Stat 451 Final Report

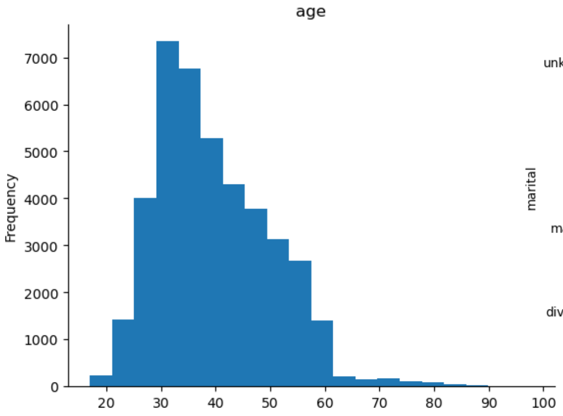
Yujun Che, Carson Trimborn, Janice Yu, Ziyi Yang
Group 2

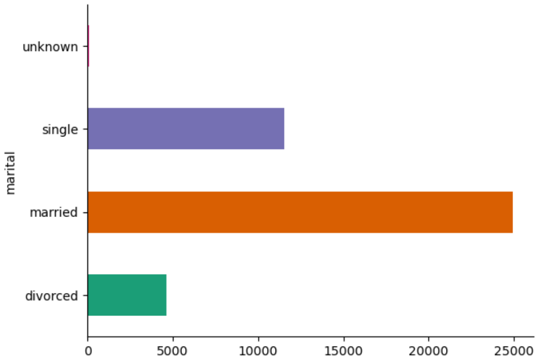
I. Introduction

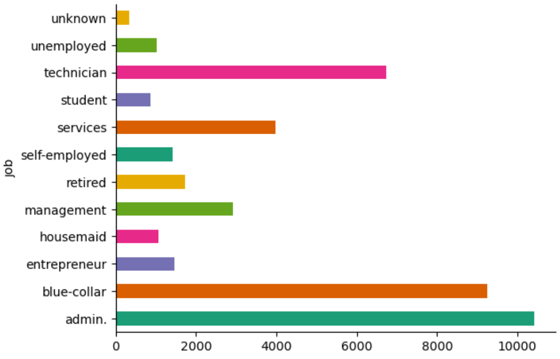
The dataset is collected from a direct marketing campaign via telephone by Portuguese bank to promote its term deposit product. Our question of interest is that whether we could predict the subscription using clients' background information and campaign interaction data. We considered four machine learning models, which are SVM, Logistic regression, PCA, and Decision Tree. We accessed the runtime, interpretability, and accuracy of models, and select 5-step decision tree as final model. We also evaluated the model by cross-validation, ROC, and confusion matrix.

II. Data Pre-procession

Our dataset records 8 categories of personal information of more than 41,000 clients, and whether they agree to subscribe to bank deposit information. The following graphs are visualizations for different variables.







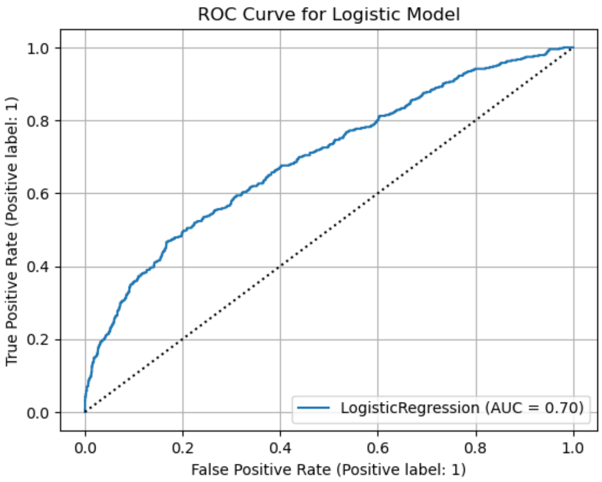
We remove missing values, transform categorical variables by one-hot encoding, replace education level with years of in school, and do data standardization.

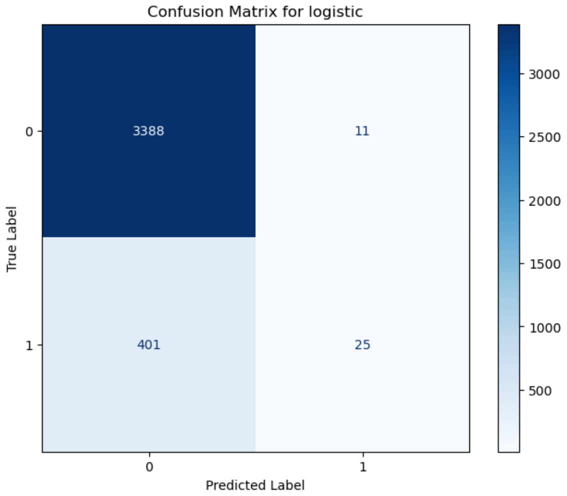
III. Model Selection and Evaluation

Logistic Regression

In building this model, we used the grid search method to tune the hyperparameter C, testing the values 0.01, 1, and 100, and finding the optimal value is 0.01. It achieves an average accuracy of about 89% for cross-validation and test scores.

The ROC curve is the weak point of this model, with an area under the curve of only 0.7. We want to minimize false negatives because the banking institution's primary concern is to minimize clients who are predicted uninterested in subscribing to a term deposit when they are. 401 false negatives potentially lead to missing opportunities in reaching out to clients, resulting in a huge loss in subscriptions and revenue for the bank.

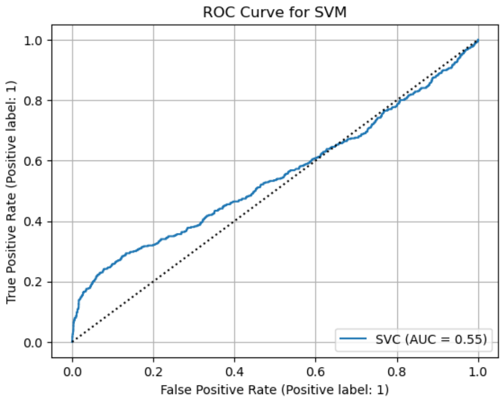


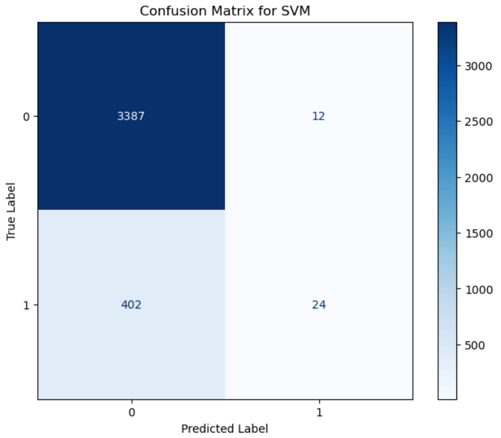


SVM

We use SVM as our second prediction model. Since we are working with non-linear data, we used the RBF (Radial Basis Function) kernel. The average cross-validation score based on the SVM is 89%, and the accuracy for test data is 89.2%.

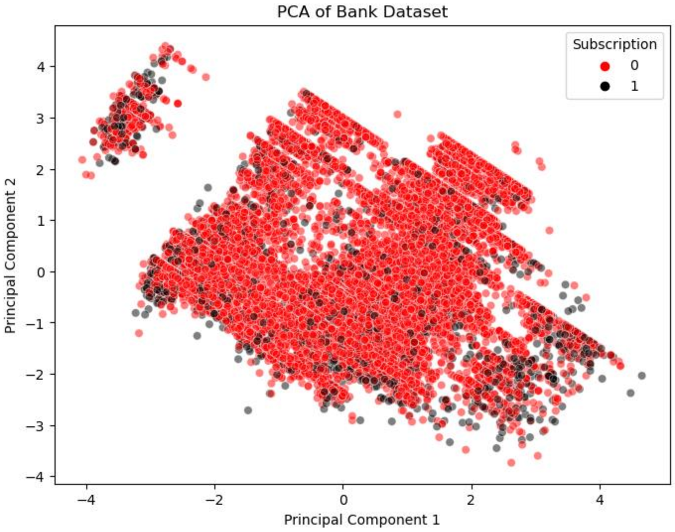
Although the SVM predicts accurately, run time and false negatives are drawbacks of the model. It took about 30 seconds to run the mode. The area under the ROC curve is only 0.55, and test data indicates there are 402 false negative results.





PCA

We construct PCA to provide visual representation of the customer groups and their characteristics. The color of points indicates subscription status. Principal Component 1 records variation related to age and housing loans; Principal Component 2 captures variation related to job type and educational background. We do not find significant distinction between people who subscribe and who are not. Furthermore, information lose is inevitable for PCA model, which decreases its interpretability.

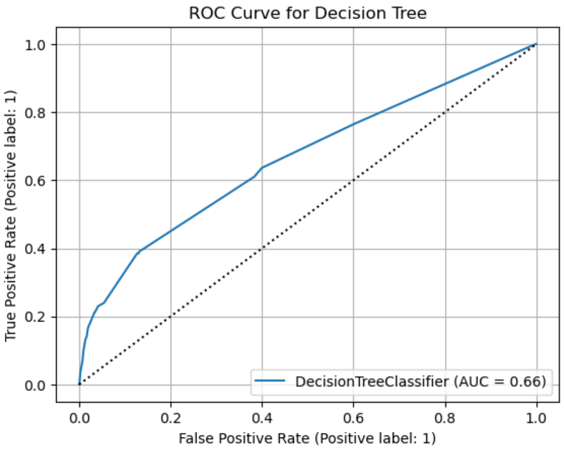


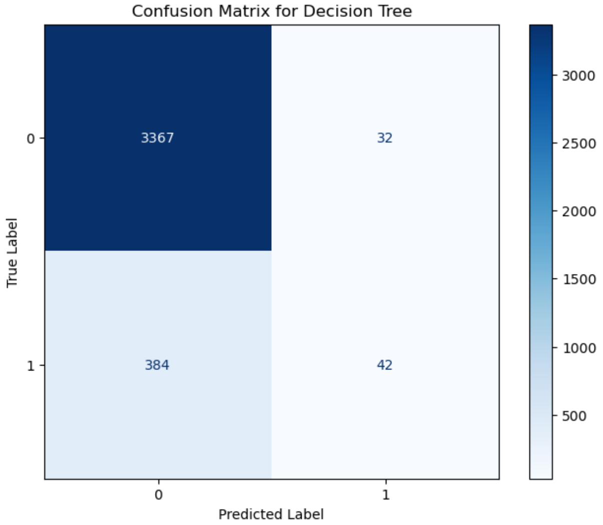
Decision Tree

We select decision tree as our final model. By setting a grid-search of max steps of 1, 3, 5, and 7, we determine to use the model with 5-maximum steps. This model has an average accuracy of 89% in cross-validation and 90% accuracy for test data.

Decision tree also has strengths in interpretability and run time. The flow chart is interpretable for bankers without data science background to decide which client to reach out to. Also, the training and prediction time is relatively short. The decision tree only requires less than 0.1s to train, and its prediction time is almost neglectable.

The weak point of decision tree is its ROC curve, and false negatives are 384.

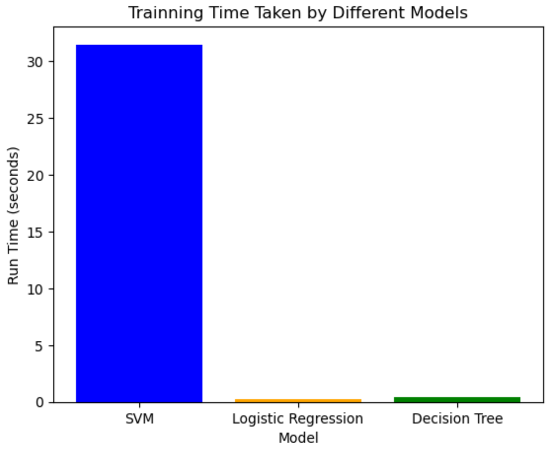


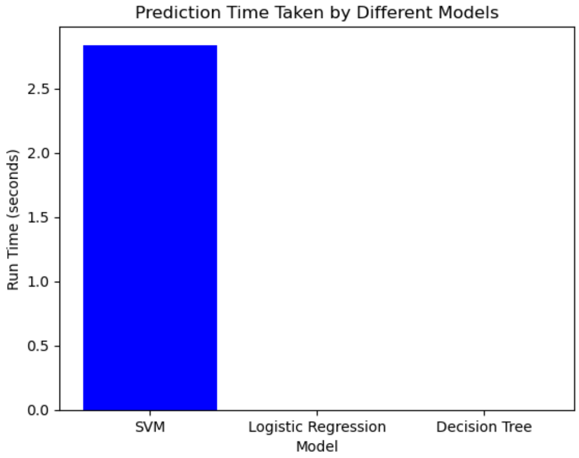


IV. Model Summary

Run Time revisited

Our final model is fast, which great fits the large dataset that banks have.

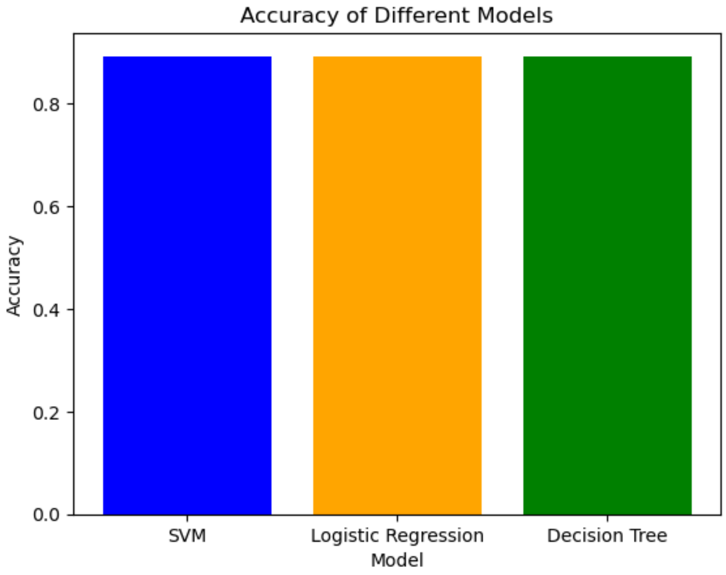




Cross validation and test accuracy revisited

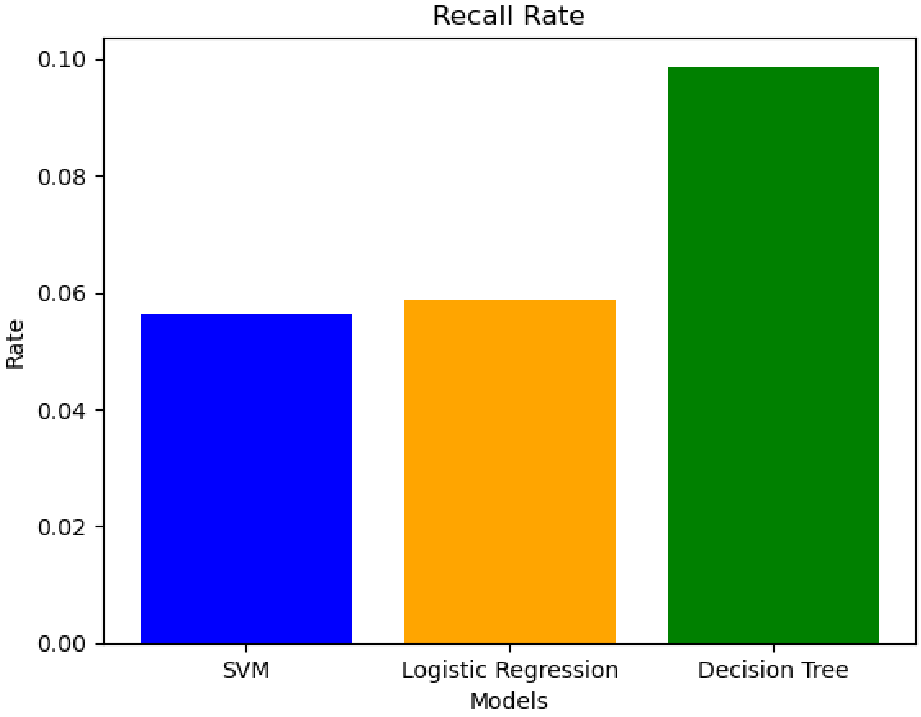
The result of test data also suggests that our final model is highly accurate. Our model correctly captures most clients' behavior.

The score of cross-validation is similar to test data. Therefore, we conclude that there is no significant underfitting or overfitting.



Recall rate revisited

Decision tree captures 9.85% of clients who are interested, which is comparatively strong across all models.



V. Conclusion

Our project reveals key factors that influence customer decisions on subscribing to the bank’s product. We considered the runtime, interpretability, and accuracy of four models, selecting decision tree as the final model. Then we evaluate overfitting problem through cross-validation, and calculate false negatives with recall rate to better understand the performance of the model.

Further work could be focused on two areas. First, add the false positive rate into the grid search step. It could help us build a model with less number of client losses. Second, interpret the result with econometric meaning. It benefits the bank to understand client behavior.

Appendix

Contribution Table

Member	Proposal	Coding	Presentation	Report
Carson Trimborn	1	1	1	1
Janice Yu	1	1	1	1
Yujun Che	1	1	1	1
Ziyi Yang	1	1	1	1, where 1 indicates full contribution

Link to the Code

GitHub: https://github.com/Bedyvere/Stat451_bank_deposit