SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

Kaggle Competition Project Report

40.016 : The Analytics Edge

Predicting the Choice of Bundles of Safety Features in Cars

Group 11

May Term 2024

Paige Trinity Tan (1006972)

Li Xing (1007031)

Tan Yan Zu, Joe (1006864)

Destor Rose Evangeline Anne Dagman (1006988)

1 Introduction

The aim of the Kaggle competition was to predict the probabilities that develop an model that accurately predicts customer preferences for new car safety features based on survey data. Customers provided their choices among various bundles of safety attributes, including price, with each answering 19 questions. The training dataset consists of 21,565 observations, while the test dataset includes 4,997 observations. The goal is to use any method or package in R to create the most effective prediction model, evaluated using the multi-class log-loss metric.

2 Approaches Utilised

Our group aimed to maximise the accuracy of our models. We tried a variety of models, namely:

- Random Forest (RF)
- Multinomial (Mlogit)
- Support Vector Machine (SVM)

2.1 Data Pre-processing

Firstly, we checked for missing data entries within the given dataset. Next, regardless of model, we added a "Choice" column and assigned a value to "Choice" based on columns Ch1, Ch2, Ch3 and Ch4. When looking through the dataset, we noticed there were variables that we could drop.

With SVM, it requires all input data to be numerical. Thus, during data preparation, all of the variables with characters will be replaced by its numerical representation which is already provided to us in the dataset. Additionally, with the different ranges of value each variable has, we normalised the value to ensure that the difference will not affect the model in deciding the importance of the variable and give a better predicted result.

In the adjusted model, we run a for loop to find out what is the most optimal number of predictors and trees that give the lowest out of bag (OOB) error for RF. From our findings, we obtained optimal number of predictors (mtry=29) and number of trees (ntrees=600), which gives the least out of bag error of 0.516 shown below (See Appendix, Figure 1). For multinomial, we dropped variables with an arbitrary benchmark with p-value of $10e^{-10}$ or higher as some variables had a much higher p-value compared to the rest, affecting the choices to a lesser extent.

2.2 Types of Approaches

2.2.1 Random Forest Model (RF)

The Random Forest model (DataCamp, 2023) is celebrated for its robustness to overfitting, achieved by creating multiple decision trees from different data subsets and features. This ensemble approach boosts accuracy by aggregating multiple tree predictions. Random Forests are versatile, suitable for both classification and regression tasks, and can handle both categorical and numerical data. They also provide feature importance estimates, handle missing values, and scale well to large datasets, enhancing efficiency for big data applications. The ensemble method reduces variance, improving generalisation, and helps in identifying outliers.

2.2.2 Multinomial (Mlogit)

A multinomial model (ScienceDirect, 2023) extends binary logistic regression to handle outcomes with more than two categories, making it suitable for multi-class classification tasks. It provides probability estimates for each class and interpretable coefficients, helping to understand the relationship between predictors and outcomes. The model is straightforward to implement and computationally efficient, allowing for comparisons against a baseline category.

2.2.3 Support Vector Machine (SVM)

A Support Vector Machine (SVM) (IBM, 2023) is a versatile algorithm primarily used for classification, but also for regression and outlier detection. Similar to Kmeans and Heuristic clustering, SVM will cluster the data into different group. However, different from the 2 clustering methods we learnt in class, SVM is supervising learning. Thus, it will clearly indicate the specific groups of the predicted result corresponding to the train data. It works by finding the optimal hyperplane that maximises the margin between classes, handling high-dimensional spaces effectively. SVMs can manage non-linear data using kernel functions such as the radial basis function (RBF). This increase the flexibility in differentiating the non-linear data with similar characteristic.

2.3 Models Implemented

During the course of the week, we implemented various combinations of models. However, we chose to focus and document on the 4 main models that provided us with the best accuracy. We had differing combinations of RF, Mlogit and SVM, which we will further explain in the sections below.

2.3.1 RF + Mlogit

For our very first model, we started out with a generic model of RF. The data we first used was not normalised. The results of the model produced were not as good as compared to the model with normalised data (Refer to table 1, Model 1 is not normalised data while Model 2 uses normalised data).

2.3.2 SVM (linear) + RF + Mlogit

When building our SVM model, we had first started of with a linear kernel, as that was the simplest kernel out of the options we had.

2.3.3 SVM (linear + radial)

Due to the unknown nature of our datasets, we decided to combine both linear and radial model to increase the accuracy for prediction. This also help us to prevent overfitting which is a common issue face when using non-linear models. At the same time, reduce the rigidness in predicting when merely using linear model.

2.3.4 Multiple Combined Models

For this model, we decided to combine the results of model 3, alongside an updated version of model 1 (which utilises normalised data instead). The combined models each have a weight of 0.5 and we decided to combine the models as the different characteristics of the model help us cross-check and balance the predicted results, particularly when some outcomes have similar predictions. By leveraging these varying aspects, we can evaluate and validate the predictions more thoroughly, ensuring they are more robust and reliable. This approach allows us to identify patterns or discrepancies that might not be immediately obvious, ultimately enhancing the accuracy and effectiveness of the model.

3 Results

The results for the 4 respective models as mentioned in Section 2.3 are as followed in Table 1, as shown below:

	Model Used	**Private Leaderboard**	**Public Leaderboard**
1	RF + Mlogit	1.24045	1.24054
2	SVM (linear) + RF + Mlogit	1.23360	1.23600
3	SVM (linear + radial)	1.22781	1.23439
4	0.5*(SVM (linear + radial)) + 0.5*(Adjusted RF + Mlogit)	1.22676	1.23042

Table 1: Tabulation of Respective Model Scores

4 Interpretability of Approaches & Models

The interpretability of machine learning models varies significantly. Random Forests are generally less interpretable than simpler models like decision trees due to their ensemble nature, though they provide some insights through feature importance (Simon, Glaum, & Valdovinos, 2023). Multinomial models, as an extension of logistic regression, offer relatively high interpretability with understandable coefficients that show the relationship between predictors and outcomes (Pattnaik, 2021). However, their complexity increases with more outcome categories.

Support Vector Machines (SVMs) are the least interpretable among the three. Their reliance on support vectors and kernel functions, especially in high-dimensional spaces, makes understanding their decision-making process difficult (Dolan, 2019). While powerful in classification, SVMs are less suitable for applications requiring high interpretability. In summary, Multinomial models are the most interpretable, followed by Random Forests, with SVMs being the least interpretable.

5 Limitations of Approaches & Models

We do acknowledge that all models come with their limitations. Random Forests can become complex with many trees, leading to slower predictions and higher memory usage, and they are less interpretable than simple decision trees. They may perform poorly on imbalanced datasets and require careful tuning of hyperparameters such as the number of trees and maximum depth (DataCamp, 2023). While generally robust, they can still overfit with noisy data. Training can be time-consuming, especially with large datasets, and they may not perform well with high-dimensional data unless dimensionality reduction techniques are used.

Similarly, multinomial models assume classes are mutually exclusive and independent, which may not always be true, and they presume a linear relationship between predictors and log-odds of outcomes. Data sparsity can be problematic, especially with few observations in some categories. While less intensive than some models, they may still face scalability issues with large datasets or many categories (ScienceDirect, 2023). Overfitting is a risk if not properly regularised, and interpreting coefficients becomes complex with numerous outcome categories.

In the case of SVMs, they can be less effective on large datasets due to computationally intensive training that scales quadratically with the number of samples. Selecting the right kernel and tuning hyperparameters such as the regularisation parameter and kernel parameters can be complex and time-consuming (IBM, 2023). Additionally, SVMs do not directly provide probability estimates, which can be a limitation in applications requiring probabilistic outputs.The standard SVM model focuses on finding the hyperplane that best separates the data into different classes. However, if probabilistic outputs are needed, one must apply additional techniques, such as Platt scaling, which fits a logistic regression model to the SVM's decision values to produce probability estimates. This additional step is necessary because the raw output of an SVM (the distance from the hyperplane) is not directly interpretable as a probability.

Mentioned above are the limitations of each individual approach, when taking the models we had built into consideration, we note that the models which had any sort of SVM variation within it required the longest training time, averaging at 3 hours for each model. It is also noted that a large amount of computing power is required for the SVM models. During the course of the week, we had visited the Data Analytics Lab multiple times, as our personal devices could not handle the long duration and computing power required by these models. We believe that with better hardware, the computing time may be cut short, and we could have produced better results as we would be able to try more combinations of things.

Appendix

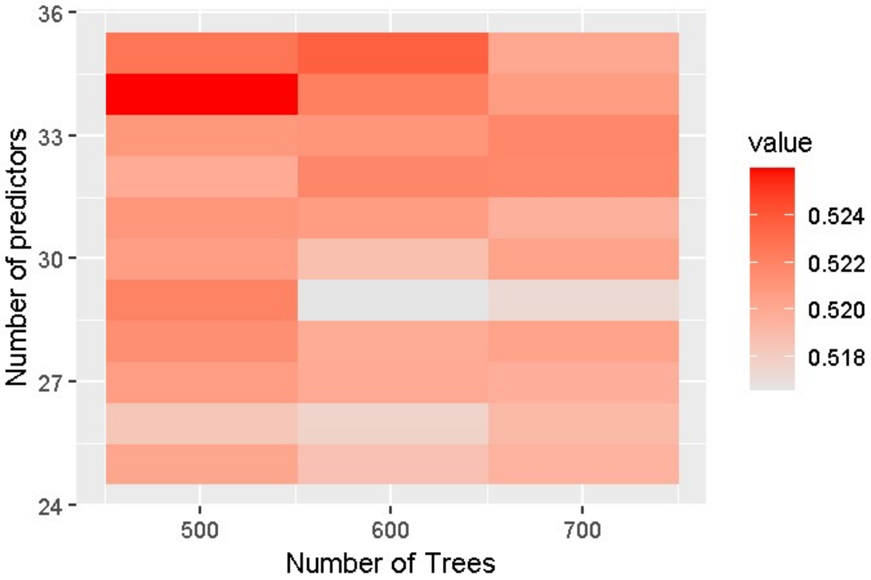

Figure 1: Chart denoting OOB error

References

DataCamp. (2023). *Random forests in python.* Retrieved from https://www.datacamp.com/tutorial/random-forests-classifier-python (Accessed: 2024-08-04)

Dolan, H. (2019). *A practical guide to interpreting and visualising support vector machines.* Retrieved from https://towardsdatascience.com/a-practical-guide-to-interpreting-and-visualising-support-vector-machines-97d2a5b0564e (Published in Towards Data Science, Accessed: 2024-08-04)

IBM. (2023). *Support vector machine.* Retrieved from https://www.ibm.com/topics/support-vector-machine (Accessed: 2024-08-04)

Pattnaik, S. (2021). *Logit — global and local interpretability in python.* Retrieved from https://towardsdatascience.com/logit-global-and-local-interpretability-in-python-f554acb541e4 (Published in Towards Data Science, Accessed: 2024-08-04)

ScienceDirect. (2023). *Multinomial logit model.* Retrieved from https://www.sciencedirect.com/topics/mathematics/multinomial-logit-model (Accessed: 2024-08-04)

Simon, S., Glaum, P., & Valdovinos, F. (2023). Interpreting random forest analysis of ecological models to move from prediction to explanation. *Scientific Reports*, *13*, 3881. Retrieved from https://doi.org/10.1038/s41598-023-30313-8 doi: 10.1038/s41598-023-30313-8