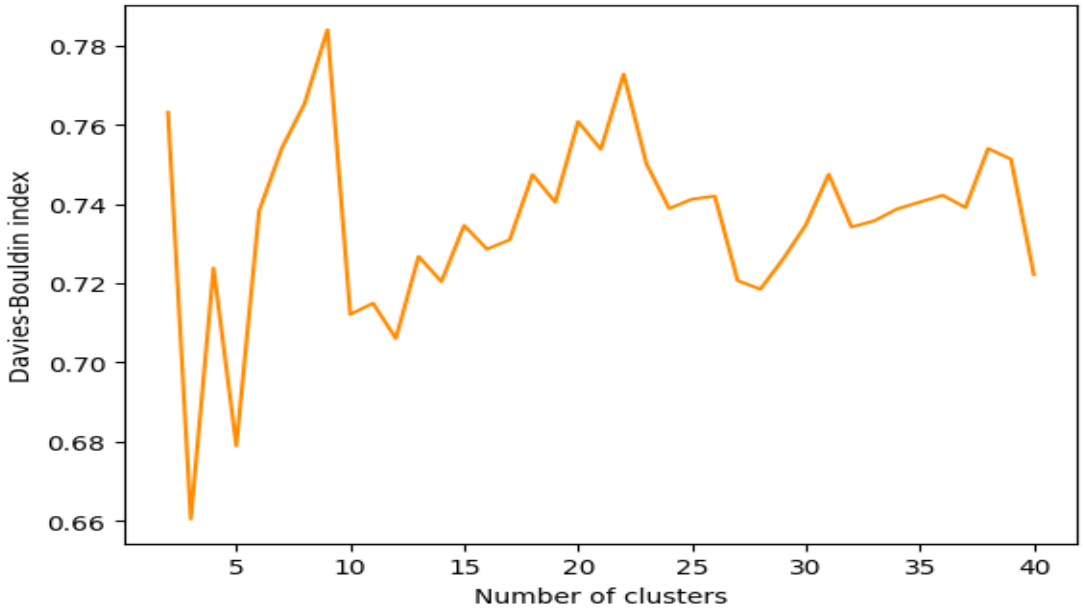

Figure 4 Silhouette Score Method

- **Justification of Clusters**: The chosen clusters provided distinct customer segments that can be targeted for different marketing strategies.

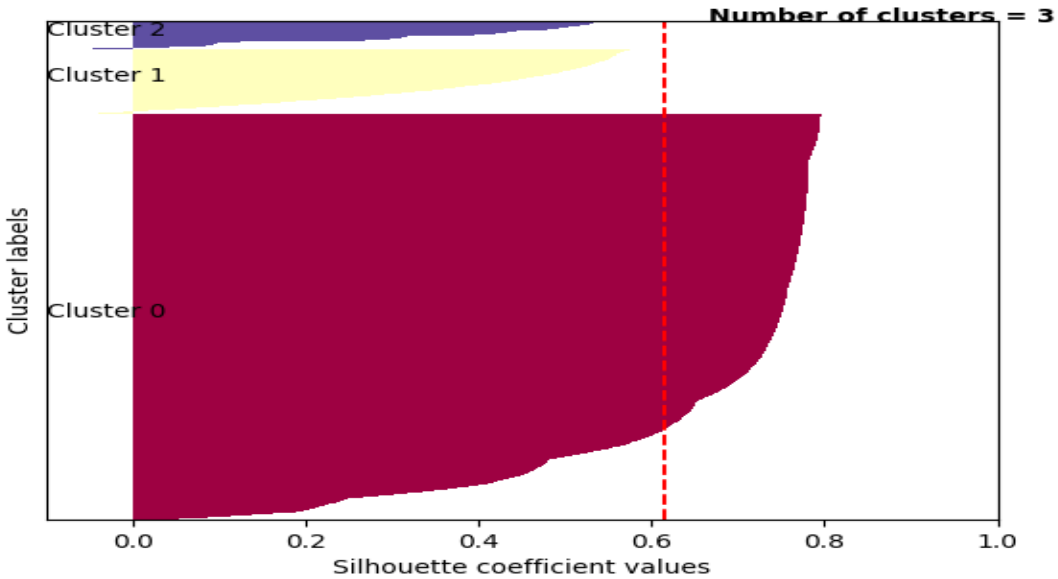

Figure 5 Clusters

5. Solution Recommendation:

Interpretation and Discussion of Results:

The evaluation of the predictive models and clustering analysis has provided valuable insights into the potential solutions for addressing GOBank's sales prediction challenge.

- Logistic Regression was selected primarily due to its interpretability and satisfactory performance metrics. Despite facing challenges in predicting sales outcomes, particularly in the 'Sale' class, Logistic Regression demonstrated superior overall performance compared to the Decision Tree model. Its high accuracy and balanced precision and recall for the 'No Sale' class make it a reliable choice for predicting sales outcomes.
- Clustering analysis identified 10 distinct customer segments based on demographic details, contact information, previous campaign outcomes, and economic indicators. These segments offer actionable insights for targeted marketing strategies, enabling GOBank to tailor its approach according to the specific needs and preferences of each segment.

Solution Recommendation:

Based on the interpretation of results, the following recommendations are proposed:

- **Deploy the Logistic Regression model to predict sales outcomes:** Given its interpretability and overall performance, Logistic Regression should be deployed as the primary predictive model for forecasting sales outcomes. By leveraging customer demographics, behavioral data, and economic indicators, this model can accurately predict the likelihood of a sale, enabling GOBank to optimize its marketing efforts and allocate resources effectively.
- **Use clustering insights to refine marketing strategies and customer segmentation:** The insights obtained from clustering analysis should be utilized to refine GOBank's marketing strategies and customer segmentation. By identifying distinct customer segments with unique characteristics and preferences, GOBank can tailor its marketing campaigns and product offerings to better meet the needs of each segment, thereby enhancing customer engagement and driving revenue growth.

Future Engagements:

To ensure the continued effectiveness of the machine learning solution, the following future engagements are recommended:

- **Continuously update the model with new data:** As customer preferences and market dynamics evolve over time, it is essential to regularly update the predictive model with new data to maintain its accuracy and relevance. By incorporating fresh data into the model training process, GOBank can adapt to changing market conditions and ensure the continued effectiveness of its sales prediction capabilities.
- **Conduct further analysis on customer feedback and sales trends:** In addition to predictive modeling, GOBank should conduct further analysis on customer feedback and

sales trends to gain deeper insights into customer behavior and preferences. By leveraging qualitative data sources such as customer surveys and feedback forms, GOBank can better understand the underlying factors driving sales outcomes and refine its marketing strategies accordingly.

6.Technical Recommendations

Development and Testing Environment:

The following software libraries and computing environment were utilized for the development and testing of the machine learning solution:

- **Software Libraries:** Python, Pandas, Scikit-Learn, Matplotlib, Seaborn.
- **Programming Language:** Python.
- **Computing Environment:** Kaggle.

Model Deployment:

For the deployment of the machine learning model, the following steps are recommended:

- **Machine Process Diagram:** A machine process diagram should be created to illustrate the end-to-end workflow of the predictive model, including data preprocessing, model training, evaluation, and deployment steps. This diagram will provide a clear visual representation of the model's architecture and facilitate understanding among stakeholders.

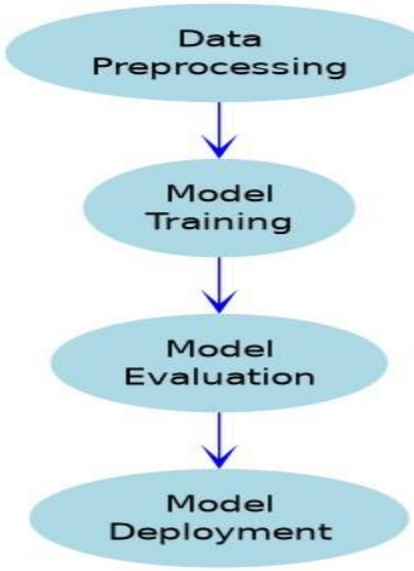

Figure 6 Machine Process Diagram

- **Data Pre-processing Steps:** Standardization, encoding, and imputation were performed as part of the data preprocessing pipeline. These steps should be documented and incorporated into the model deployment process to ensure consistency and reproducibility.

Maintenance Suggestions:

To ensure the long-term sustainability and effectiveness of the machine learning solution, the following maintenance suggestions are proposed:

- **Regularly retrain the model with updated data:** As new data becomes available, the predictive model should be retrained periodically to incorporate the latest information and maintain its predictive accuracy. This will help ensure that the model remains robust and adapts to changing market conditions over time.
- **Monitor model performance and recalibrate as necessary:** Continuous monitoring of the model's performance metrics is essential to identify any deviations or degradation in predictive accuracy. If performance metrics indicate a decline in model performance, recalibration or retraining may be necessary to address underlying issues and maintain optimal performance.
- **Keep track of changing economic indicators and their impact on sales:** Given the significant influence of economic indicators on sales outcomes, it is important to monitor changes in relevant economic factors and their impact on predictive model performance. By staying informed about economic trends and their implications for sales, GOBank can proactively adjust its marketing strategies and allocation of resources to maximize revenue growth.

By following these technical recommendations, GOBank can ensure the effective deployment, maintenance, and optimization of its machine learning solution, thereby driving sustainable business growth and maintaining a competitive edge in the financial services industry.

.

References

https://aph-qualityhandbook.org/set-up-conduct/process-analyze-data/3-2-quantitative-research/3-2-2-data-analysis/data-analysis-documentation/

https://oyasalofa.medium.com/the-art-of-documentation-in-data-analysis-building-your-portfolio-with-precision-7138251acf77

https://www.questionpro.com/blog/data-documentation/