MSc in Data Science

SCC460 Group Project Cover Sheet

***All* coursework must have a cover sheet properly filled in.**
Any work not possessing a cover sheet will be returned unmarked.
To be completed by the Student:

Names and Library Card Numbers:	XINJI.WANG 32026312 SHUXING.ZHOU 32071193 JIALU.XU 32049520 ZHENGYU.WANG 32243164
Course Tutor:	YEHIA ELKHATIB
Project name:	Project A: Feature Engineering and Selection
Group name:	Group F
Company involved:	Square, Inc.

I declare that the attached coursework is my own work:
…….ZHENGYU.WANG, XINJI.WANG, SHUXING.ZHOU, JIALU.XU……(signed)

Overview of work carried out by each group member

We shared project work together.
We agree that all group members will have the same mark for this project.

Project A: Feature Engineering and SelectionDetecting Insults in Social Commentary

Group F
XINJI.WANG 32026312
SHUXING.ZHOU 32071193
JIALU.XU 32049520
ZHENGYU.WANG 32243164

1 Introduction

A project of in-bag model accuracy and feature engineering and selection is provided by Square, Inc. The Square, Inc. markets point-of-sale (POS), mobile payments application software and offers multiple financial services including payment management and processing world-wide.

A payment prediction model with high accuracy on Framed will help the Square gain the potential customers and minimize the loss of customers. The prediction of customer payment activities could be produced by machine learning model which obtained from their dataset. Square company could make marketing scheme based on this prediction.

Though the customer payment activities, massive data were collected and two datasets were provided by Square, Inc. Events and usage of Framed form the incidence dataset, which each row is an observation (a user), and each column is a feature. The customers behaviour whether they purchased a paid plan on the Squares platform is stored in the response dataset.

Our project is to investigate a high dimensional dataset to find the valid features and apply these features into Support-vector-machine (SVM), Neural Network and Random Forest models to predict whether the customers will purchase a paid plan. We aim to analysis whether the features selected are enough to optimize model accuracy. To achieve this goal, we followed a standard pipeline, which contains data exploration, data prepossessing, modeling and feature selection and feature validation evaluation.

Owing to the poor quality of dataset, which has no column descriptions at all, we could only do basic analysis of the dataset and then chose the corresponding preprocess methods and analysis technique. After data preprocessing, random forests will be used to select important features between the severe imbalanced original-dataset and balanced dataset produced from the Synthetic Minority Over-sampling Technique (SMOTE) algorithm. We obtained two sets of significant features. Then we trained the several models with the dataset only containing these non-trivial features, and use the criteria including accuracy, precision, recall and F1 to evaluate the model performance. In this way, we figure out that our prediction models and dimension-reduced dataset are authentic. Hence, by using the sets of machine learning models and technique above, we could measure the trend of customers payment (which is the observation predicted to be 1 but actually 0).

2 Methodology

The project methodology could be described by the following subsections.

2.1 Data exploration

The data exploration is aimed at finding information of the dataset and providing the clue of suitable model and research method in future analysis. Our dataset is provided with 100279 observations and 1818 anonymous attributes. Besides, the raw dataset has little validity with various random errors, misclassification and missing data. Simultaneously, some constant columns were found in the dataset. We examined whether the general model and the traditional machine learning model could provide an accurate prediction on customers behavior using the raw dataset which obtained from the Square Company. Facing a binary classification problem, we used SVM, Neural network and random forest classifiers to test on the split data. The performance of the models will then

be compared.

2.2 Data prepossessing

The data contains two CSV files, the response dataset and the incidence one. We cleaned the dataset in three aspects as followed:

- The missing value will lead to considerable influences on the classification. As the result of detection, no missing value is founded. In other words, no gapes should be filled in in the dataset.
- The description of each column are removed. In this case, columns with the same number should be removed as they provide no different information between observations. there is no more detailed information in these constant columns. It reduced the dimension of incidence dataset from 1818 to 1393 columns.
- We deleted the duplicated data.The majority class is about 400 times of minority class. After deleting the repeated observations, the ratio drops to 200:1 level and we get the imbalanced dataset A (1393*48157).
- To unravel the imbalance between two classes, SMOTE algorithm is chosen to eliminate the problem of imbalance. SMOTE is a common method of oversampling to deal with imbalanced data. After this process we get dataset B (1393*14869) which is balanced.

After data cleaning process, we derived two datasets, the imbalanced dataset A and balanced dataset B.

2.3 Feature selection and modeling

Random forest is an ensemble of unpruned regression or classification trees. All the trees in the random forest are created by using bootstrap samples of the training data and random feature selection in tree induction. With numerous decision trees working together, a random forest could prevent overfitting compared to one single classification tree.

Our intention is to select the top 50 important features, which could composite as an ideal representation of the original dataset. To achieve this goal, we build several random forests on the imbalanced dataset A and the balanced dataset B.

Despite that deeper random forests have better performance in classification, shallow random forests with only 3-4 nodes but vast number trees could also figure out the important features which are truly useful in separating two class of customers. Constrained by computation budget, the enormous random forest was split into several smaller random forests. By weighting and summing up smaller random forests, we finally obtain the rank of features according to the importance.

We use node purity to evaluate the importance of each feature. A node purity importance is an evaluation criterion of a certain feature. Each time a tree in a forest is built, a node in the tree is replaced by a randomly selected feature. This changes the Gini index of the tree. If the feature is useful in classifying two class of observations, the change of Gini index will be massive.

A Gini index in a single tree is calculated as followed:

$$G = \sum_{i=1}^{n_c} p_i(1 - p_i)$$

p_i represent the probability of one class separated by feature V_i.
Given an impurity function $i(t)$, we split at some node if the change of GINI index is significant, where t_L is the node on the left, and t_R is the node on the right.

$$\Delta i(t) = i(t) - \frac{N_L}{N} i(t_L) - \frac{N_R}{N} i(t_R)$$

Summing up all the result from random forest models, we analysis the significance of features using the following algorithm and plot the top 50 importance feature in the plot. The feature value function is described by:

$$Y = \sum_{i}^{100} (100 - X_i)$$

In this case, Y stands for feature importance. X_i represents the rank of a certain feature among top 100 important features in each random forest (if X feature is not in 100 then counts it as 100). Top 100 features are selected in order to figure out top 50. For example, in a single forest, V995 ranks one. Then we

assign 99 scores(100-rank) to this feature. The total value Y of V995 sums up all the result of random forests.

Besides getting purity of each feature, it is possible to evaluate the importance of some variableX_k. If $i.$ is Gini index, then $I.$ is called Mean Decrease Gini function.

$$I(X_k) = \frac{1}{M} \sum_m \sum_t \frac{N_t}{N} \Delta i(t)$$

We extracted features using the method above from both our imbalanced dataset and the balanced dataset derived from SMOTE algorithm.

Part of the result was presented in figure one and figure two.

Figure 1: Node Purity

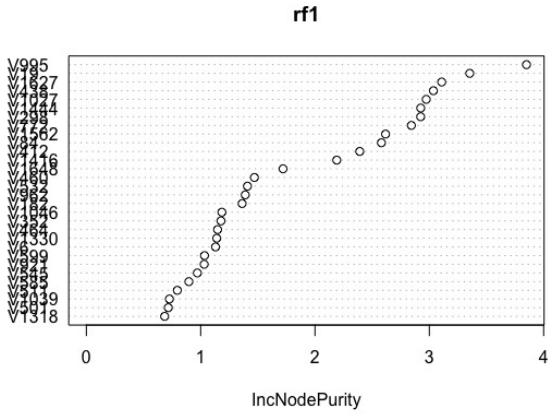


Figure 2: Valued Feature Importance

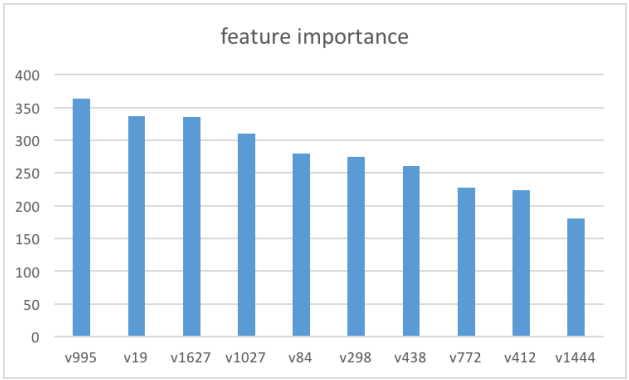


Figure one showed node purity from one of the random forests, the more important a feature is, the bigger the node purity would be. Figure 2 indicated the valued feature importance from several random forests.

2.4 Feature validation evaluation

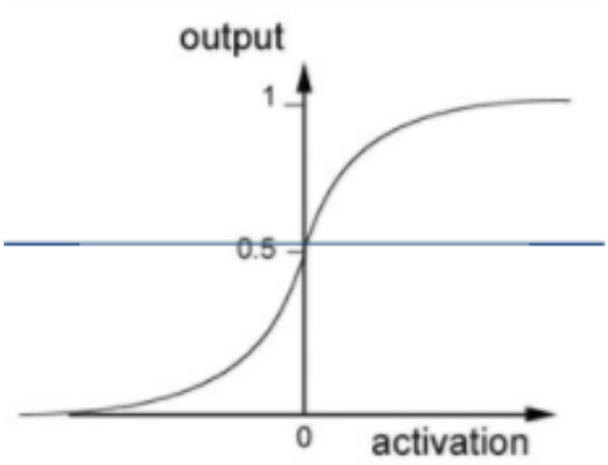
In this subsection, we will discuss how features are tested to be valid. A good representation of raw data aggregated by valid features could provide most of the information we require to predict a customer's behaviour. In another word, if we train effective prediction models using dataset only containing the top-50 features, the features selected are proved to be valid.

Generally, after selecting the top-50 important features, we trained and tested our features using machine learning models including SVM, neural network and k-nearest neighbour classifier. To compare the performance of each model, we used performance metrics. Meanwhile, using this sets of models, we got the estimation of a customers probability of consuming (in some of the models like the neural network). Cross-validation was applied to training stage to increase the robustness of the prediction. Finally, a identification of observation values (0 or 1) is carried out by multiple models. The value should be re-assigned by following majority acceptances. For example, if an observation was predicted to be 1 by the majority of the models, the new observation would be considered as 1.

To implement the above steps, the normalized features is required. we could use standardized function to obtain the normalized features: $z = \frac{x-\mu}{\sigma}$ μ stands for the mean of features value x and σ represents the standard deviation of x. Besides, different value ranges in features normalizing would not influence the accuracy but decrease the runtime of the models. In the neural network model, a five layers neural network is built by using a different combination of activate function. Using the sigmoid function(figure 3) in the final layer, which outputs a prediction between 0 and 1. During modeling, all the models are trained by a balanced dataset which derived from SMOTE algorithm, then tested by a cleaned-imbalanced dataset.

The result of neural network turns out to be accumulated around 0.5. Thus, by setting up a probability boundary p_i, which means predictions over p_i would be re-assigned to 1.

Figure 3: Sigmoid



We could adjust the model as time goes and situation change.

In the k-nearest neighbor (KNN) classifier, an object is assigned to the KNN by a majority of model identification. By adjusting the number of neighbors, we managed to found the optimal number of neighbors.

SVM is a supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis. By changing the kernel function and other parameters,the performance and fitness of SVM model is improved.

We use the balanced and imbalanced datasets to train SVM, KNN, neural network and then used the imbalanced dataset to test our models. As the majority of test data are 0 (might be no purchase behavior), the accuracy metric is no longer appropriate for the test since it is reasonable that we predict 0 as 1(predict potential customers as ones with purchase behavior). By contrast, the evaluation of precision give more guidance on exploring the potential customers with the prediction of 1.

In this case, we use the precision, recall and F1 scores calculated by the following formulas:

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1scores = \frac{2TP}{2TP+FP+FN}$$

Where TP, FP and FN stand for true positive, false positive and false negative separately.

3 Result

We have selected top-50 features from both the imbalanced dataset and the balanced dataset using random forest. Here we compare the difference between them.

3.1 Feature selection comparison

First, we compare whether there are any difference between this two sets of features.

Figure 4: Feature Importance

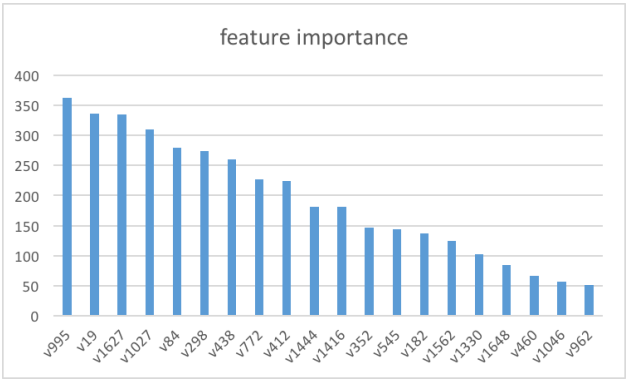
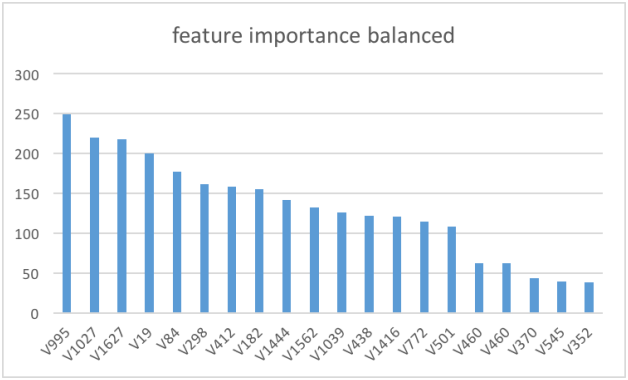


Figure 5: Feature Importance balanced



The first chart is the features extracted from the imbalanced dataset while the second one comes from the balanced dataset. We can see the difference is relatively small. Only several rankings of the features are different and most of the features reserve.

3.2 Model performance comparison

First, we used cleaned dataset(1393*14869) to train the SVM, neural network and KNN model and use the value of precision, recall, F1-score and accuracy as reference.

That means a prediction model works well

when its performance metrics get close to the these four values as followed:

As mentioned above, by comparing model performance metrics between two sets of features selected in the imbalanced and balanced dataset.Compare the result from the cleaned data.

The table below was present as the performance and will be compared. The features

Figure 6: dataset with 1393 features

models	train dataset with 1393 features		
	SVM	NN	KNN
Precision	92.75%	81.30%	99.00%
recall	94.92%	0.41%	45.00%
F1-score	93.82%	0.80%	62.00%
accuracy	94.50%	95.40%	93.45%

from the balanced dataset perform slightly better in all of the models we built. Although the accuracy of neural network drops significantly from 95.40% to 81.12%, the other performance metrics like recall and f1-score improve as we expected.

Secondly we compare the performance between two sets of features. Apart from neural network, the other two models performed better in the metrics we concerned about. Hence we conclude that the features we select is a good representation of the dataset. Therefore the selected models are able to identify customers purchasing activity.

Figure 7: Comparation between two sets of features

models	feature from imbalanced dataset(50)			feature from balanced dataset(50)		
	SVM	NN	KNN	SVM	NN	KNN
Precision	96.79%	82.00%	95.00%	99.60%	99.00%	100.00%
recall	98.52%	81.00%	96.00%	99.20%	42.80%	97.00%
F1-score	97.65%	84.00%	94.00%	99.40%	59.90%	98.00%
accuracy	97.80%	81.12%	91.19%	30.20%	70.10%	98.50%

3.3 Further exploration

To further verify our conclusion, we analyze some of the features we selected. So the most important feature in two datasets is V995, which gets the highest score in two purity importance table. The table clearly shows a pattern that while values in feature V995 is not 0, the customer will pay plan on the platform. So as soon as we detect this behavior, a high chance of customer being labeled 1 can be

Figure 8: Summary of V995

Level of V995		0	1	2	3	6
RESPONSE	0	47920	0	0	0	0
	1	70	147	15	3	1

assured.

We are showing other results in the following table, they all follows a similar pattern,which prove the validation of our results.

Figure 9: Summary of V1627

Level of V1627		0	1	2	3	4	6
response	0	47920	0	0	0	0	0
	1	102	111	16	5	1	1

Figure 10: Summary of V1027

Level of V1027		0	1	2	3	4	6	7
response	0	47882	25	10	1	2	0	0
	123	32	42	17	10	8	3	1

Figure 11: Summary of V84

Level of V84		0	1	2	3	4
response	0	47916	4	0	0	0
	1	106	67	48	13	2

3.4 A brief summary to the answers of our original questions:

The first objective of our project is to feature engineering in high dimension dataset. For feature selection, we use the random forest to select top-50 importance features and test its validation by using sets of models. The second objective of our project is to test whether traditional feature selection heuristics are enough to optimize model accuracy. Finally, by testing imbalanced dataset which is more close to the raw dataset, we can generalize the models to reality and get confirmation that the Square, Inc. gain more financial profits from our model.

3.5 Bias and Validity:

We follow a validity principle to organize our research. Firstly, we use cross-evaluation in training stage of modeling to improve the model performance. Secondly, the imbalanced

dataset which is close to the raw dataset is used to test our model, from which we conclude that our model is able to generalize to reality. Thirdly, the construct of feature validation evaluation follows control variable method and is authentic.

It deserves to discuss the bias in our project. For the data cleaning, we did not remove the misclassification, random error as we do not know column description. For the modeling process, the parameters like activation function in neural network and N cluster in KNN we choose to apply in the machine learning model is empirical and will produce inevitable bias in classification.

4 Conclusion

The purpose of this project is selecting the important features which can be engineered. In addition, we use traditional feature selection heuristics to check if it is enough to optimize the model accuracy. In this project, we use the methods of data exploration and preprocessing to clean the dataset. After that, we used data science methods such as random forest, SVM, neural network to deal with a real dataset. We have confirmed that the features of the random forest are the good representation of the original Framed data and our prediction model is valid and robust in predicting customers purchase behaviour. The outcome of this project shows that the features we selected from original dataset are significant since the accuracy of models in our project is relatively high. These findings of the project are prospective and can be used to predict the potential customers.

However, there were some limitations in our project.In the stage of feature selection, we did not consider the process of dimension reduction technique like LDA and PCA which will keep more information from the original data. In the stage of modeling, we have not considered other classifiers like naive bayes which also perform well on binary classification.In the same time, there is no standard while we are tuning parameters of our model. So we do not know whether we have already constructed the best model. Moreover, we could make an assumption that the top-25 or even fewer features could perform well in predictions. It could be quite interesting to find out the differences between top-25 and top-50 features. Logistic regression is not appropriate to model our balanced or imbalanced datasets because this algorithm is not aggregated since some estimated variables are not existing. We could change several parameters and make features standardized to solve this problem.

References

[1] Chen, C., Liaw, A., Breiman, L. (2004). *Using Random Forest to Learn Imbalanced Data.* (Available online from :http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf)

[2] Grave, E., Ghaoui, L. (2014). *Fast imbalanced binary classification: a moment-based approach.* (Available online from: https://hal.inria.fr/hal-01087452/document)

[3] Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. (2002). *"SMOTE: Synthetic minority over-sampling technique."*. Journal of Artificial Intelligence Research, vol. 16,pp. 321–357.

[4] Svetnik, V., Liaw, A., Tong, C.et al. (2003). *Random forest: a classification and regression tool for compound classification and QSAR modeling[J].* Journal of chemical information 43(6): 1947-1958.

[5] Pardoe, I. (2006). *Data Mining Techniques Chapter 6: Decision Trees.* (Available online from: http://www.iainpardoe.com/teaching/dsc433/handouts/chapter6h.pdf)