DATA 622 Assignment 3

Bank Marketing

Susanna Wong
Professor Joseph Sabeljia
March 24, 2025

Introduction

The Bank Marketing dataset was obtained from UCI Machine Learning Repository. It contains 45,211 client data related to direct marketing campaign conducted by a Portuguese banking institution, which is carried out via phone calls. The client data contains 17 features that includes their demographics, i.e., age, education, job, and martial status, their financial status, i.e., bank balance in euros, personal and housing loan status, and credit default status, and campaign related variables. The target variable indicates whether the client subscribed to term deposit or not. For this assignment, we conducted experiments using SVM kernels. The objective is to choose the optimal model for predicting where a client will subscribe to a bank term deposit.

SVM Models

Support Vector Machines are powerful supervised learning used for both classification and regression. They are great with high-dimensional data as it draw a boundary between different groups of data, making it easier to tell which points belong to which class.

All models were scaled as there were some skewed distribution and SVM models are sensitive to scaling as it relies on distance calculations. The linear SVM model default predicts the outcome y (yes or no to term deposits) based on the features in the train data with default setting of `cost` = 1 (regularization parameter), cross = 0 and with scaling. The linear kernel models both default and tuned performed consistently well.

The Radial SVM model predicts the outcome y (yes or no to term deposits) based on the features in the train data with default setting of `cost` = 1 (regularization parameter), cross = 0, gamma = 1/(number of features) and with scaling. It outperformed the others overall, especially once it was tuned, which makes sense since its better at handling complex, nonlinear patterns, as expected in a real world data like this data set. The tuned radial model showed the highest accuracy (0.898) and f1-score (0.944). However, the AUC decreased slightly after tuning the radial SVM model, suggesting a possibility of an overfitting to the majority class.

The polynomial kernel is capable of modeling complex relationships.It did not outperformed the other models.Tuning was not applied to this kernel due to time constraints.

All Model Comparisons

This dataset contains 16 features including categorical and numerical values.It is a moderately large dataset that includes over 45,000 observations.Moreover, class imbalance exists with a subscription rate around 11.7% ("yes" is the minority class,) and there are a large amount of high unknowns in the categorical variables (ie: about 81% of `poutcome` has unknown values.) Considering these dataset characteristics, Random Forest and Adaboost are the recommend algorithm to use to predict the categorical outcome of this dataset (yes or no to term deposit). Both models are robust to mixed data types, class imbalances, and missing values.

Observing the results, the results aligns with the conclusions above. The Random forest and AdaBoost clearly outperformed the other algorithms. They consistently achieve the highest F1-score, which is the most reliable metric for an imbalanced dataset as it balances precision and recall. The particular lower AOC score for Decision Tree is a strong indicator of its struggle with identifying the minority class. Below is a summary of the performance of the models ranked.

Random Forest models:

- Highest F1-Scores: 0.947
- 0.948 - Highest AOC: 0.926-0.928
- Highest Accuracy: 0.904-0.905

AdaBoost models:

- Strong F1-Scores:0.945
- 0.946 (close to Random Forest)
- Strong AOC: 0.920 - 0.926 (close to Random Forest)
- High Accuracy: 0.903 - 0.904

Decision Tree models:

- F1-Scores:0.941 - 0.944 (lower than ensembles)
- AOC: 0.744 - 0.746 (significantly lower than ensembles, which highlights the weakness on imbalance data despite a good F1-score)
- Accuracy: 0.903 - 0.904

SVM models:

- F1-Scores:0.941 - 0.944 (similar to Decision Trees)
- AOC: 0.855 - 0.905 - Accuracy: 0.891-0.898

Random Forest is recommended to get more accurate result as all various Random Forest models performed the best in all metrics compared to the other models.

	Accuracy	**Precision**	**F1_Score**	**AOC**	**Comment on Performance**
Decision Tree (default)	0.8990562	0.9173027	0.9445412	0.7465744	Baseline Performance
Decision Tree (Max-depth = 3)	0.8946321	0.9165153	0.9419978	0.7445256	Slight drop due to restricted depth
Decision Tree (Max-depth = 5)	0.8990562	0.9173027	0.9445412	0.7465744	same as default
Decision Tree (Pruned)	0.8990562	0.9173027	0.9445412	0.7465744	simplified but similar performance as default
Random Forest (default)	0.9053974	0.9203554	0.9480462	0.9282924	Baseline Performance; highest F1-score
Random Forest (ntree=200)	0.9047338	0.9201007	0.9476754	0.9304559	Slight increase in AUC
Random Forest (Tuned-Mty=6)	0.9046601	0.9209552	0.9475775	0.9283361	similar to default

	Accuracy	Precision	F1_Score	AOC	Comment on Performance
AdaBoost (Default)	0.9042177	0.9302119	0.9467295	0.9247494	Baseline Performance with high presion
AdaBoost (mfinal=100,cp= 0.001	0.9018581	0.9326181	0.9451790	0.9209830	lower AUc but has slightly better precision
SVM linear	0.8916826	0.9003887	0.9414671	0.9036006	Baseline Performance
SVM linear (tune d)	0.8916826	0.9003887	0.9414671	0.9089608	similar to default
SVM Radial	0.8952957	0.9085772	0.9429581	0.9059097	Baseline Performance
SVM Radical (tun ed)	0.8980239	0.9125321	0.9442676	0.8909411	Performed the best out of all SVM models
SVM Polynomial	0.8925675	0.9008460	0.9419453	0.8850769	

Article Review

1. Decision Tree Ensembles to Predict Coronavirus Disease 2019 Infection: A Comparative Study This study shows that class imbalance present a challenge when working with medical data, in this case predicting Covid 19, where about 13% of the cases are actually positive. Researchers used techniques like SMOTE and RUS to address class imbalance. Ensemble models like Random Forest, XGBoost, AdaBoost, Decision tree were used along with evaluating metrics like Recall, F1-measure, precision, AUROC as accuracy alone is not the best measure for an imbalanced dataset. Models designed for imbalance data like balanced random forest performed the best and better a single decision tree.

Relevant to our assignment: This article supports my strong recommendation of Random Forest and AdaBoost as they are robust for imbalanced dataset. It also validates my usage of the following metrics besides Accuracy to evaluate the models: Precision, AUC, and F1-score.

2. A novel approach to predict COVID-19 using support vector machine This article compares the usage of SVM against other models like Random Forest and Decision Tree for predicting Covid infections severity. SVM outperform all other models like Naive Bayes, KNN , Random Forest and more. This contrasts with the results I found in our assignments, highlighting model performance could be dataset specific. This articles proves SVM are still super powerful for classifying thing as they are great with high-dimensional data and draw a boundary between different groups of data, making it easier to tell which points belong to which class.

3. [A Comparative Analysis of Decision Tree and Support Vector Machine on Suicide Ideation Detection](https://www.sciencedirect.com/science/article/pii/S1877050923017209) This articles from 2023 provides a comparison between decision tree and SVM in suicide detection using social media data. Again in this article SVM outperforms single decision tree in all metrics (precision, recall, accuracy, f1-score)

4. [Are Random Forests Better than Support Vector Machines for Microarray-Based Cancer Classification](https://pmc.ncbi.nlm.nih.gov/articles/PMC2655823/)?

This article compares two algorithm: Random Forest and SVM. The author challenges a previous belief that Random Forest is superior. This is important as there is no single algorithm that is the best as performance is based on the dataset and the methodology like we learn in article 2. The author discusses about the bias found in prior works due to using the wrong metric.

Previous study relied heavily on accuracy alone which is flawed as the metric is sensitive to imbalance class (like our bank dataset).They also pointed that previous studies did not property tune SVM model.It is important to optimize the models with the right parameters to see the models true potential.

5. [SVM vs Decision Tree Algorithm Cost Effective Comparison to Enhance Crime Detection and Prevention](https://sifisheriessciences.com/journal/index.php/journal/article/view/514/497)

This article focuses on examining how well Decision Tree and SVM perform when used to predict crime patterns.Both models were used on a dataset from 56 states with 30 features. The results showed that Decision Tree significantly performed better than SVM in terms of accuracy with p-value of 0.025. This is a sharp contrast to some of the articles (suicide detection and covid 19 severity prediction) I found where SVM outperforms Decision Tree The articles notes Decision Tree's advantages that includes low computational cost, robust to different data types and its interpretability.