Index

1. Introduction

World Plus Pvt Bank has approached VisionSphere Consultancy (our analytics consultancy company) to develop and deploy a lead prediction system to improve their lead conversion by efficiently identifying and aiming at their target customers who will most likely purchase their banking products. The bank offers various services such as credit products, saving accounts, loans, investments, etc. World Plus is facing challenges in identifying their target customers accurately and thus wants us to formulate the lead prediction system for their upcoming new term deposit product. Their motive is to save both time and money. The project will make use of a provided dataset of 220,000 historical customer data records.

2. Background

In this project, we focus heavily on the data mining process to better understand the provided data set and ensure it is ready for use by our selected models: Logistic Regression, Decision Tree, Random Forest and Support Vector Machine (SVM). We apply encoding techniques to text fields and convert those values into factors. The main issues in the data set include some erroneous values in dependent status and missing values in the credit product column. We decided to remove the errors in dependent status as it did not provide us with a significant information gain. Therefore, we assumed that it would not impact the predicted results.

On the other hand, we applied hot deck imputation to the missing values in the credit products column [6]. By observing the data set, we assume that these values are missing completely at random. Thus, this method would give us some variation in the predicted value compared to mode imputation. We also created another data set where all the missing values were removed, and we applied SMOTE balancing techniques to get the target value back to the same proportion as the original data set. This project will compare the results of each model and data set to identify the most effective prediction model.

3. Relevant Academic Papers

3.1. The prevention and handling of the missing data [1]

There is a very common issue with data across domains - missing data. It concerns data values not stored or recorded for a given variable. The major reason to choose this research paper is that we, too, have such issues in our dataset, and it could have various implications, including reducing the statistical power, causing data bias, and sometimes making the analysis more complicated.

The research focuses on understanding different types of missing data, followed by multiple data handling techniques used to rectify them.

Data quality can be ensured by planning the data collection phase upfront, conducting participant training, and efficiently designing studies. We can implement some missing data handling techniques like listwise and pairwise deletion, regression imputation and maximum likelihood.

Even after making these necessary adjustments, potential bias remains. Therefore, the best solution is to maximise data collection. Multiple imputations provide robust solutions for datasets with missing data, especially for larger datasets.

3.2. Data Mining based Handling Missing Data [2]

The main reason this paper was selected was to give us a better idea of how to handle missing values. This paper presents techniques for handling different types of missing values within the dataset. These techniques can be categorised into two types: missing data deletion and missing data imputation. The data deletion technique basically removes those with missing values. However, it is applicable only when the missing ratio is small, and the result produced by the analysis of the complete dataset will not be biased. On the other hand, there are many techniques classified as missing data imputation. There are three approaches that different techniques use to impute missing variables. First, the global approach applies the global correlation information from the whole dataset. Second, the local approach utilises the local similarity structure in the dataset. Finally, the hybrid approach considers both local and global structures.

3.3. Bagging for linear classifiers (Skurichina and Duin, 1998) [4]

This paper hypothesises that bagging, the combination of bootstrapping and aggregating, could be promising as a classification technique as it is robust and may provide a more stable solution. Bootstrapping is a random sampling technique, allowing for more accurate statistical estimators. The method involves replacement, allowing the bootstrap training data to include fewer outliers. Aggregating, on the other hand, involves combining classifiers as they give better results than individual classifiers.

Overall, the paper fails to identify whether the techniques stabilise classifiers or just improve their performance. It also concludes that the method may only be useful on more unstable data sets.

From this, it could be time-consuming to implement and may still be biased or unstable if the data lacks proper representation. To avoid these issues, we should utilise a larger training data set, which is what we implemented.

3.4. Comparison of RF and SVM for regional land cover mapping using coarse resolution FY-3C images [5]

This paper aims to determine which algorithm among RF and SVM produces more accurate results and is more efficient for large-area mapping of parts of Africa using coarse-resolution FY-3C images. The two most important parameters were defined for each algorithm, namely, the number of trees and the number of variables for RF and penalty value and gamma for SVM, used to gain the best model for each of the algorithms. The findings showed that RF outperformed SVM in terms of accuracy, where RF achieved an OA of 0.86 and k of 0.83, and these values were 1-2% and 3% higher than the best SVM model, respectively. The study concluded that RF was a better model for heterogeneous large-area mapping and more effective in handling large datasets. Moreover, SVM was memory expensive and generated noisy products because of more confusion among classes. This paper clearly summarised the implications and limitations of both algorithms, which helped us get clarity for our assignment.

3.5. Modelling Purchase Behaviour at E-Commerce Website: A Task-Completion Approach [7]

This literature paper was chosen to showcase the predictive modelling approach to identify potential buyers for an e-commerce website. The article aimed to develop a modelling approach to predict the online buying behaviour of web users and examine the clickstream data collected by the operator of a major commercial website. The analysts proposed a conditional probability approach to breaking down the website activities into sequential nominal user tasks (NUTs) to model the probabilities based on task completion rates at each level. Therefore, instead of developing a model to predict the buy/non-buy outcome, this approach was based on the result of a series of conditional probabilities to predict the results at each level of NUTs defined. In this article, they have used the Bayesian approach, implemented with Markov chain Monte Carlo (MCMC) methods, to estimate the joint model of task completion. Their proposed modelling approach consisted of MULTI, outperforming the remaining approaches, such as SINGLE 1 and SINGLE 2. The article also highlighted that the multilevel task completion approach provides superior targeting capability and better measures of predictive accuracy.

3.6. Predicting students' performance using ID3 and C4.5 classification algorithms [3]

To predict students' performance based on their previous records, ID3 and C4.5 classification algorithms were used. During the data preparation stage, missing values were ignored because they had less negative impact on the accuracy of the model. ID3 and C4.5 classified algorithms were used to generate a decision tree. Finally, a web page was introduced after modelling and evaluation to map the results to faculties. Lastly, creating a visual display such as a web page is beneficial to show the results for designated people. The research paper had a similar objective to predict future results using a decision tree. It helped us to understand the whole process of developing and deploying a lead prediction system.

4. Data Analysis (Modelling)

A classification model is needed to create a lead prediction system according to the business requirements. In addition, we wanted to avoid unnecessary expenses for uninterested customers and at the same time, we wanted to capture potential buyers as much as possible. For this reason, high precision, recall and F-1 Score should be the main criteria for this model. Based on our group research paper, we have created four models, including Linear Regression, Decision Tree, Random Forest and SVM, to evaluate their performance on the two data sets: removed NAs and Imputed NAs. In accordance with our research, all models performed significantly better on the imputed NAs data set in every aspect. We also observed that building the model exclusively with positive information gain values yielded results similar to the original training set. To prevent overfitting and streamline model complexity, we concluded by favouring the model constructed solely with positive information gain.

Below is a detailed comparison between the results of each model.

4.1) Logistic regression

Model	Accuracy	Kappa	Precision	Recall	F1	Balance Accuracy
Linear Regression	90.07%	53.77%	75.37%	48.65%	59.12%	72.95%

- With fair recall, 48.64%, which shows a limitation in avoiding false negatives. It can be concluded as having moderate performance.

4.2) Decision Tree C50

Model	Accuracy	Kappa	Precision	Recall	F1	Balance Accuracy
Decision Tree C50	90.74%	56.9%	78.96%	50.88%	61.89%	74.27%

- Better overall performance with slightly higher recall compared to Linear regression.

4.3) Decision Tree Ctree

Model	Accuracy	Kappa	Precision	Recall	F1	Balance Accuracy
Decision Tree Ctree	90.76%	56.46%	80.27%	49.68%	61.38%	73.78%

- Highest precision across all models however, recall is only 50%, which shows a limitation in avoiding false negatives.

4.4) Random Forest

Model	Accuracy	Kappa	Precision	Recall	F1	Balance Accuracy
Random Forest	90.79%	57.98%	77.36%	53.18%	63.03%	75.24%

- Maintained a good balance between recall and precision. The Kappa were also significantly higher than those of other models.

4.5) SVM

Model	Accuracy	Kappa	Precision	Recall	F1	Balance Accuracy
SVM	90.71%	57.48%	77.2%	52.57%	62.566%	74.94%

- The Model summary was moderate however, it took significantly longer than any other model.
- There is a moderate recall of 52.57%

Figure 1. Comparison of Predictive Model

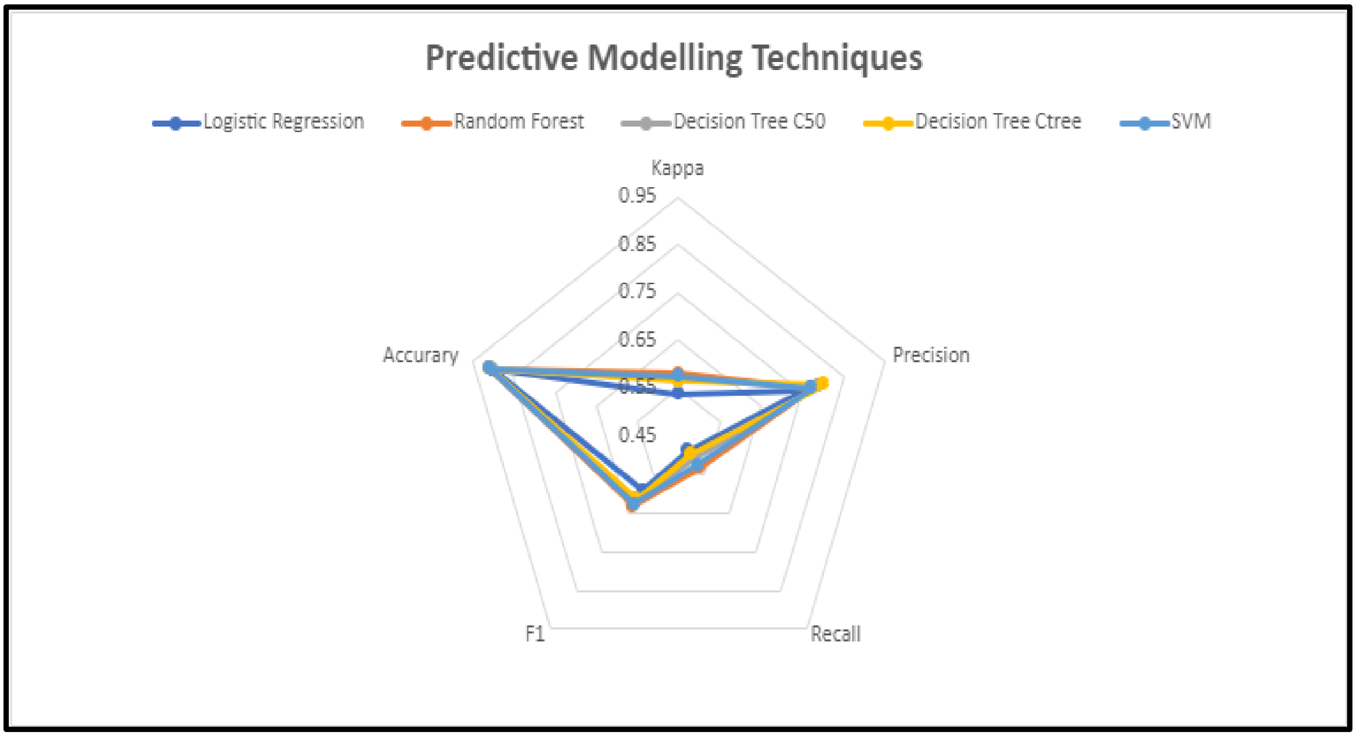

Figure 1.

Figure 2. ROC Graph of each model

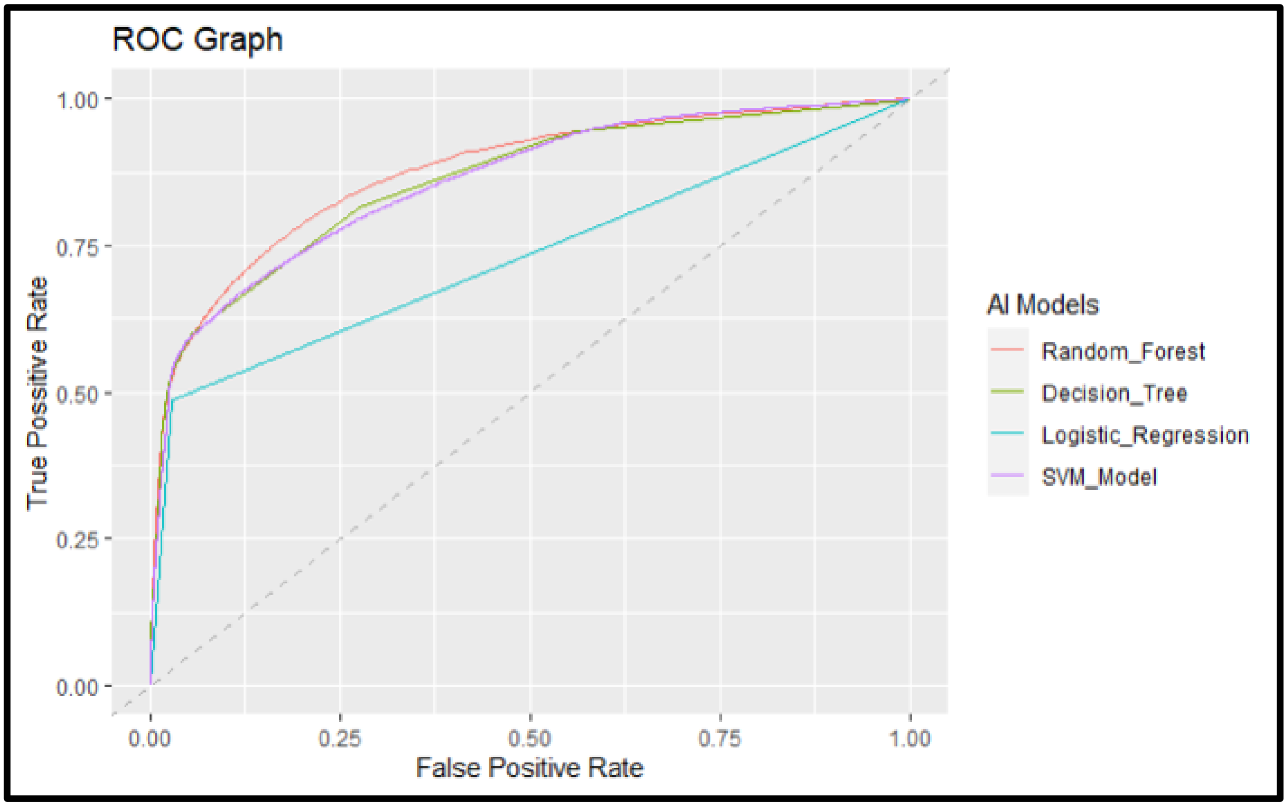

Figure 2.

Table 1. Area under Curve of each model.

(Table 1)

Model	Area Under Curve
Logistic Regression	0.9007
Decision Tree	0.8596
Random Forest	0.8780
SVM	0.8604

Figure 3. Bar graph comparing each model by metrics.

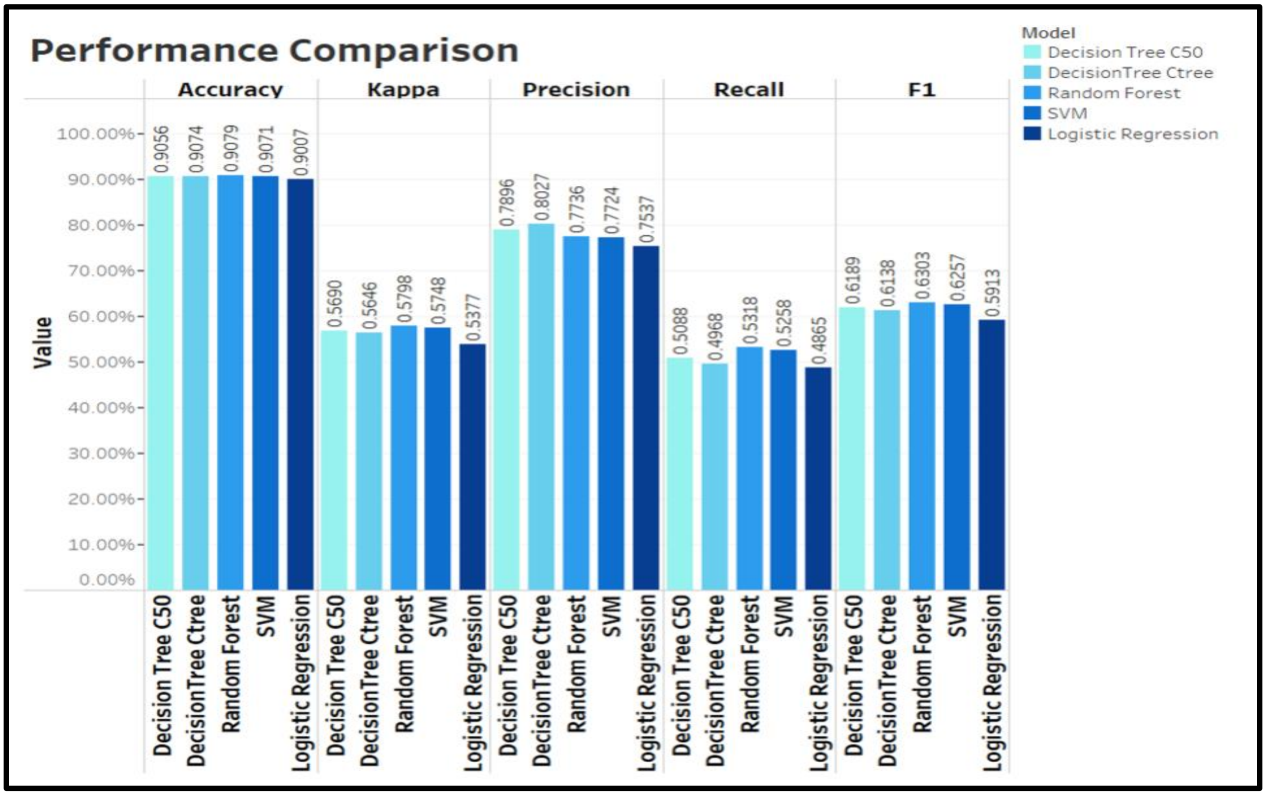

Figure 3.

From this, Random Forest appears to perform well across multiple evaluation metrics regarding precision, recall and F1 followed by SVM, Decision tree, Linear Regression and SVM.

5. Constraints Encountered/ Identified

- The company did not provide information on the pricing model, which could have enhanced our prediction performance and improved our data balancing method and evaluation process.
- The SVM model took much more time to learn and provide results compared to other models.
- Models are based on some imputed values via the hot deck imputation method, which the results may varies when used with actual dataset.

6. Conclusion

Overall, we arrived at the conclusion that random forest was the best model to move forward with. We found that other models that we tried lacked the same levels of precision and recall for the project. These models were built with imputed values as the dataset contained missing data and some incorrectly formatted. Another problem we faced was a trade-off between precision and recall. This forced us to reconsider which feature we value the most, leading us to collectively decide we wanted to prioritise precision. The SVM model was especially limited as it was also very time-consuming to carry out. It was found that the random forest model was more stable than the decision tree as it combined predictions from diverse trees whilst also involving less risk of overfitting. Through this, there was an overall balance between precision and recall, leading to us selecting the random forest model as the most optimal.

7. References

1.) Kang, H., 2013. The prevention and handling of the missing data. Korean journal of anesthesiology, 64(5), pp.402-406.

2.) Dubey, A. and Rasool, A., 2019, December. Data mining based handling missing data. In 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 483-489). IEEE.

3.) Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R. and Honrao, V., 2013. Predicting students' performance using ID3 and C4. 5 classification algorithms. arXiv preprint arXiv:1310.2071.

4.) Skurichina, M. and Duin, R.P., 1998. Bagging for linear classifiers. Pattern Recognition, 31(7), pp.909-930.

5.) Adugna, T., Xu, W. and Fan, J., 2022. Comparison of random forest and support vector machine classifiers for regional land cover mapping using coarse resolution FY-3C images. Remote Sensing, 14(3), p.574.

6.) Andridge, R.R. and Little, R.J., 2010. A review of hot deck imputation for survey non-response. International statistical review, 78(1), pp.40-64.

7.) Sismeiro, C. and Bucklin, R.E., 2004. Modeling purchase behavior at an e-commerce web site: A task-completion approach. Journal of marketing research, 41(3), pp.306-323.