Speedstar is a vehicle insurance service provider with approximately 20k clients. They did a recent telemarketing campaign which resulted in acquiring 1604 clients from 4000 calls. Based on the current campaign data client wants to develop a predictive model which can help in designing better campaign for next phase. The dataset consists customers’ demographics, specific information about the campaign like call time and previous attempts to acquire the customer.

Insurance company has given the dataset of their marketing results where 3200 datapoints have been selected for the training and 800 for testing which doesn’t contain the predictor variable Insurance. After carefully determining the dataset, we come to know that the dataset have customers’ average yearly balance. Due to average some of the customers have negative values in their balance which suggests that those customers are having more outgoing of money than incoming. This can be a determining factor in whether customers will buy the insurance or not.

As we need to transform the data and then getting the predictive model for the same, we have merged both the datasets after removing Insurance variable from training dataset and storing it into separate variable.

Most of the customers have their job profile of management followed by blue-collar and the least numbers are the customers who are housemaid. If we talk about education of the customers most of the customers have secondary education which holds for almost 50% of entire customer base. Almost 95% of the customers do prefer contact via cellular which is quite understandable as telephones are now on the verge of disappear.

## **Missing value handling:**

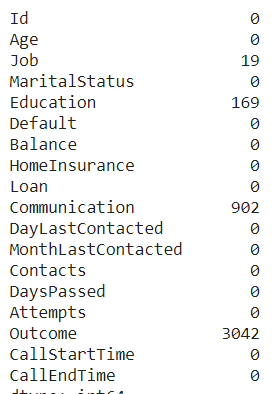
 The first thing comes in data transformation is missing value handling. As we can see in the figure – 1 we do not have missing values in continuous variables. But we do have missing values in categorical variables which are Job, Education, Communication and Outcome.

Figure 1

The outcome variable contains almost 3042 null values which represents 75% of null values. Missing value handling of this variable can lead to a biased value, so we removed this variable.

All the other three variables’ missing values are given the highest frequent values in their represented column. With Job null values replaced with management, Communication with cellular and Education with Secondary.

## **Categorical Transformation:**

Id, Job, MaritalStatus, Education, Communication, MonthLastContacted, CallStartTime and CallEndTime are the categorical variables in the dataset. We removed Id as it will not make any sense in predictive modelling. Then we created another variable called CallTime which is time difference in seconds between CallStartTime and CallEndTime.

MonthLastContacted also removed as there are 12 different values with not much difference. Education and Communication have been treated as ordinal values and transformed accordingly. MaritalStatus and Job after generating different combinations came up as backward difference encoding has the highest accuracy in predicting the target variable.

But we have a problem of scaling as different variables have different value ranges. So transformed the continuous variables based on their z-scores.

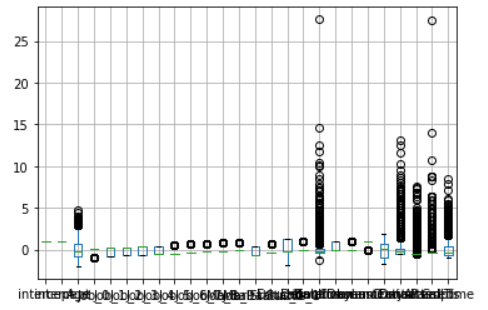


Figure 2

After transformation we now have less outliers than the previous one.

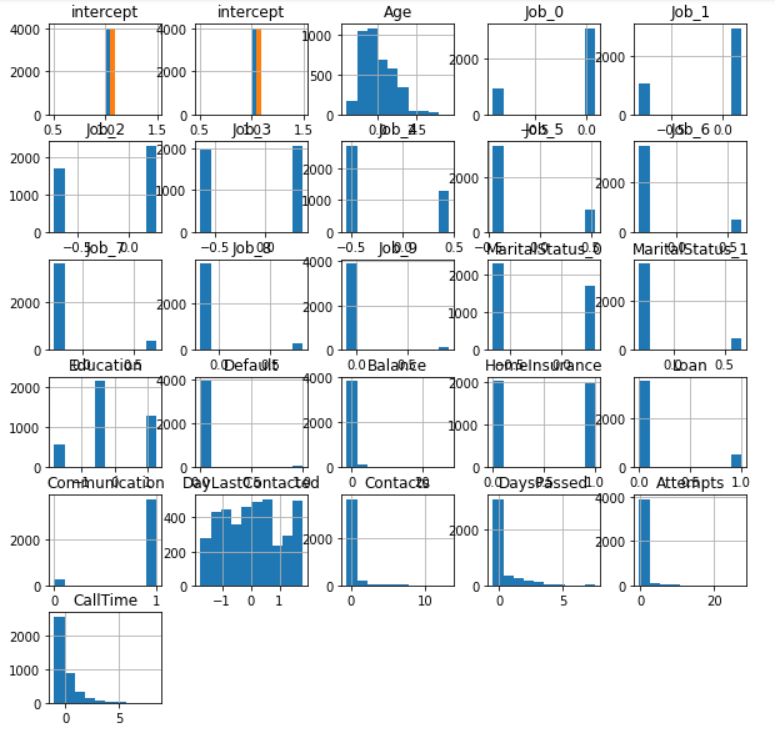


Figure 3

Not only the transformation have removed the outliers but have almost made all the variables normally distributed which is good sign as we can now have unbiased results of predictions.

# **Modelling Outcomes:**

Now, for the modelling we have split the data in 70-30 ratio of training and validation purposes after separating the combined dataset.

We ran six different models:

1. Logistic Regression (for better performance number of iterations have been changed to 10000 from default and solver is set to lbfgs)
2. Neural Network
3. K-fold cross validated logistic regression (with 10 folds)
4. Random Forest
5. Grid Search based parameter tunning of decision tree classifier
6. Gradient boosting

As the problem is of classification, we have selected AUC (Area Under Curve) as our standard measure of identifying the best model. Because the number of customers who have bought the insurance after marketing campaign is 1200 which is almost one third of the total number of customers. The accuracy itself gives biased result as let’s say model predicts all the customers are not buying insurance still, we will get almost 67% accurate model which is not a good sign as model is not a penny good while AUC measures the trade off between true positive rate and false positive rate so we can have better understanding of best model.

After running the model in different settings of transformation and feature selection we get the below AUC scores for validation dataset of all the models:

|  |  |
| --- | --- |
| **Model** | **AUC Score** |
| Logistic Regression | 0.77 |
| Neural Network | 0.78 |
| K-fold cross validated logistic regression | 0.77 |
| Random Forest | 0.73 |
| Grid Search based parameter tunning of decision tree classifier | 0.75 |
| Gradient Boosting | 0.81 |

As we can see in the above table gradient boosting has performed the best among all the models and thus, we selected this model as out final model. After uploading the predictions of testing dataset to Kaggle we got 0.8123 percentage of AUC score.

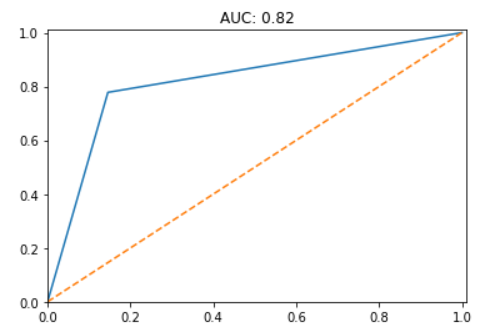


Figure 4 AUC Score of Gradient Boosting Algorithm