Berkeley Engineering | BerkeleyHaas

UC-Berkeley-ML-AI - Practical Assignment 3
Non-Technical Documentation

Version 1.0.0

Duy Nguyen

July 26, 2024

UC-Berkeley-ML-AI - Practical Assignment 3 Non-Technical Documentation

Client Name	CORE
Project Name / Engagement ID	Predictive Modeling for Customer Subscription Optimization
Version	1.0

Revision History

Date	Version	Description	Author	Authorized By	Approved By
05/06/2024	1.0	Draft version	Duy Nguyen	Vikesh Koul	Vikesh Koul
25/07/2024	1.1	Edit option	Duy Nguyen	Vikesh Koul	Vikesh Koul

Feedback and Acknowledgments

You scored 100%.

Reviewer	Vikesh Koul
Feedback	"Good work Duy!"

Confidentiality Notice: This document contains confidential and proprietary information of Core. Unauthorized use, disclosure, or distribution of this document or its contents is strictly prohibited. If you are not the intended recipient, please notify Core immediately and delete all copies of this document.

Contents

1 Overview 3

2 Data Description 3

2.1 Numerical Features . . . 3

2.2 Categorical Features . . . 3

3 Exploratory Data Analysis (EDA) 4

3.1 Key Visualizations . . . 4

4 Model Comparison and Observing Overfitting 10

4.1 Visualizations . . . 11

4.2 Observations . . . 12

5 Classification Performance and Confusion Matrix Analysis 13

5.1 Logistic Regression . . . 13

5.2 Decision Tree . . . 15

5.3 K-Nearest Neighbors (KNN) . . . 18

5.4 Support Vector Machine (SVM) . . . 21

6 Feature Importance Analysis 24

6.1 SVM Feature Importance . . . 24

6.2 Logistic Regression Feature Importance . . . 25

6.3 Partial Dependence Plots (PDPs) . . . 26

7 Business Implications 27

8 Comparison with the Paper Findings 27

9 Recommendations Moving Forward 28

9.1 Enhanced Feature Engineering . . . 28

9.2 Model Improvement . . . 28

9.3 Business Strategy Optimization . . . 28

10 Conclusion 29

List of Listings

Confidentiality Notice: This document contains confidential and proprietary information of Core. Unauthorized use, disclosure, or distribution of this document or its contents is strictly prohibited. If you are not the intended recipient, please notify Core immediately and delete all copies of this document.

1 Overview

This project aims to build and evaluate predictive models for a marketing campaign to optimize customer subscriptions for long-term deposit products. It involves data preprocessing, exploratory data analysis (EDA), model training, and evaluation. For detail code base:

Project Repository

For more details, you can visit the project repository on GitHub: UC-Berkeley-ML-AI Practical Assignment 3

2 Data Description

The dataset includes both numerical and categorical features relevant to customer behavior.

2.1 Numerical Features

- **age:** Age of the customer
- **duration:** Last contact duration in seconds
- **campaign:** Number of contacts performed during this campaign
- **pdays:** Number of days since the client was last contacted
- **previous:** Number of contacts performed before this campaign
- **emp.var.rate:** Employment variation rate
- **cons.price.idx:** Consumer price index
- **cons.conf.idx:** Consumer confidence index
- **euribor3m:** Euribor 3-month rate
- **nr.employed:** Number of employees

2.2 Categorical Features

- **job:** Type of job
- **marital:** Marital status
- **education:** Education level
- **default:** Has credit in default?
- **housing:** Has housing loan?

- **loan:** Has personal loan?
- **contact:** Type of communication contact
- **month:** Last contact month
- **day_of_week:** Last contact day of the week
- **poutcome:** Outcome of the previous marketing campaign

3 Exploratory Data Analysis (EDA)

3.1 Key Visualizations

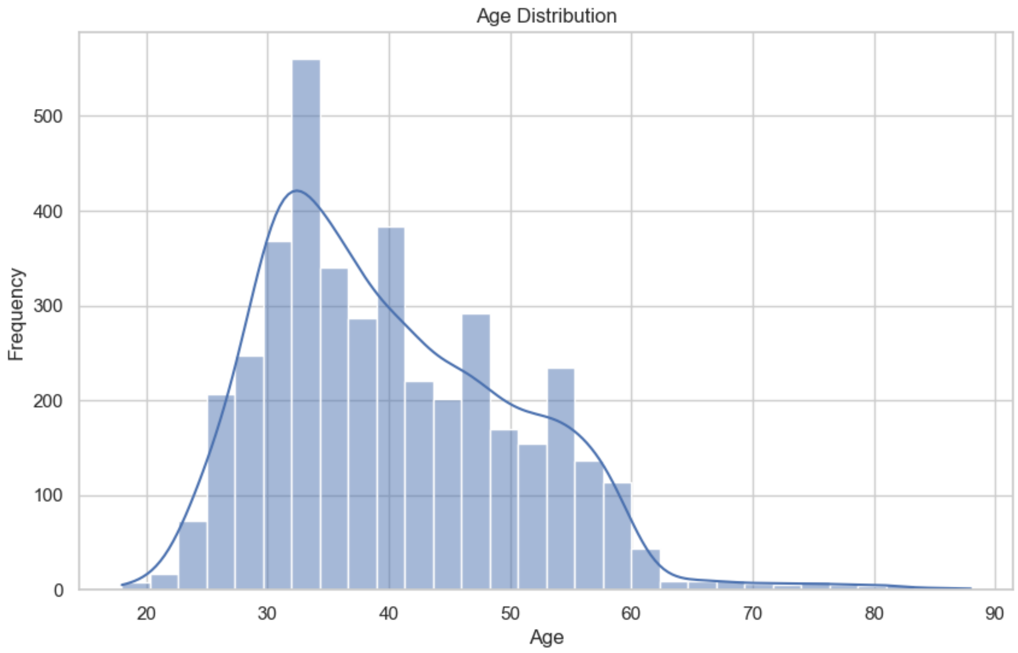


Figure 1: Age Distribution

- Most customers are around 30 years old.
- The age distribution shows a right skew, indicating fewer older customers.

Confidentiality Notice: This document contains confidential and proprietary information of Core. Unauthorized use, disclosure, or distribution of this document or its contents is strictly prohibited. If you are not the intended recipient, please notify Core immediately and delete all copies of this document.

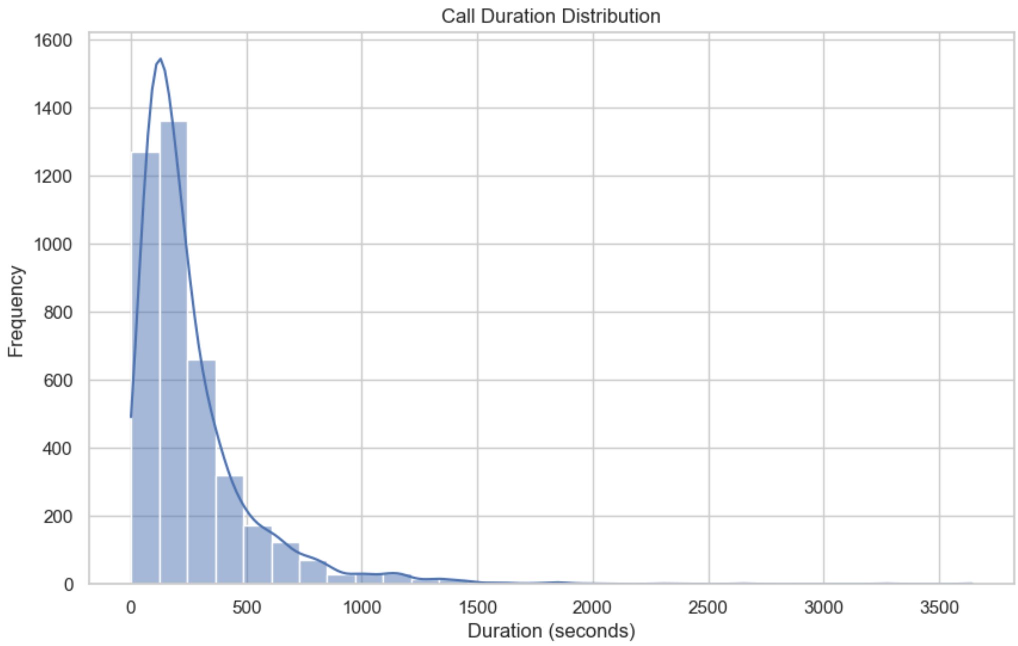
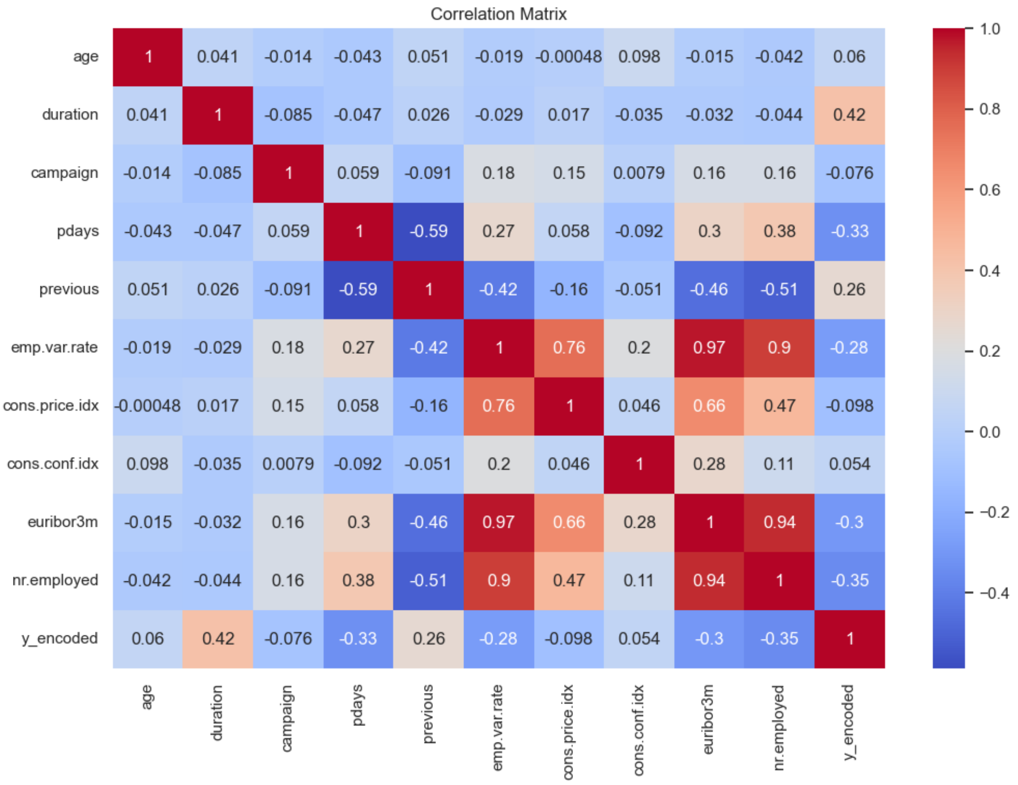


Figure 2: Call Duration Distribution

- Majority of calls are short, typically under 500 seconds.
- The distribution is heavily right-skewed, with a long tail of longer calls.



	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y_encoded
age	1	0.041	-0.014	-0.043	0.051	-0.019	-0.00048	0.098	-0.015	-0.042	0.06
duration	0.041	1	-0.085	-0.047	0.026	-0.029	0.017	-0.035	-0.032	-0.044	0.42
campaign	-0.014	-0.085	1	0.059	-0.091	0.18	0.15	0.0079	0.16	0.16	-0.076
pdays	-0.043	-0.047	0.059	1	-0.59	0.27	0.058	-0.092	0.3	0.38	-0.33
previous	0.051	0.026	-0.091	-0.59	1	-0.42	-0.16	-0.051	-0.46	-0.51	0.26
emp.var.rate	-0.019	-0.029	0.18	0.27	-0.42	1	0.76	0.2	0.97	0.9	-0.28
cons.price.idx	-0.00048	0.017	0.15	0.058	-0.16	0.76	1	0.046	0.66	0.47	-0.098
cons.conf.idx	0.098	-0.035	0.0079	-0.092	-0.051	0.2	0.046	1	0.28	0.11	0.054
euribor3m	-0.015	-0.032	0.16	0.3	-0.46	0.97	0.66	0.28	1	0.94	-0.3
nr.employed	-0.042	-0.044	0.16	0.38	-0.51	0.9	0.47	0.11	0.94	1	-0.35
y_encoded	0.06	0.42	-0.076	-0.33	0.26	-0.28	-0.098	0.054	-0.3	-0.35	1

Figure 3: Correlation Matrix

- Highlights positive and negative relationships between numerical features.
- Duration and y_encoded have a moderate positive correlation, suggesting longer calls are associated with higher subscription rates. (Note that y_encoded is the subscription rate)

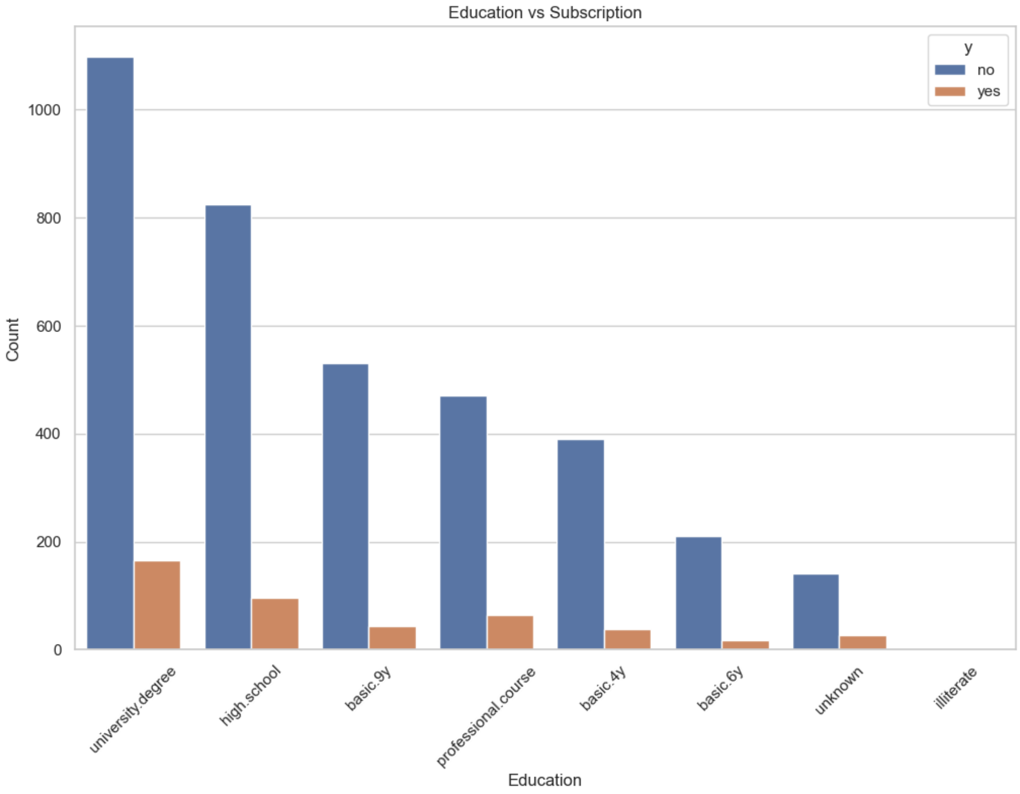


Figure 4: Education vs Subscription

- Higher subscription rates are observed among customers with university degrees.
- Customers with lower educational levels tend to have lower subscription rates.

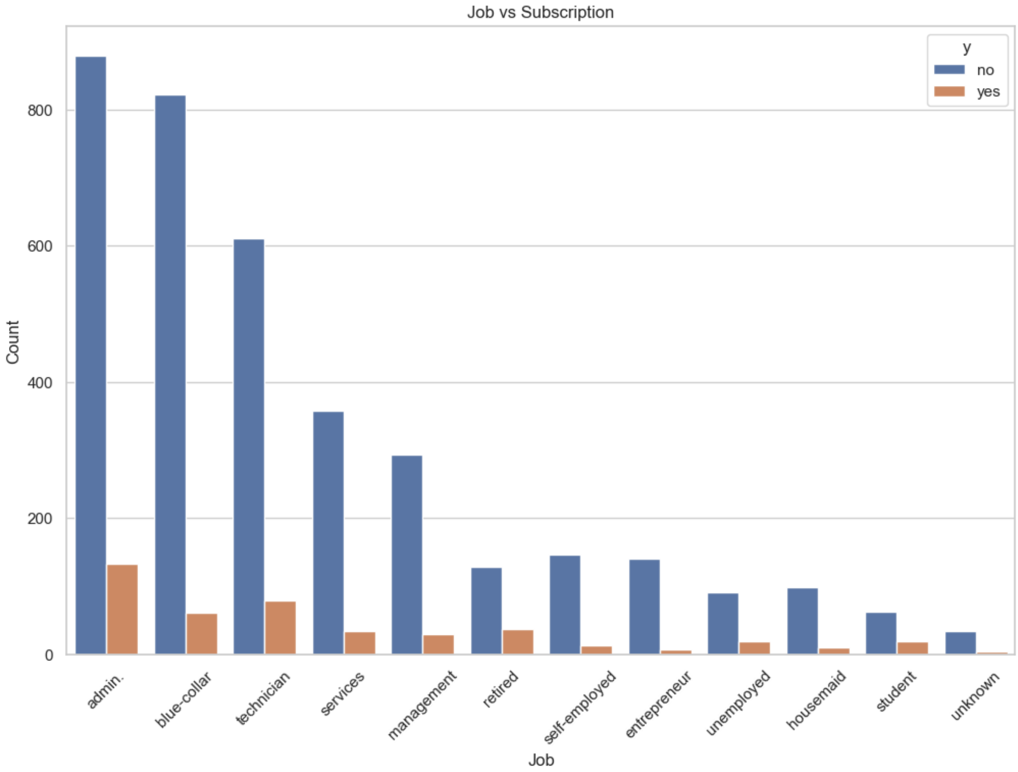


Figure 5: Job vs Subscription

- Job types like ’admin.’ and ’blue-collar’ have lower subscription rates.
- Higher subscription rates are observed for job types like ’management’ and ’retired’.

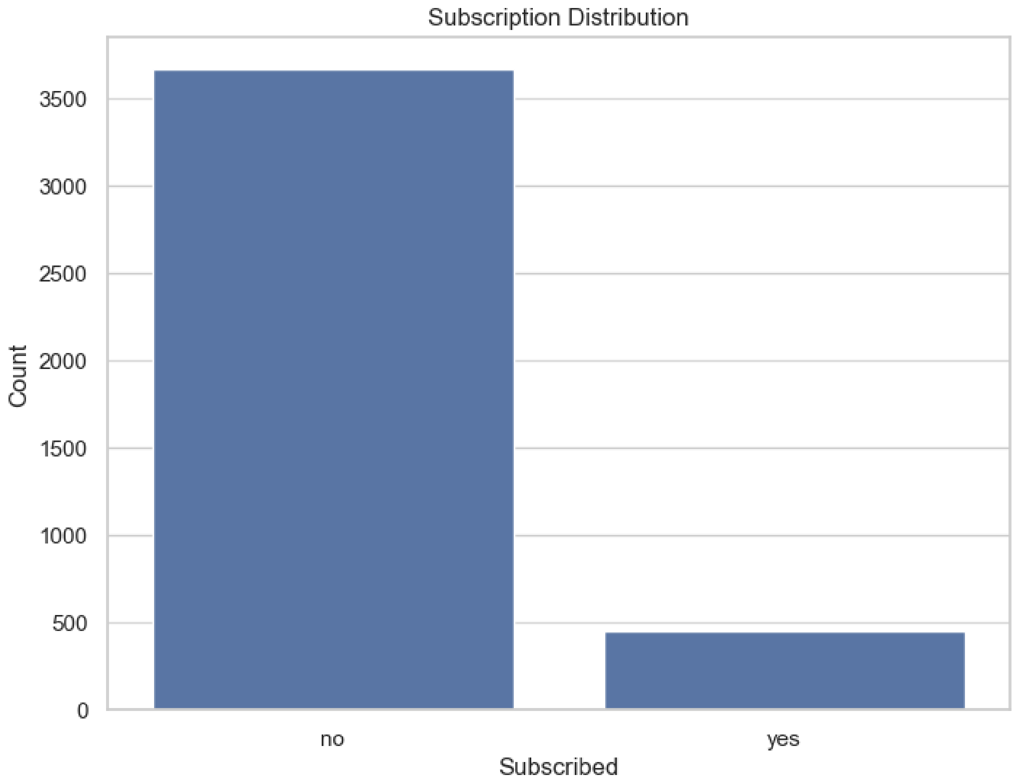


Figure 6: Subscription Distribution

- Majority of the customers did not subscribe.
- There is a significant class imbalance with far fewer subscribers.

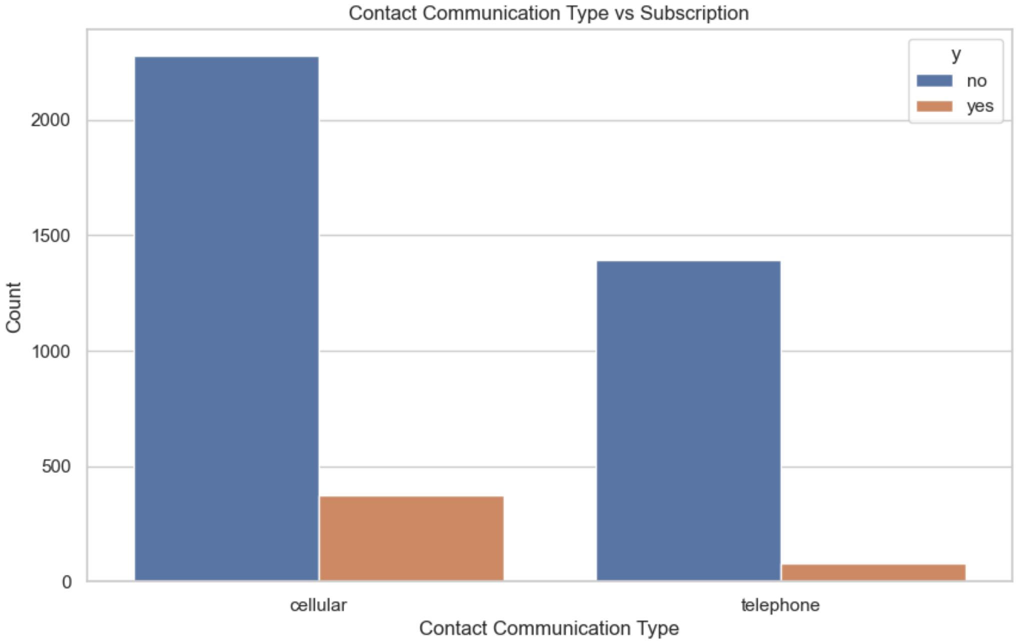


Figure 7: Contact Communication Type vs Subscription

- Higher subscription rates are seen for cellular contacts compared to telephone.
- The majority of contacts were made via cellular phones.

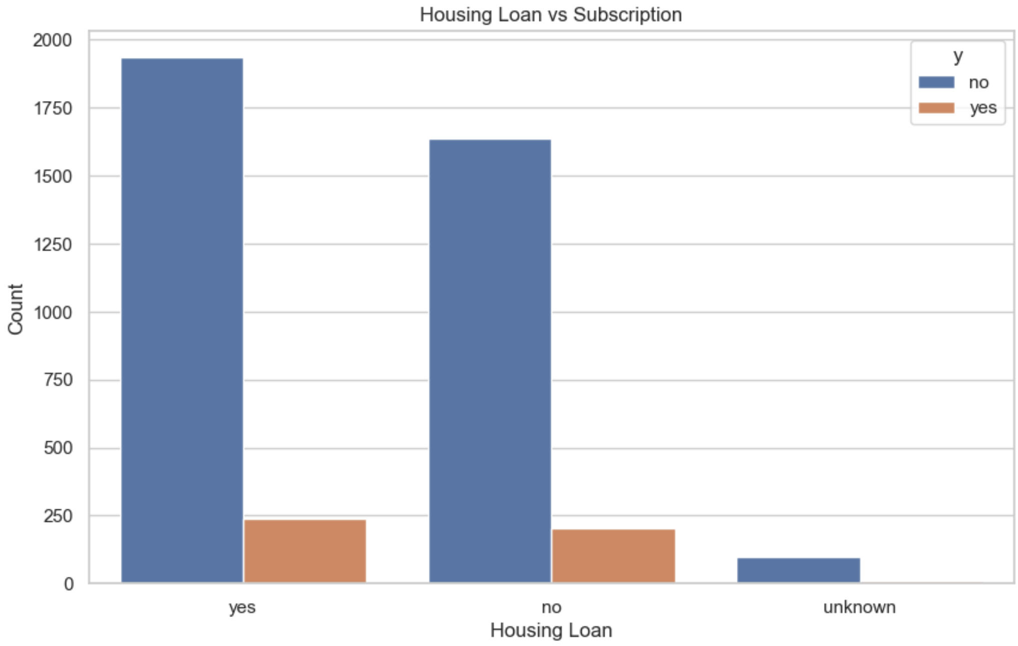


Figure 8: Housing Loan vs Subscription

- Customers without housing loans have higher subscription rates.
- Most customers, irrespective of subscription status, do have housing loans.

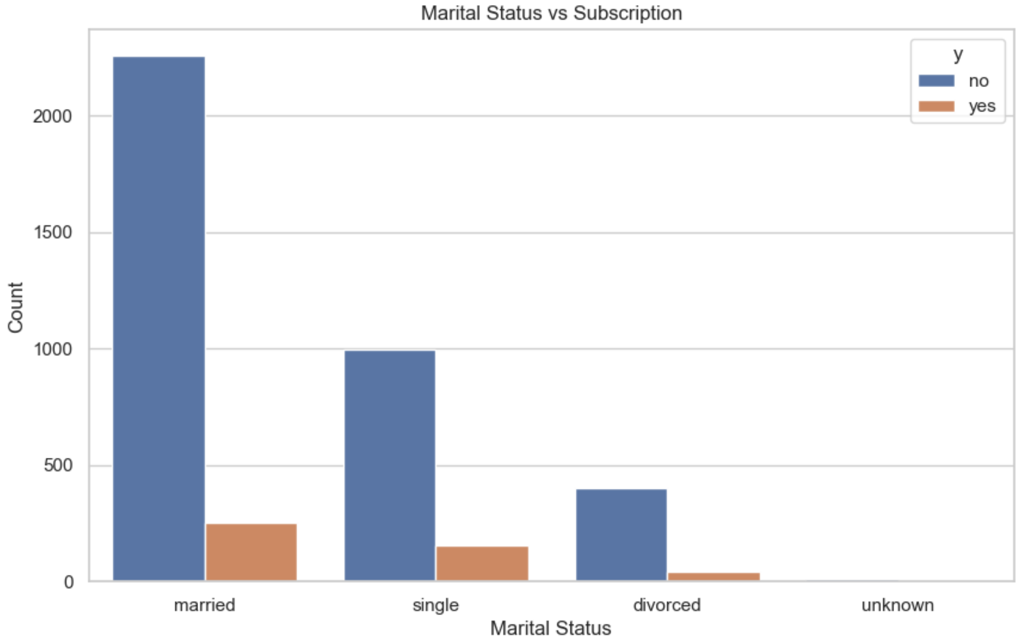


Figure 9: Marital Status vs Subscription

- Single customers have higher subscription rates compared to married or divorced ones.
- Married customers form the largest group but have lower subscription rates.

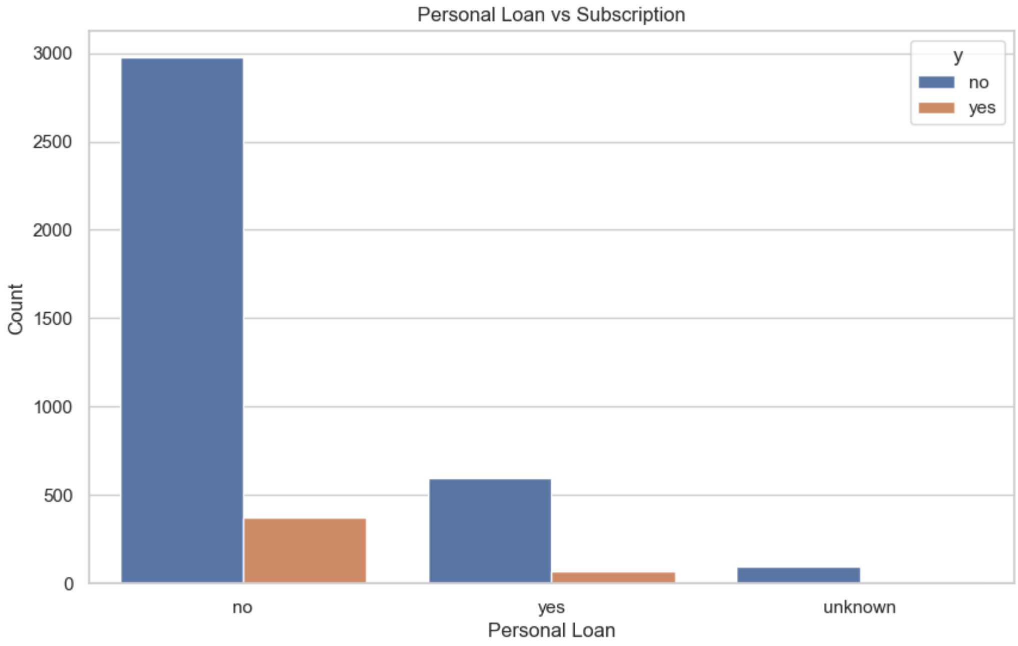


Figure 10: Personal Loan vs Subscription

- Higher subscription rates are observed for customers without personal loans.
- Most customers, especially those who did not subscribe, do not have personal loans.

4 Model Comparison and Observing Overfitting

Four models were evaluated: Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM).

Model	Train Time (seconds)	Train Accuracy	Test Accuracy	CV Mean Accura
Logistic Regression	81.01	0.878	0.877	0.874
Decision Tree	2.32	0.991	0.960	0.944
K-Nearest Neighbors	4.75	1.000	0.909	0.913
Support Vector Machine	875.95	1.000	0.996	0.992

Table 1: Model Comparison

4.1 Visualizations

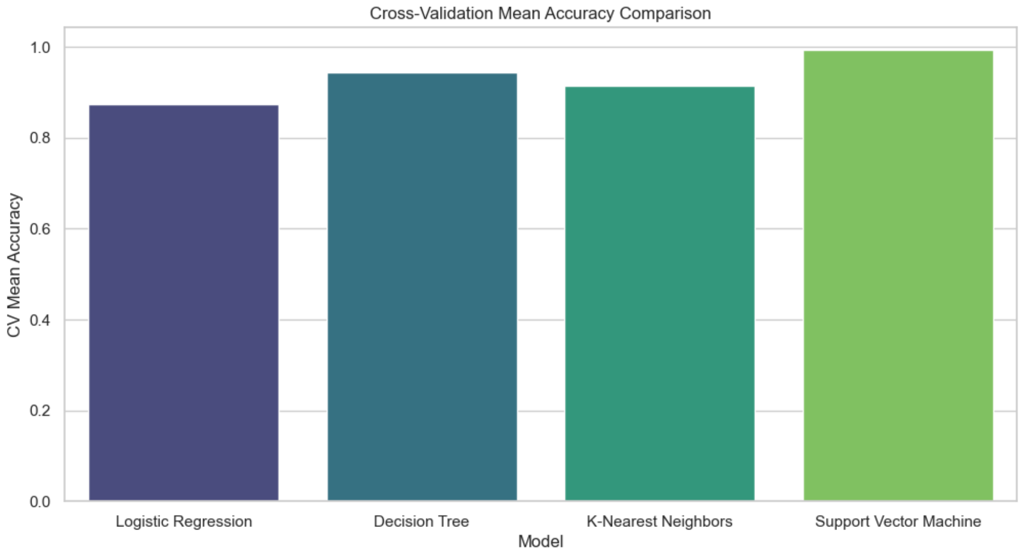


Figure 11: Cross-Validation Mean Accuracy Comparison

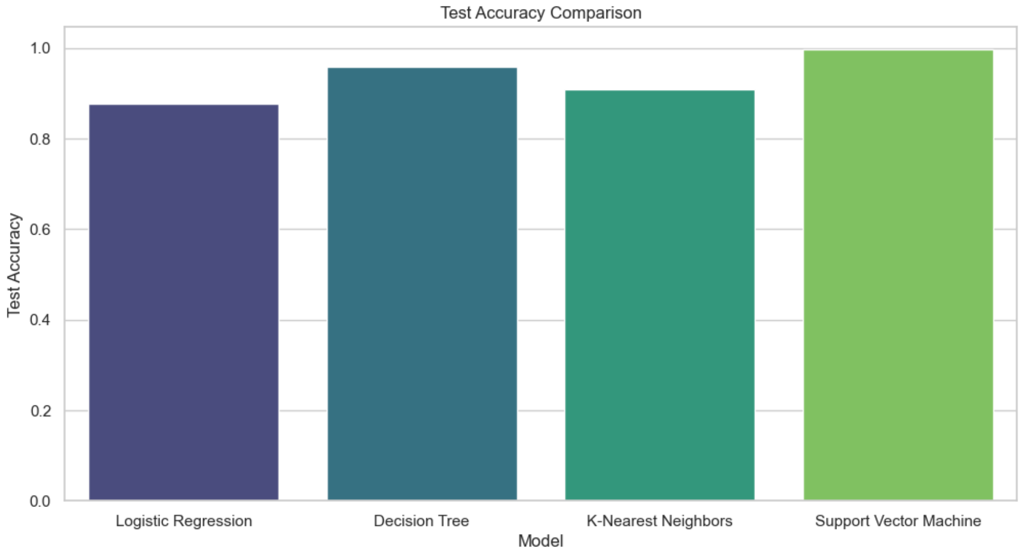


Figure 12: Test Accuracy Comparison

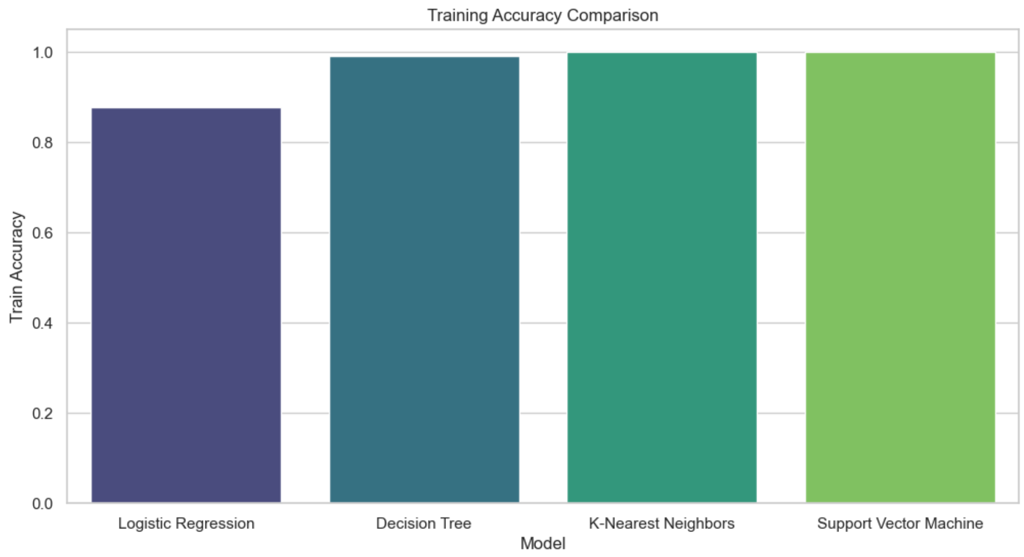


Figure 13: Training Accuracy Comparison

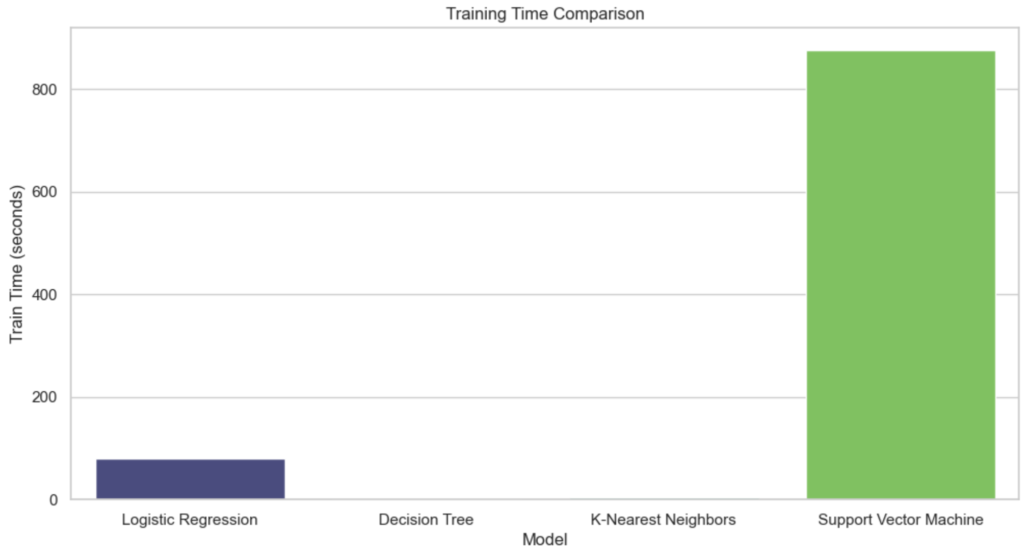


Figure 14: Training Time Comparison

4.2 Observations

- SVM shows the highest accuracy but requires more training time.
- Logistic Regression and Decision Tree models have strong performance with faster training times.
- KNN and Decision Tree models indicate potential overfitting.

The Support Vector Machine (SVM) is the best choice for optimizing marketing campaigns due to its superior predictive performance and ability to generalize well on new data, despite its higher computational cost.

5 Classification Performance and Confusion Matrix Analysis

5.1 Logistic Regression

- **Training Time:** 81.01 seconds
- **Confusion Matrix:**

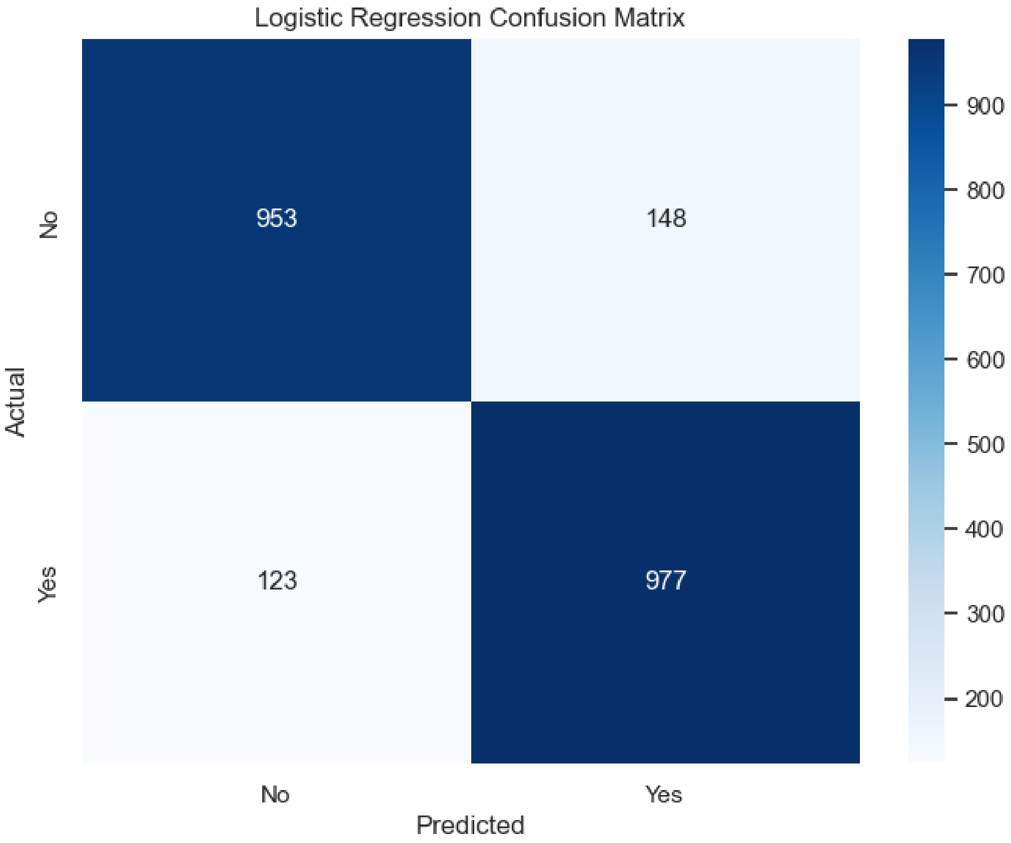


Figure 15: Logistic Regression Confusion Matrix

- **Classification Report:**

Metric	Precision (No)	Precision (Yes)	Recall (No)	Recall (Yes)	F1-score (No)	F1-score (Yes)
Values	0.89	0.87	0.87	0.89	0.88	0.88

Table 2: Logistic Regression Classification Report

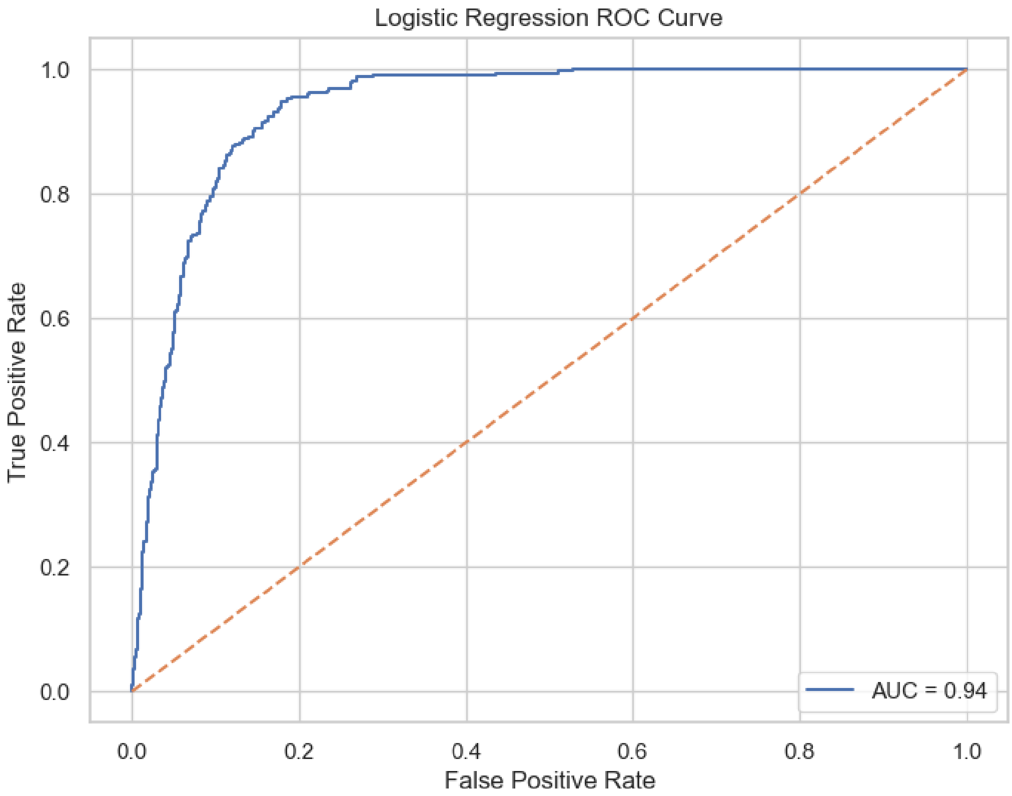


Figure 16: Logistic Regression ROC Curve

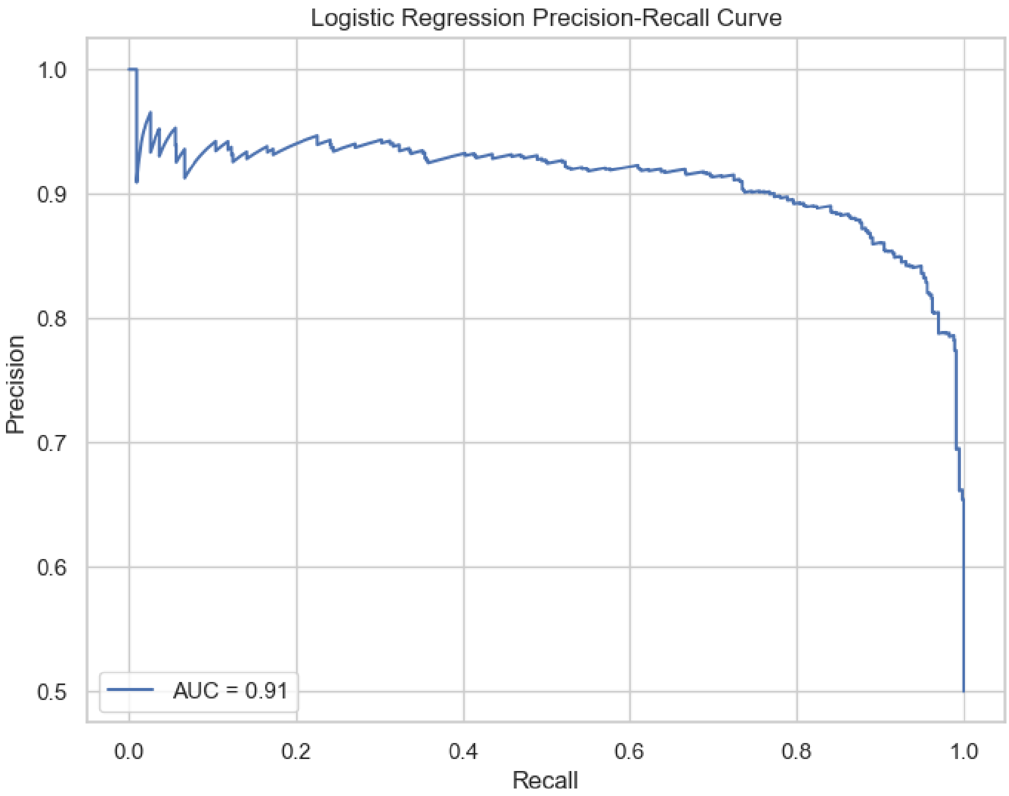


Figure 17: Logistic Regression Precision-Recall Curve

5.2 Decision Tree

- **Training Time:** 2.32 seconds
- **Confusion Matrix:**

Confidentiality Notice: This document contains confidential and proprietary information of Core. Unauthorized use, disclosure, or distribution of this document or its contents is strictly prohibited. If you are not the intended recipient, please notify Core immediately and delete all copies of this document.

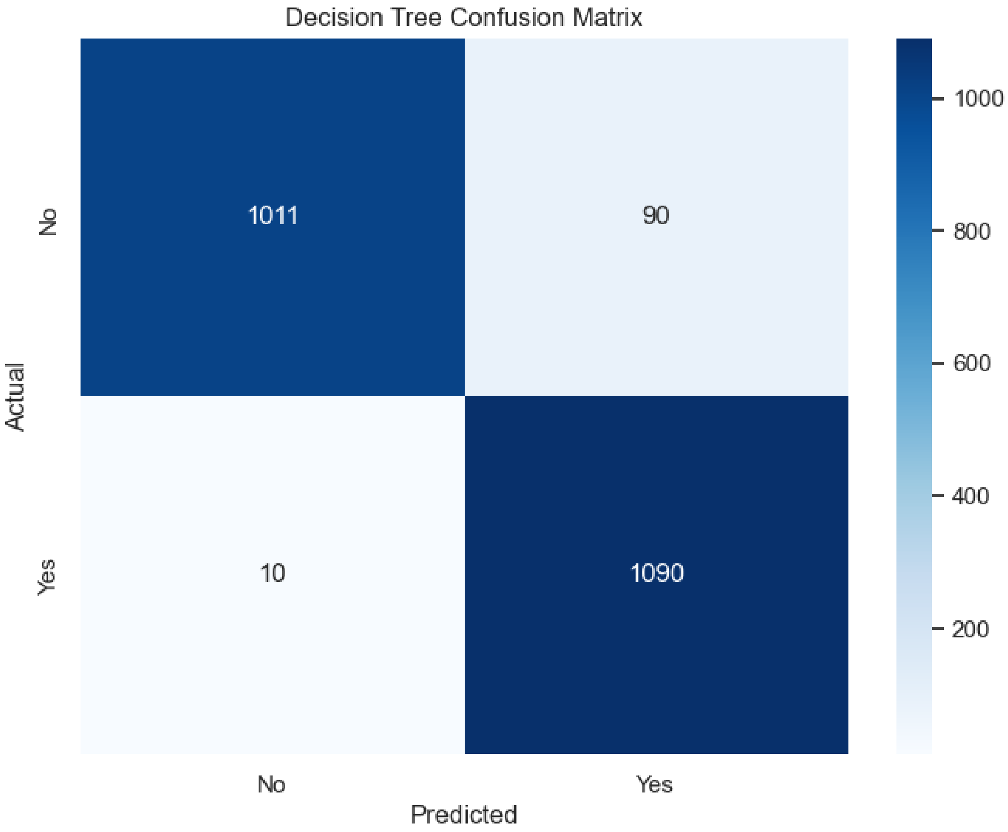


Figure 18: Decision Tree Confusion Matrix

- **Classification Report:**

Metric	Precision (No)	Precision (Yes)	Recall (No)	Recall (Yes)	F1-score (No)	F1-score (Yes)
Values	0.99	0.93	0.93	0.99	0.96	0.96

Table 3: Decision Tree Classification Report

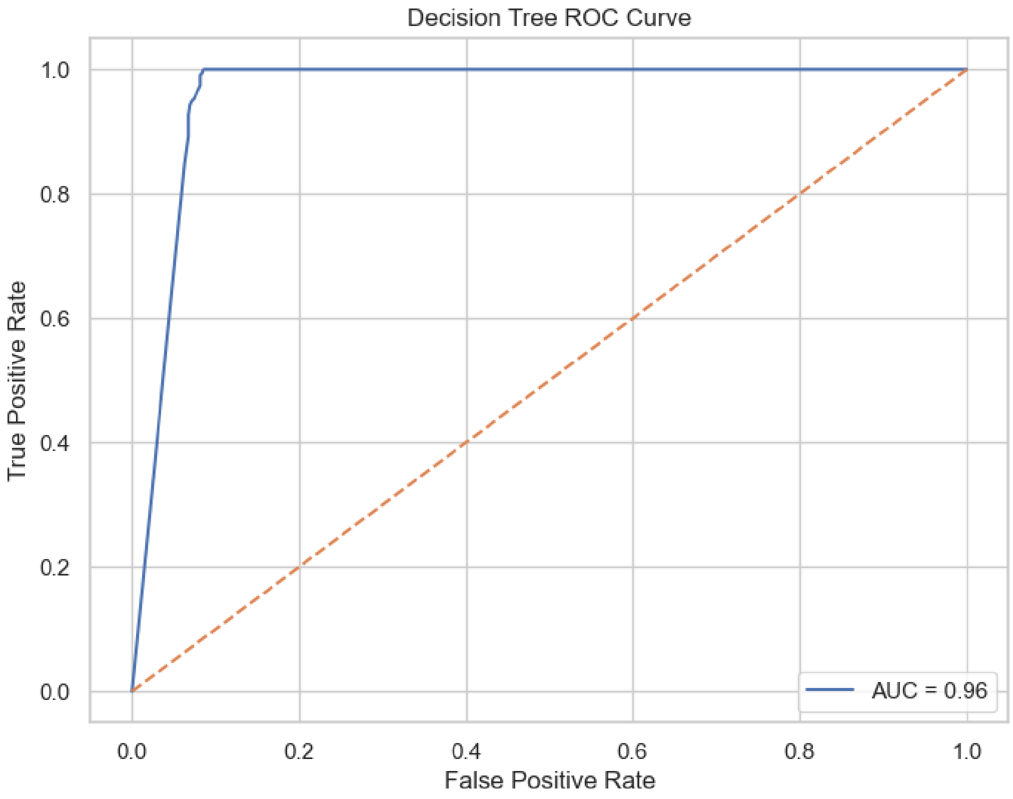


Figure 19: Decision Tree ROC Curve

Confidentiality Notice: This document contains confidential and proprietary information of Core. Unauthorized use, disclosure, or distribution of this document or its contents is strictly prohibited. If you are not the intended recipient, please notify Core immediately and delete all copies of this document.

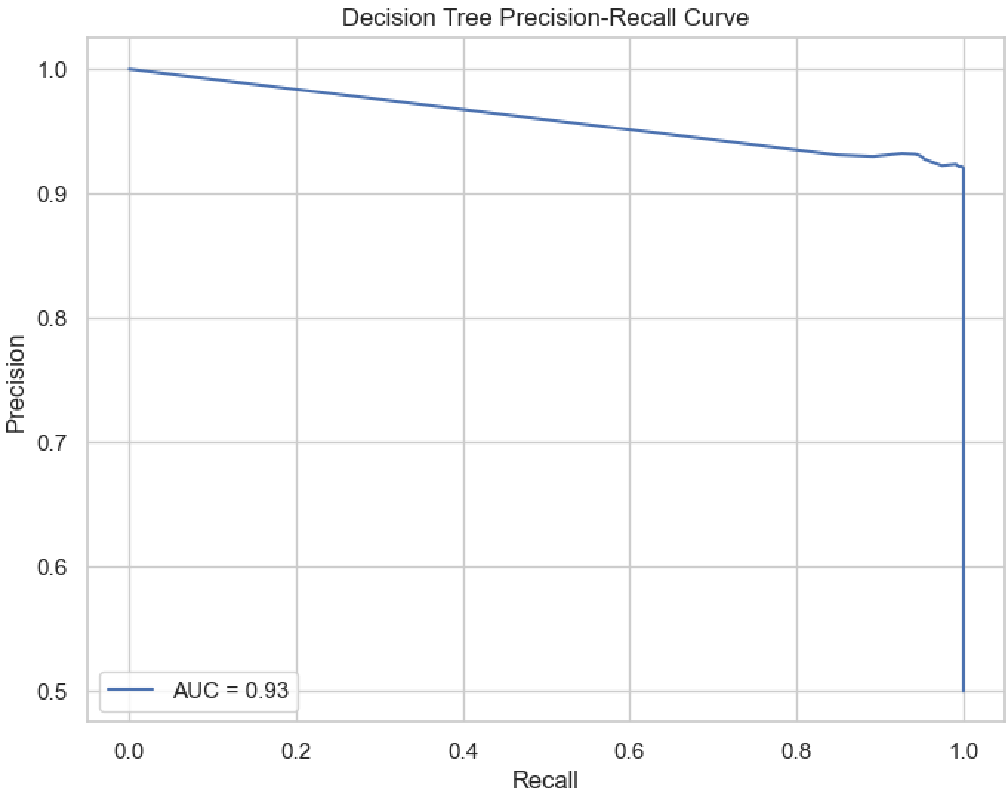


Figure 20: Decision Tree Precision-Recall Curve

5.3 K-Nearest Neighbors (KNN)

- **Training Time:** 4.75 seconds
- **Confusion Matrix:**

Confidentiality Notice: This document contains confidential and proprietary information of Core. Unauthorized use, disclosure, or distribution of this document or its contents is strictly prohibited. If you are not the intended recipient, please notify Core immediately and delete all copies of this document.

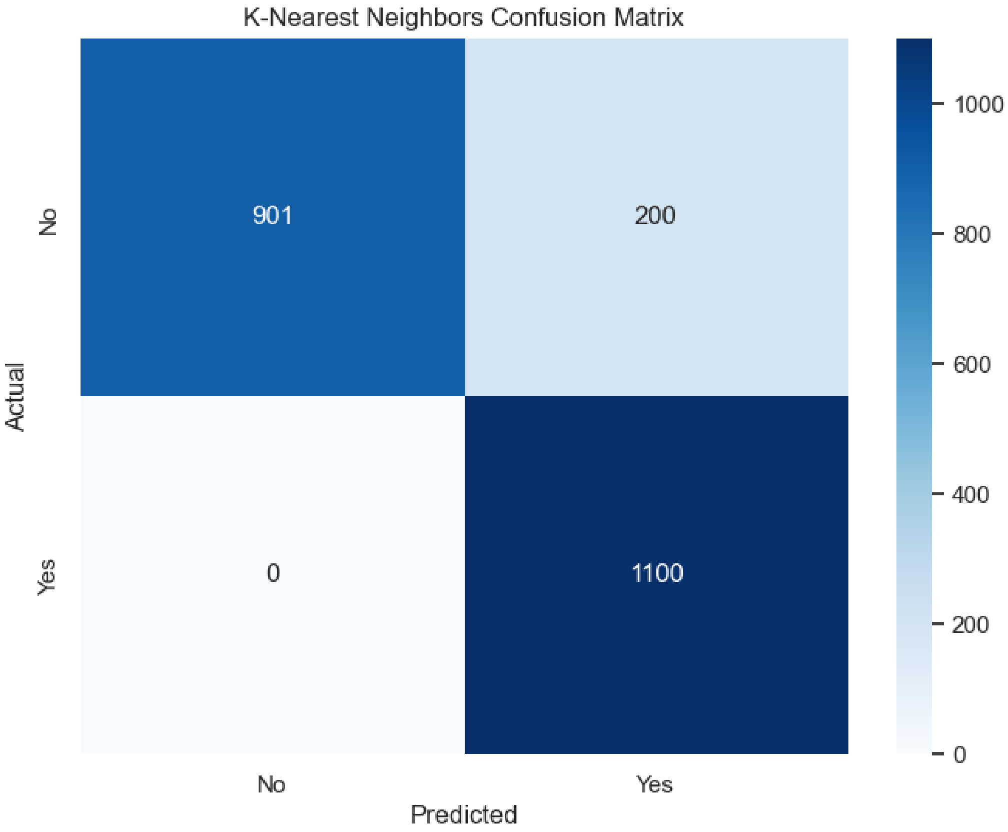


Figure 21: K-Nearest Neighbors Confusion Matrix

- **Classification Report:**

Metric	Precision (No)	Precision (Yes)	Recall (No)	Recall (Yes)	F1-score (No)	F1-score (Yes)
Values	1.00	0.85	0.82	1.00	0.90	0.92

Table 4: K-Nearest Neighbors Classification Report

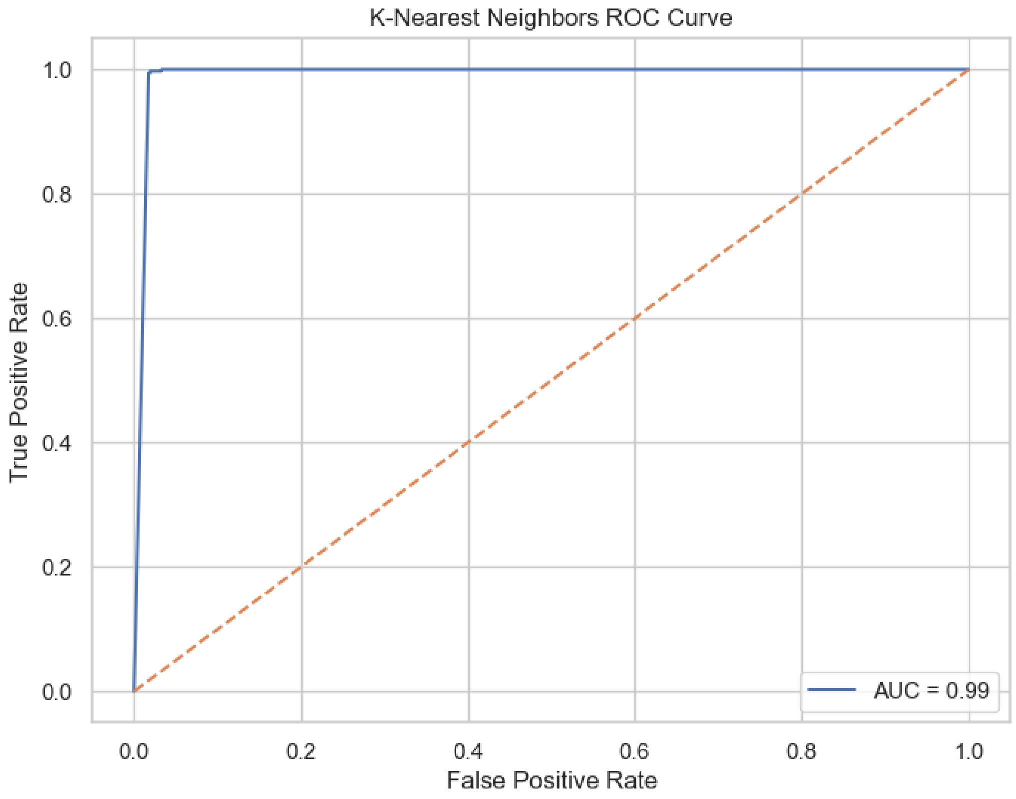


Figure 22: K-Nearest Neighbors ROC Curve

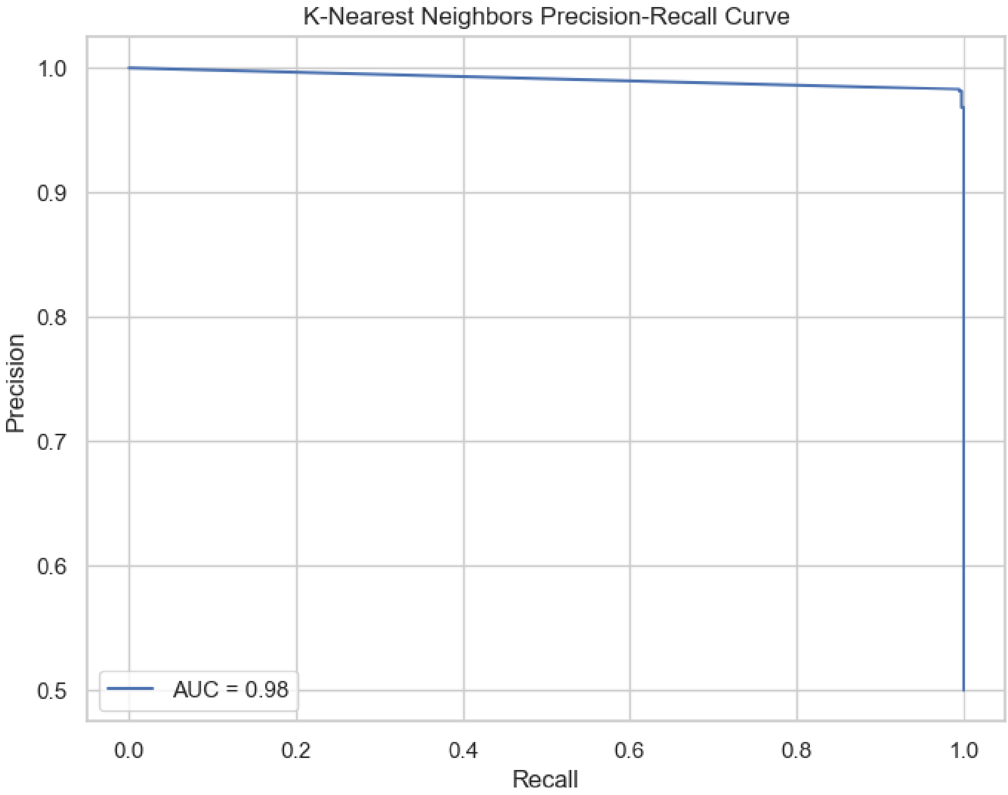


Figure 23: K-Nearest Neighbors Precision-Recall Curve

5.4 Support Vector Machine (SVM)

- **Training Time:** 875.95 seconds
- **Confusion Matrix:**

Confidentiality Notice: This document contains confidential and proprietary information of Core. Unauthorized use, disclosure, or distribution of this document or its contents is strictly prohibited. If you are not the intended recipient, please notify Core immediately and delete all copies of this document.

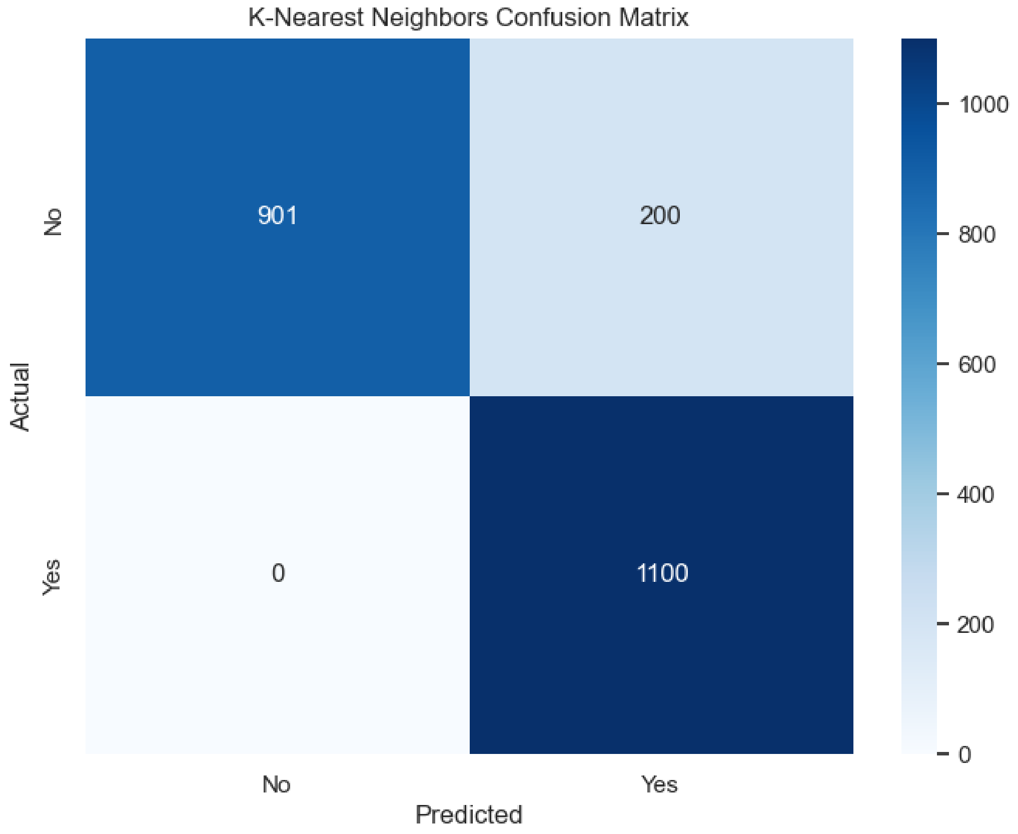


Figure 24: Support Vector Machine Confusion Matrix

- **Classification Report:**

Metric	Precision (No)	Precision (Yes)	Recall (No)	Recall (Yes)	F1-score (No)	F1-score (Yes)
Values	0.99	1.00	1.00	0.99	1.00	1.00

Table 5: Support Vector Machine Classification Report

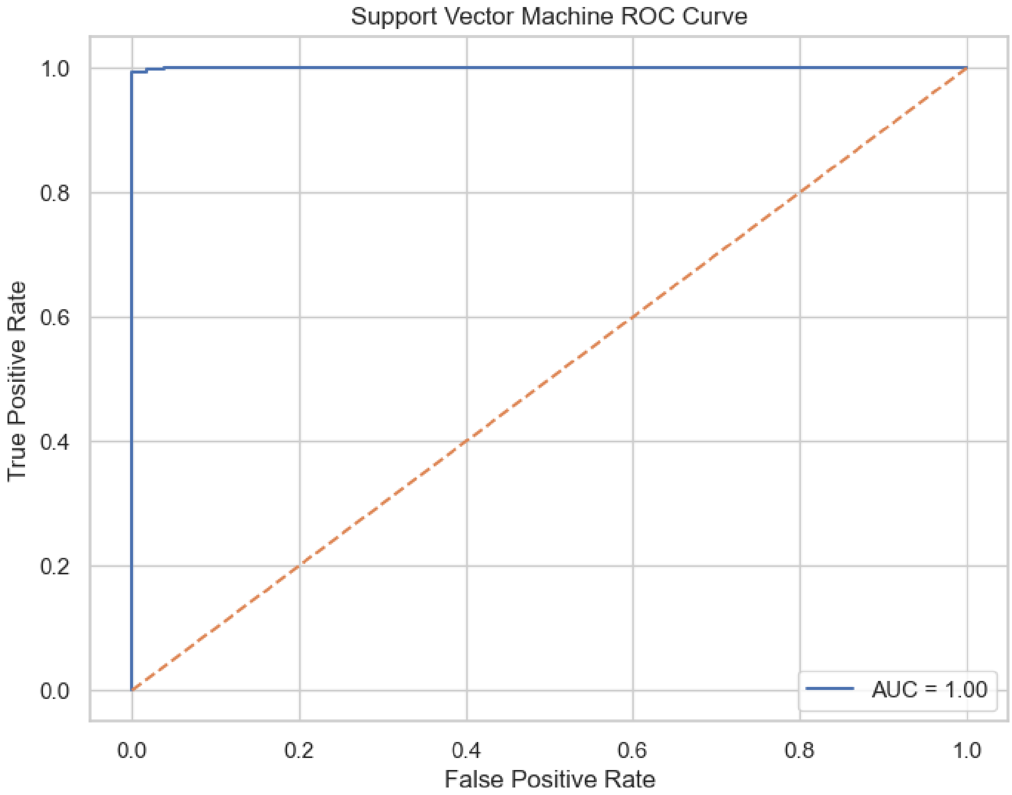


Figure 25: Support Vector Machine ROC Curve

Confidentiality Notice: This document contains confidential and proprietary information of Core. Unauthorized use, disclosure, or distribution of this document or its contents is strictly prohibited. If you are not the intended recipient, please notify Core immediately and delete all copies of this document.

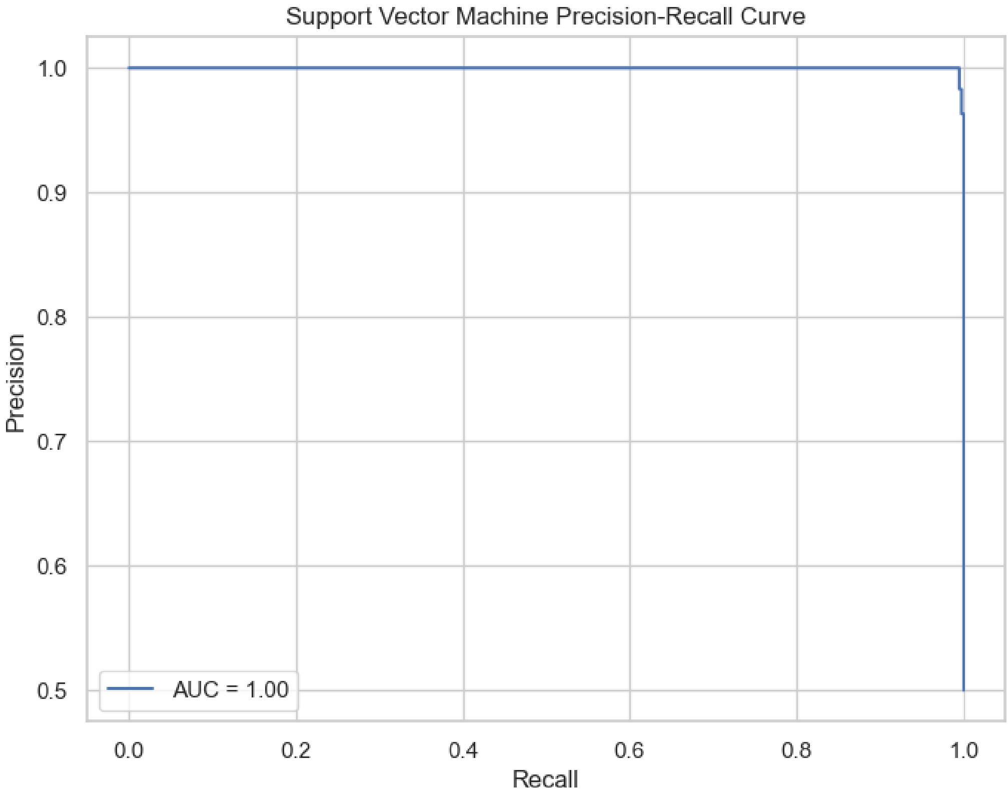


Figure 26: Support Vector Machine Precision-Recall Curve

The SVM model achieved the highest accuracy and AUC but required the longest training time. The Decision Tree and KNN models also performed well, with the Decision Tree showing strong accuracy and KNN demonstrating high recall for the positive class. Logistic Regression, while slightly lower in performance, remains a robust and interpretable model.

6 Feature Importance Analysis

Permutation feature importance was analyzed for SVM and Logistic Regression models due to the lack of overfitting evidence.

6.1 SVM Feature Importance

Top features:

- num_emp.var.rate
- num_duration
- num_cons.price.idx
- num_nr.employed

- num_euribor3m

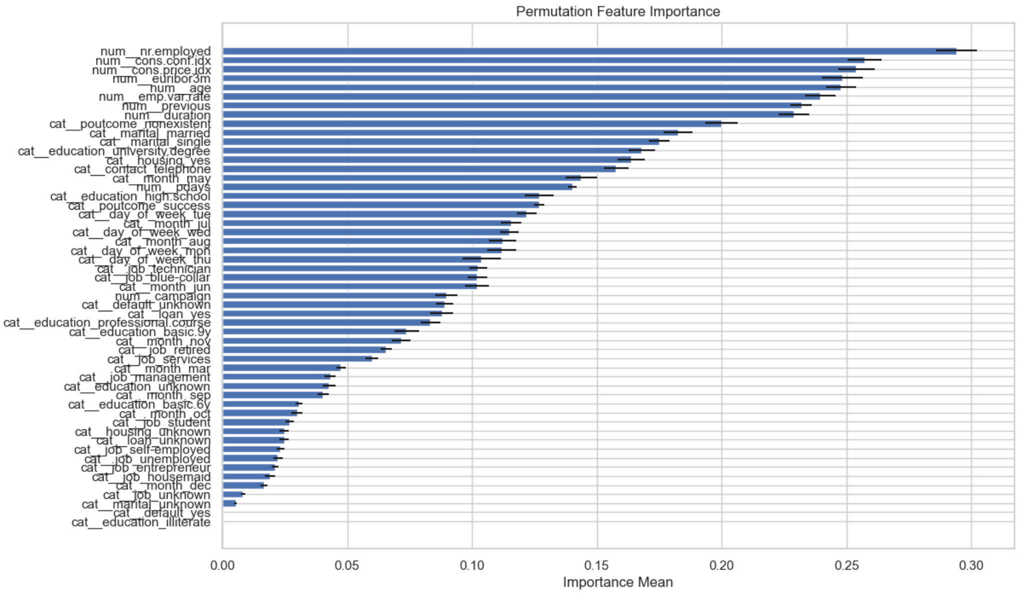


Figure 27: SVM Feature Importance

6.2 Logistic Regression Feature Importance

Top features:

- num_nr.employed
- num_cons.conf.idx
- num_cons.price.idx
- num_euribor3m
- num_emp.var.rate

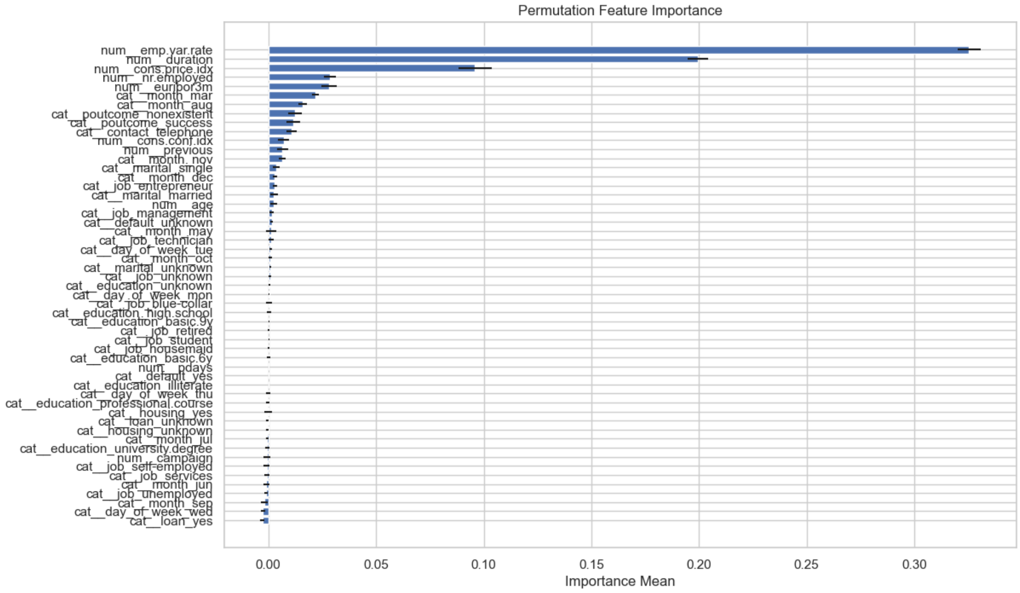


Figure 28: Logistic Regression Feature Importance

6.3 Partial Dependence Plots (PDPs)

PDPs help understand the relationship between features and the target variable. Key Insights:

- **Call Duration:** Moderate durations (6-8 minutes) increase subscription likelihood.
- **Number of Contacts During Campaign:** Fewer contacts are better.
- **Days Since Last Contact:** More days since the last contact increases likelihood.
- **Number of Previous Contacts:** Fewer previous contacts are better.
- **Employment Variation Rate:** Lower rates are better.
- **Consumer Price Index:** Moderate values are better.
- **Consumer Confidence Index:** Moderate values are better.
- **Month of Contact:** Specific months like March, June, September, and December might be less favorable.

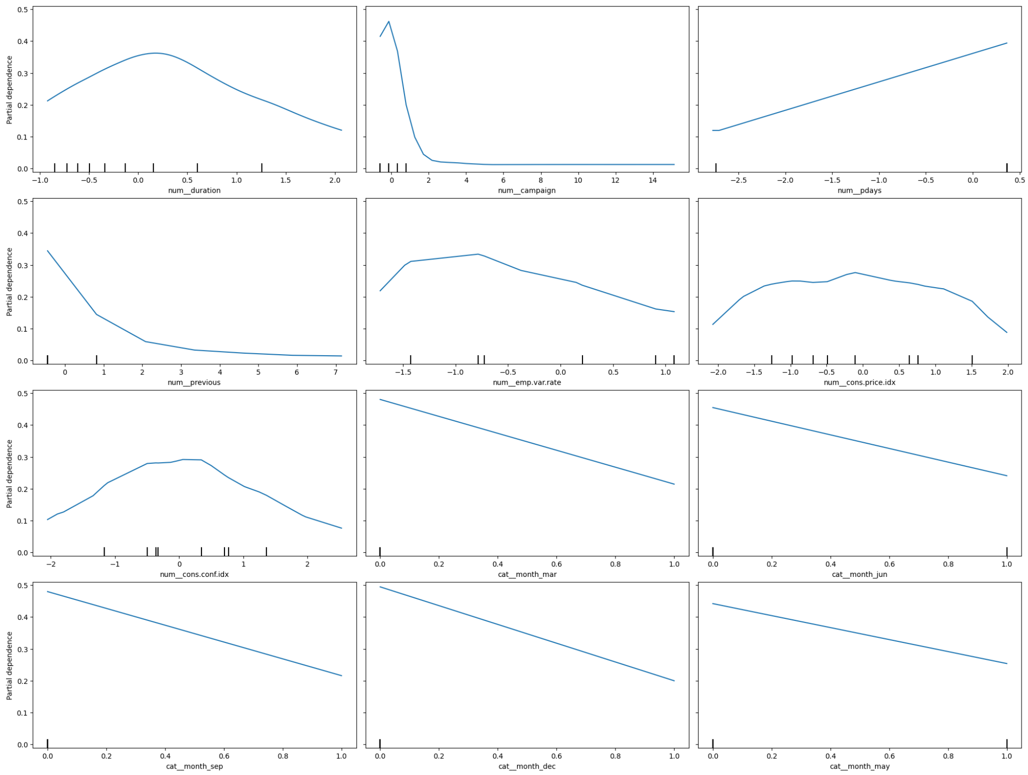


Figure 29: Partial Dependence Plots

7 Business Implications

- **Optimal Call Duration:** Maintain calls between 6 and 8 minutes.
- **Contact Frequency:** Reduce the number of campaign contacts.
- **Re-contact Timing:** Allow more days between contacts.
- **Previous Contacts:** Minimize to avoid contact fatigue.
- **Economic Indicators:** Monitor to identify favorable conditions.
- **Month of Contact:** Adjust timing strategies based on further data validation.

8 Comparison with the Paper Findings

- **Model Performance:** Both my findings and the paper highlight SVM as the best performer [1].
- **Key Features:** Both analyses emphasize the importance of call duration and month of contact [1].

- **Insights:** Align on the effectiveness of fewer contacts and strategic timing for higher subscription rates [1].

9 Recommendations Moving Forward

9.1 Enhanced Feature Engineering

- **Interaction Features:** Develop interaction terms between important features, such as num_duration and num_previous.
- **Temporal Features:** Create features like recency of contact (recent_contact if num_pdays < 7).
- **Aggregate Features:** Calculate mean/median call duration for each customer.
- **Binning Features:** Bin continuous variables like call duration into categories (e.g., short, medium, long).
- **Lag Features:** Generate lag features for temporal data, like the difference in days between consecutive contacts.
- **Encoding Categorical Variables:** Use frequency encoding for categorical variables to retain distribution information.
- **Macro-Economic Trends:** Create moving averages for economic indicators like num_emp.var.rate.

9.2 Model Improvement

- **Ensemble Methods:** Combine models using techniques like Random Forest, XGBoost, Gradient Boosting, or possibly neural network to capture its hidden complexity.

9.3 Business Strategy Optimization

- **Call Duration Management:** Focus on maintaining call durations within the optimal range of 6-8 minutes to maximize subscription likelihood.
- **Contact Frequency Reduction:** Reduce the number of campaign contacts to avoid customer annoyance and improve effectiveness.
- **Strategic Re-Contact Timing:** Plan re-contacts after optimal periods (e.g., 20-30 days) and focus on months that show higher success rates based on historical data.

By implementing these recommendations, the marketing campaigns can become more efficient, targeted, and effective, leading to higher subscription rates and better resource management.

10 Conclusion

The SVM model is recommended for its superior performance and actionable insights. The use of permutation importance and PDPs enhances the model's interpretability, guiding strategic decisions for optimizing marketing campaigns. Further feature engineering can improve model performance and provide deeper insights into customer behavior.

References

[1] Moro, Sérgio and Laureano, Raul M. S. and Cortez, Paulo, Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology, Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal and Universidade do Minho, Guimarães, Portugal, 2011, `https://core.ac.uk/display/55616194`