Predicting Charitable Donations

Rajasekhar Malireddy
s1929109
s1929109@ed.ac.uk

Anna Ivahnenko
s1243073
s1243073@ed.ac.uk

Indraneel Brahma
s1981533
s1981533@ed.ac.uk

Abstract

The KDD Cup '98 data mining challenge aimed to identify loyal charity donors and maximise the amount that charity can raise in the direct mail campaigns. We compared 3 methods: Logistic Regression, Random Forest, XGBoost and their ensemble on how well they are able to identify future donors and estimate how much they will donate. The Logistic Regression method performed better than other methods and allow us to predict the outcome of campaigns. Also during analysis were identified the demographic factors that characterise individuals who contribute to good causes allowing the charity to save time and money.

1 Background

The second KDD knowledge discovery and data mining competition was held in June - August 1998. Where the aim was to analyse data collected by Paralyzed Veterans of America (PVA), a non-profit organisation which provides programs and services for veterans with spinal cord injuries or disease [1]. Direct mailing campaigns were used to raise funds, as they allowed the charity to contact people who gave in the past. The charity collected data on "over 13 million donors, PVA is also one of the largest direct mail fundraisers" in America [2]. This allowed the participants to apply an ML data-driven approach for problem-solving.

The charity was interested in contacting lapsed donors, people who made their last donation 13 -24 month prior to June 1997 and therefore stopped donating for at least 12 months. The available dataset contains a record for nearly 200 thousand lapsed donors, who received the mailing campaign letter in 1997. The data contains information collected about each donor along with their response to the latest campaign. Since there is a $0.75 cost of mailing each campaign the goal of this project is to save the time and money for the charity by allowing them to target only the most probable potential donors based on past response data.

This project aims to improve this analysis using modern and robust methods of feature selection, classification and regression to maximize profits. This approach could be also applied to similar marketing, e-commerce, medical procedure, credit assessment, fraud detection and many more real-life problems in the future [3].

2 Exploratory Analysis

The original dataset contained both learning and test sets that were provided in the original KDD Cup'98 competition. The target variables were added to the test set separately to allow validate model results with the previous competition entries. Although the test set had 48% of records, it was not used for training or validation. To get better final results, a part of the test data could be used to adjust the model, which structure was defined on the learning sample.

The dataset consists of 190,000+ records and 481 categorical, ordered or quantitative variables. This information was overlaid with 3rd party demographic data which increased the size of the data set, but added more noise. There was a clear necessity to preprocess the original dataset since it contained many missing values and some formatting errors. Encoding and data transformations were required for ordered and categorical variables where information might be lost otherwise.

The dataset contains two target variables, one binary indicating if a person responded (Target_B) and the other is the donation amount the person gave to the charity as a response to the campaign (Target_D). It is ideal to send direct mail to all the people who will respond however it is useful to target people who are more likely to donate more also. This might be useful to target new possible donors who have similar demographic characteristics to the top donors as well.

From the training set, it is apparent that only 5% (4,843 out of 95,412) previous donors gave a donation. Working with a large unbalanced dataset presents a big challenge in terms of a signal to noise ratio and ability to fit a model to account for such disproportional response.

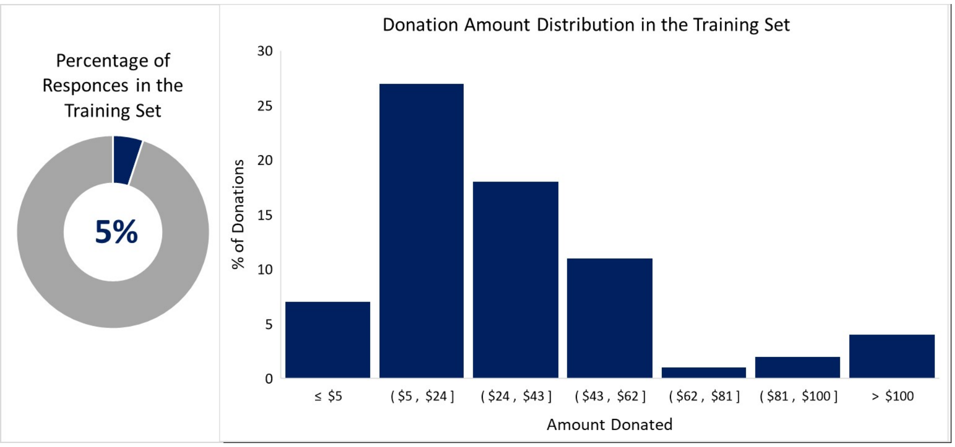

Figure 1: The figure is showing how unbalanced is the response rate to the direct mailing campaign and of those 5% of people who responded the donation amount is depicted on the right as a histogram.

Figure 1 is showing the percentage distribution of the returns of the campaign, where we see the inverse relationship between the probability to donate and the amount donated, making it difficult to build a returns maximization model. Only about 7

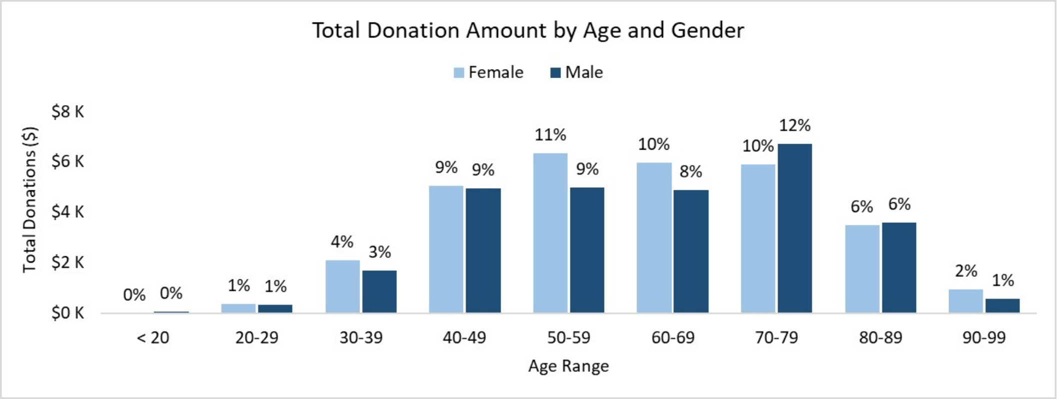

Figure 2: Shows how donation amounts differed with age and gender as part of the exploratory analysis.

Figure 2 shows one of the graphs produced during exploratory analysis so that we can better understand the demographics of contributors. The graph shows that men and women roughly contribute the same amounts (whilst 52% of women mostly in middle age).

3 Preprocessing

The preprocessing step addressed each variable, in turn, looking at missing values, decoding and transforming specific variables so that majority of recorded information can be used in the analysis. The following steps were done for both training and test sets by Microsoft Excel since it allowed us to easily view the data, create lookups, filter and analyse summary pivot tables. This allowed us to explore, plot and transform variables very easily, creating the best possible encoding with respect to the data dictionary. However, if this work needed to be reproduced in production for a similar data set, it is better to develop code (either Python or VBA) that transforms this data up to a certain standard.

At first, categorical variables needed to be standardized and then transformed or encoded. For example, the "TITLE" variable that used 100 different encodings such as '15002 = Commander Mrs.' and by summarising the data, it became apparent that it is a categorical variable. This field was used to create groupings: female, male, in the army, religious figure, government person, a couple or used a neutral title. It enabled to apply standardization, information retainment and instead of creating 100 new variables only 8 were created. Similar groupings were done for "STATE" and "DATE OF BIRTH" variables. The other group of coded variables, such as "DOMAIN" used two levels of encoding to represent both urbanicity level of the donor's neighbourhood and socioeconomic status of the neighbourhood were split into two separate categorical or ordered variables. This required for the understanding of the data representations and variables significance during analysis later.

Secondly, ordered categorical variables were given numerical values, where the value of -1 was assigned to any missing values so that no significance is placed on them. Whereas the missing values of categorical variables were left blank since categorical variables were later transformed by one-hot encoding. The missing values in binary variables were filled with '0' and for quantitative variables were filled with median value, since it is less sensitive to outliers. The missing values were carefully identified for each of the variables.

During the preprocessing step, the number of features grew from 481 to 1,388 due to categorical variables encoding. This highlighted the need for feature selection and dimensionality reduction so that the learning method would be more robust and to avoid fitting to the noise in data. Feature reduction was needed to discard the least important variables and to reduce dimensionality. Firstly variance across all features were calculated,- those features that had variance less than 0.01 were discarded since they were unlikely to influence the target variable. Secondly, looking at the redundant variables by calculating Pearson's correlation between the targets and the arguments, were dropped features with a correlation less ± 0.01, since they were unlikely to influence the target variables and can add noise to the dataset also.

The final clean data set contained 294 features only was split into learning and test samples once again, according to the original data. Then the dataset was shuffled to prevent any inherent ordering that could affect later results.

4 Feature Selection

In order to get the optimal set of variables, that describes the response to the mailing campaign, the feature selection analysis was implemented. At first, it was important to try Principal Component Analysis, however, it was not applicable to this data set, since it contained many binary variables for which it was not possible to compute the variance along with the principal component. In the end, we implemented two types of feature selection methods, one is correlation-based and the other using feature importance property which can be extracted by the Extra Tree Classifier model. Correlation-based model outcomes the interdependency of the inputs on the target either positively or negatively. Tree-based feature selection model also uses the concept of random forest trees, that computes the importance of the features, which in turn helps to remove the irrelevant features and to identify important features for prediction.

Since our dataset is biased towards negative class instances, direct application of the techniques, described above on the entire dataset does not give accurate results. So for non-biased selection of features, the training dataset was randomised initially to prevent any inherent ordering. Then the same number of positive and negative instances were extracted and used in the methods on those congregated instances to get the best possible results. Since the dataset is randomised, it is worth noting that the number of important features differs by ± 10 features (at max) for different runtimes.

Also to ensure that the dataset is balanced with an equal number of negative samples as the positive samples, upsampling of the responder class was used to prevent any bias in the data analysis.

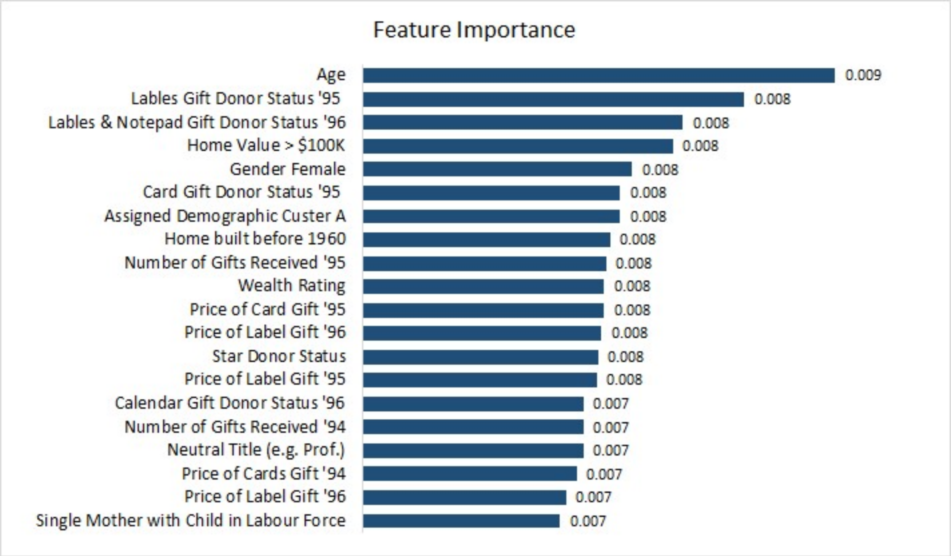

Figure 3: Ranked importance of features according to tree-based feature selection model.

In tree-based feature selection, a different number of features (i.e. 10, 20, 50, 80, 100) were evaluated. As shown in Fig.3 the importance of the 20 input features is ranked for the best model, that describes the response to the direct mailing campaign.

5 UpSampling

This dataset couldn't be used for model training and prediction as the models always show a proclivity towards negative class prediction. So to mitigate this problem, various methods needed to be explored: such as upsampling, downsampling, smote etc. In this project, we had implemented two types of samplings - upsampling and downsampling. Upsampling is a method of increasing the lower class instances to the count of higher class instances by duplicating the already present instances in the dataset. Downsampling method decreases the higher class instances to the count of lower-class instances by deleting some of the instances. In the downsampling method, a dataset might lose the potential information required for training of the model in deletion. Different models were evaluated for the downsampling method, but the performances come out to be mediocre. However, the upsampling method illustrated potential improvement in the models' performance. In our project, as there are a lot of 0 classes, we have increased the count of the 1 class instance with respect to the count of 0 class.

6 Returns Prediction

To optimise the returns made through this campaign it is important to find out who are the donors that contribute the most and what are their identifying demographic factors.

The feature selection was done for output Target_D which is the donation amount made. As we had discussed before, before training the models, preprocessing such as converting the categorical values to numeric, removing unwanted features, selecting important features and finally upsampling of the instances was performed on the dataset. Since we had used two types of feature selection methods, correlation-based and tree-based, the models described below are trained and tested after preprocessing by these two feature selection methods.

6.1 Logistic Regression

Logistic regression was used with 10 values of C parameter between 10^-4 to 10^5. Preliminary the data was split into 75% training set and 25% validation set. Optimal values of C parameters were founded during the analysis by assessing the F1 and accuracy score of the model for 10 values of C. It was found that optimal C parameter is equal to 10^3 when using correlation-based feature selection method and equal to 10^5 when tree-based feature selection is applied. The model performance is listed in the table below. The Logistic regression achieved the lowest test-error rate and highest recall and F1 score when tree-based selection method was applied.

6.2 XGBoost

Optimized distributed gradient boosting classifier also used the same ratio of train-validation split into 75% and 25% samples for model training and testing. The classifier based on XGBoost method was optimised on the two parameters: learning rate and the depth. XGBoost performed better when features are selected by the correlation method at learning rate 0.3 and depth 2. In a nutshell, XGBoost does not perform well for all variants of features selection in comparison with other models.

6.3 Random Forest

Random Forest classifier constructs and fits different tree classifiers on the subsamples of the data and computes the average to control the overfitting and to improve the accuracy and precision of the model. Random forest classifier in this project was optimised on the depth parameter. The model was trained and validated at various depths on the split training and validation set to achieve optimal depth. This classifier achieved optimal performance at depth 2. And during the correlation feature selection method this classifier, this model outperformed all of the models used and achieved the highest performance which can be seen in the ROC curve image.

6.4 Ensemble Model

The ensemble method averaged the results of Logistic Regressing, Random Forest and XGBoost. This combined prediction was produced in the hope that it will be better than any of the individual method prediction since the weaknesses of one method could be balanced out with the strength of others. However, later results show that XGBoost far surpasses the other two methods and therefore the ensemble method actually performs worse.

7 Results

The tables compare the performance results with correlation and tree-based feature selection methods.

Model Performance (correlation based feature selection)	Test-Error Rate	Precision	Recall	F1_score
XGBOOST at learning rate 0.3 and depth 3	60.0	5.17	39.7	9.15
XGBOOST at learning rate 0.3 and depth 2	56.0	5.33	45.6	9.55
Ensemble Model	57.2	5.40	45.0	9.64
Logistic Regression Model	54.7	5.41	48.2	9.73
Random Forest Model	54.0	5.43	49.1	9.78

Model Performance (tree based feature selection)	Test-Error Rate	Precision	Recall	F1_score
XGBOOST at learning rate 0.3 and depth 3	60.49	5.19	39.32	9.17
Ensemble Model	53.67	5.22	47.36	9.40
Random Forest Model	51.12	4.97	47.68	9.01
Logistic Regression Model	47.52	5.64	59.36	10.30

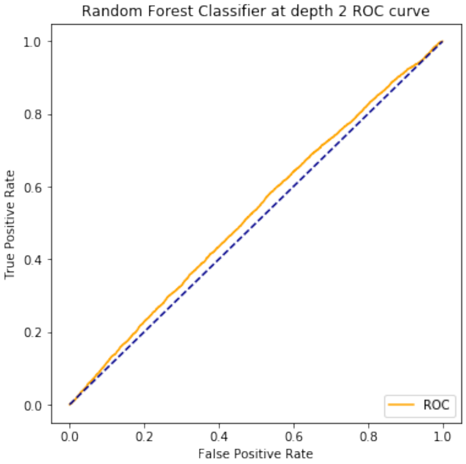

Figure 4: Best correlation-based selection method

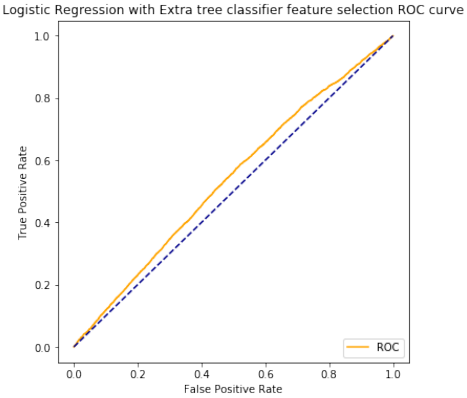

Figure 5: Best tree-based selection method

The following figures 4 and 5 show that although there is an improvement over the baseline for the both correlation and tree-based methods it is very small due to the signal to noise ratio in the dataset.

8 Prediction

Later on, we moved to the job of predicting the profit by scoring the model on the test sample. It was mentioned in the competition that the cost of each email sent is 0.68 dollars. So for calculating the profit 0.68 dollars should be deduced from the amount over 0.68 dollars. Similarly just like before in classification, unwanted columns are removed and important features are selected based on the correlation between the features and the Target_D. By doing this 286 features were selected. And on these data linear regression algorithm was applied and achieved RMSE test error around 4.7. Overall on the Target_D column, 10467 dollars profit was achieved after deducting the email sending cost 0.68 dollars. One of the potential improvements in this task can be done is to apply hierarchical prediction such as classifying the donors based on the logistic regression and apply a linear regression algorithm on the potential donor instances for better prediction

9 Future Work

In the future, it would be interesting to explore other models such as univariate, chi square methods for features selection. Smote upsampling method that perturbs some of the features during upsampling could be implemented and compared with the current results.

10 Individual contribution

Rajasekhar Reddy: I had done the upsampling, feature selction, done classificaton using different models,performed prediction and report preparation.

Anna Ivanhenko: The report initial draft, data preprocessing, variable transformation and discussion.

Indraneel Brahma: exploratory data analysis and ROC curves.

11 References

[1] Kdnuggets.com. 2020. KDD-CUP-98 Results. [online] Available at: <https://www.kdnuggets.com/meetings-past/kdd98/kdd-cup-98-results.htm>l.

[2] Kdd.org. 2020. SIGKDD. [online] Available at: <https://www.kdd.org>.

[3] Li, C. and Ling, C., 1998. Data Mining for Direct Marketing: Problems and Solutions. American Association for Artificial Intelligence.

[4] Zadrozny, B. and Elkan, C., 2001. Learning and Making Decisions When Costs and Probabilities are Both Unknown. Proceedings of the seventh ACM SIGKDD.

[5] Scikit-learn.org. 2020. 1.13. Feature Selection — Scikit-Learn 0.22.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/feature$_selection.html >$.