Will this client subscribe a term deposits?

UCR CS 235 Project Report, Spring 2019

Manli Ran
Computer Science
University of California, Riverside
Riverside, CA, USA
manli.ran@email.ucr.edu

Hsin-Ta Li
Computer Science
University of California, Riverside
Riverside, CA, USA
hsinta.li@email.ucr.edu

Dandi Wang
Computer Science
University of California, Riverside
Riverside, CA, USA
dandi.wang.doris@gmail.com

ABSTRACT

Our project aims to generate a model that helps evaluate the probability that a client will subscribe to a term deposit. For this binary classification problem, we use k nearest neighbors, logistic regression and neural networks to generate predicting models. In addition, we apply principal components analysis and forward selection to select a feature set to find the most representative features and reduce feature dimension. To handle the imbalance problem of our raw dataset, we use oversampling to increase the portion of positive entries. Finally, we compare the performance of each classifier by accuracy, precision and recall. Among the three methods, neural networks have the best performance in all cases in terms of recall, the key metrics for evaluation in our project. We found that oversampling greatly improve the results and forward selection also helps optimize the classifier.

CCS CONCEPTS

• Information systems • Information systems applications • Data mining

KEYWORDS

Data mining, PCA, kNN, Logistic Regression, Neural Networks

ACM Reference format:

Manli Ran, Hsin-Ta Li and Dandi Wang. 2019. Will this client subscribe a term deposit? UCR Data Mining Techniques *(UCRDMT'19). Riverside, CA, USA, 8 pages.*

1 Introduction

Data mining is applied to multi-industries; such as analyze business information, identify, analyze and prevent crises improve customer relationship management and so on. A term deposit is an investment that offers a higher rate of interest compared to normal saving but in a fixed term customer may not freely withdraw money from their account. In our project, based on a bank customer dataset, we try to figure out the features that related to the decision of subscribing to a term deposit most and how to predict if a customer will subscribe a term deposit or not. If we know how much possible the person will become a client, we can decide how much effort the bank worker should spend on people trying to convince them, which will be more efficient.

Our project is generally divided into three parts. First, for data pre-processing, we focus on doing feature reduction and selection to eliminate those features are not highly correlated to the subscription because the dataset contains up to 20 features. Next, since the dataset provides the results of subscription, our work is supervised learning, and the prediction can be viewed as a binary classification problem, we plan to apply k nearest neighbors, logistic regression and neural networks to generate predicting models. Finally, we will evaluate and compare those models in terms of accuracy. After our analysis, we want to find out a better way to predict the subscription of term deposit.

2 Related Work

2.1 Data Preprocessing

Our goal is to use data mining methods for bank business. Refer to [1] as the most important method. CRISP-DM is a great method because it puts business understanding front and center at the beginning of the project. Although there are 2 data mining programs working on the same dataset, the result of business understanding might be absolutely different from each other, one single method can't solve it. but this is also the big problem within pre-processing. Because business understanding is different from person to person. In this paper, authors delete the incomplete data which are valuable, we choose to fill the missing data with a label and analyze the rest. If the missing characteristic data actually has little effect on the result, then the sample will be useful.

There are plenty of methods for data pre-processing, it is detailed and We learned new method from related papers such as FP-Growth algorithms. [2] did a great job in summarized methods in data reduction part, which includes Data reduction strategies, Aggregation, Sampling, Dimensionality Reduction, Feature subset selection, Feature creation, Discretization and Binarization, Attribute Transformation. However, this paper still has some problems. First of all, it just uses 163 records to train the data, it is possible that the result from Apriori and FP-Growth algorithm looks same. Secondly, PCA is a very important method for data reduction, the paper would be more complete if PCA is included. Third, it would be great if author can create a table for comparing different methods after listing method. If working on the same data is hard, then list the conditions which method applied best, but beside above, this paper gives us a lot of help,

2.2 Classification Techniques

This paper [3] is the one that the dataset our project going to use comes from. It introduces using data mining techniques to form the decision support systems that helps to predict if a client would subscribe the long-term deposits. The raw dataset has 150 potential features, so the preprocessing focuses on the feature reduction. This paper conducted semi-automatic approach, first it filtered relevant features according the experience of bank campaign manager, then it applied forward selection to discard more features. This is a binary classification problem and the paper chooses several models including logistic regression (LR), decision trees (DTs), neural network (NN), support vector machine (SVM), to test and compare their performances. For complex black-box models, NN and SVM, they also use rule extraction (with DTs) and sensitivity analysis to better illustrate the relationship between features and outputs. To evaluate the performances, this paper uses two criteria, the area under the receiver operating characteristic curve (AUC) and the area under the Lift cumulative curve (ALIFT). Since the outputs of those models is a probability, a threshold is needed for deciding whether assign true or false for it. AUC is to add up the accuracy for models using different thresholds, while ALIFT is to add up the accuracy for models obtained with different size of data. This paper implemented those models with the rminer package of the R tool. After preprocessing, 22 features were selected. Among the four techniques, NN performed the best, with the highest AUC as well as ALIFT. After sensitivity analysis and rule extraction, it turned out that the three-month Euribor rate is the most relevant attribute.

Credit risk evaluation system with supervised neural network [4]. The dataset has 1000 cases with 24 features. The neural network generates complex nonlinear models that fits the data nicely, but it usually needs considerable computation power that grows over the complexity of the models. Cross validation is often used to find out the appropriate design of neural networks models, and since the dataset is divided into training set and validation set, a higher proportion of training set also implies higher computation cost. To find the trade-off between accuracy and computation cost, this paper tries three neural network models and nine learning schemes. Those three artificial neural network (ANN) have 18, 23, 27 neurons respectively in the only hidden layer, labeled as ANN-1, ANN-2, and ANN-3. And the different learning schemes differ in the ratio of the size of training to validation dataset, specifically from 0.1 to 0.9. First, for those four non-categorical features, this paper normalized them using the maximum. Then applied the three neural networks models under different learning schemes. Here they used 0.5 as the threshold. First, for those four non-categorical features, this paper normalized them using the maximum. Then applied the three neural networks models under different learning schemes. Here they used 0.5 as the threshold. They compare performances by run time and accuracy rates, also recorded the change of error over the increase of iterations. According to this paper, ANN-2 under the learning scheme with 40% for training and 60% for validation works best. It got considerable accuracy within acceptable short processing time.

2.3 Previous Works on Bank Marketing Data Set

Classification techniques such as Naïve Bayes (NB), Decision Trees (DT), Artificial Neural Network (ANN) and Support Vector Machines (SVM) have been widely applied

to the Portuguese bank database. *Koumétio et al*. [5] proposed a technique to pre-process different type of features, then applied 1-Nearest Neighbor to predict clients. Based on their experiments, they claimed that the proposed technique can prove stable and accurate regardless of feature normalization. Their results showed that the proposed method has the best performance in terms of f-measure comparing to NB, DT, ANN and SVM. However, the authors asserted that the proposed method is better than ANN since it consumed the most execution time. In fact, the performance of ANN is better than the proposed method. As the number of data increases, the cost to compute their nearest neighbor based method (or similarity) will increase too. As a result, the proposed technique will consume the most execution time instead of ANN. In sum, this paper provided a method to transfer the features which may also be used in our project.

Elsalamony [6] also evaluated and applied several important classification methods such as Multilayer Perceptron Neural Network (MLPNN), Tree Augmented Naïve Bayes (TAN) known as Bayesian Networks, Logistic Regression (LR) and Ross Quinlan new decision tree model (C5.0) to the Portuguese Bank Marketing Data Set. The aim of his work is to compare the performance of these methods by three statistical measures; classification accuracy, sensitivity, and specificity. The experimental results showed that C5.0 can achieve slightly better performance than MLPNN, LR and TAN. In addition, the importance analysis showed that the most significant attribute in C5.0, LR, and MLPNN is "Duration" whereas the attribute "Age" is more important than the other attributes in TAN. The author demonstrated a significant analysis of attributes to find out the most important attribute in each classification technique. However, the author did not give an explanation of the reason why the attribute is the most important. Moreover, it would be better if the author also analyzed the order of importance for all attributes and provided explanations. In sum, this paper inspires us to perform a similar analysis for the importance of attributes.

3 Data Preprocessing

We got the Bank Marketing Data Set from the UCI Machine Learning Repository. It has 20 features and one binary output. For all the 41188 entries, we divided it into a training dataset of 37070 entries and a test dataset of 4118 entries. Its features and output are shown in Table 1.

Input variable
1 - age: numeric
2 - job: categorical
3 - marital: categorical
4 - education: categorical
5 - default: categorical (has credit in default?)
6 - housing: categorical (has housing loan?)
7 - loan: categorical (has personal loan?)
8 - contact: categorical (contact communication type)
9 - month: categorical (last contact month of year)
10 - day_of_week: categorical (last contact day of the week)
11 - duration: numeric (last contact duration)
12 - campaign: numeric (number of contacts during this campaign)
13 - pdays: numeric (number of days since last contacted)
14 - previous: numeric (number of contacts performed before this campaign)
15 - poutcome: categorical (outcome of the previous marketing campaign)
16 - emp.var.rate: numeric (employment variation rate)
17 - cons.price.idx: numeric (consumer price index)
18 - cons.conf.idx: numeric (consumer confidence index - monthly indicator)
19 - euribor3m: numeric (euribor 3 month rate - daily indicator)
20 - nr.employed: numeric (number of employees - quarterly indicator)
Output variable (desired target):
21 - y - binary (has the client subscribed a term deposit)

Table 1: Dataset Information

3.1 Noise Data Cleaning

We have plenty of noise data from our raw dataset, which is normal for real world data. The noise data contains missing data, unknown data and so on. For missing data records, we decided to delete those samples, because except those data, we still have more than 400,000 sample. And for unknown data, we chose to set label for them, so that we can continue calculating the dataset without worrying about them. For next step we transfer the string to numerical and then classify them. Last but not least, in order to have more accurate result, we normalize the dataset.

3.2 Principal Component Analysis & K-means

Because the futures of raw data are too much, we found out that some features have little power with the result, so we want to mine the data through the PCA method to remove those non-using features. After PCA analysis using Matlab, please see the corresponding accuracy with Eigenvalues in Table 2. We get that more than 90% information could be represented by only 2 vectors, you can see the detail in Figure 1.

Index	Eigenvalue
1	67400.0
2	35600.0
3	4399.0
4	109.00
5	20.900
6	12.900
7	7.490
8	4.620
9	3.540
10	1.990
11	0.758
12	0.586
13	0.333
14	0.309
15	0.201
16	0.167
17	0.142
18	0.029
19	0.023
20	0.019

Table 2: Corresponding accuracy with Eigenvalues

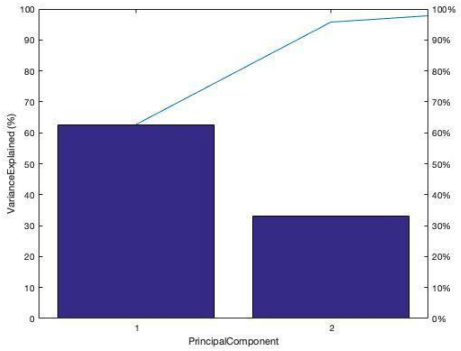

Figure 1: Value and accuracy of first two Eigenvectors

Thus, we use K-means to reduce the dataset. In order to have more accuracy, we set K as 10 and choose one feature from each class. As a result, we reduce 20 features to 10 features at last.The results are shown in Table 3.

Feature Index	Assigned Class by k-means
1	2
2	8
3	10
4	10
5	10
6	2
7	8
8	7
9	8
10	9
11	2
12	5
13	7
14	1
15	6
16	6
17	4
18	3
19	4
20	2

Table 3: Classify result of 20 features

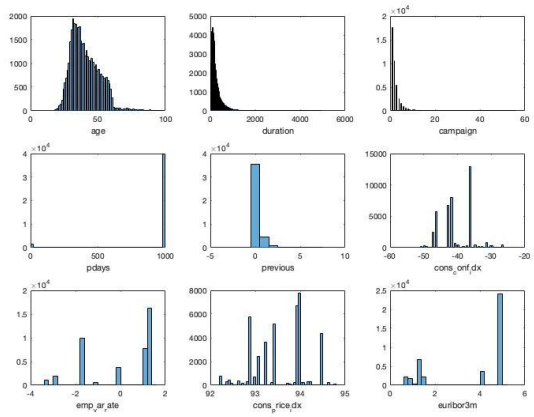

Figure 2: Histogram

We can see from Figure 2 the histogram, although some of the features have a great different results based on different value, some are not, combine with the classification we did in section 2.2, we can easily understand how PCA and k-means methods work.

3.3 Forward Selection

The aim of feature selection is to select the most correlative features from the original feature set to obtain a better performance of classification tasks. In order to solve the feature selection problem, we implement forward selection in this project. The forward selection is a heuristic method for feature selection. It begins with an empty selected set, then adds feature one by one. In each forward step, the feature with the best recall in the testing dataset will be added to the selected set. This selection process will continue until all features are selected, then return the feature set with the best performance. In this project, we will apply forward selection on the oversampled data classified by ridge logistic regression model. The selected features are [1, 2, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16, 19, 20].

3.4 Oversampling

While inspecting our dataset, we find that it is seriously imbalanced. There are only about 10% positive entries in the original dataset. This imbalance problem can mislead those training model to predict most entries to be negative so as to minimize the square-error loss. However, especially in our case, we tend to find more positive entry and achieve higher true positive. Therefore, according to [9], we found oversampling aims at solving the imbalance problem. In our case, we simply copied those positive entries over and over again and added them back to the original dataset. In this

way, we got an oversampling dataset with 66183 entries and over 50% of them are positive.

4 Classification Technique

In order to run the classification tasks, we should construct a classifier that can classify unseen data (testing set) based on a given dataset with labeled classes (training set). Here, we briefly introduce the classifiers used in this project.

4.1 k-Nearest Neighbors

The k-nearest neighbor (k-NN) algorithm is a popular classification technique because of its simple and intuition. The k-NN algorithm classifies the test sample vector ***x*** to one of the classes according to determine the k nearest training data to the sample ***x***, using a distance metric. Afterward assign ***x*** to the class with the most votes within the set of k nearest training data. The only pre-requiring setting of k-NN algorithm are the number of neighbors, k and the distance metric.

Assume that x_i is the *i*th training data vector, and z_i is the corresponding class indicator, where $i = 1, \ldots, n$. That is, $z_i = j$ if the *i*th training data is a sample from class ω_j. Let the distance between ***x*** and ***y*** be denoted as $d(x,y) = |x - y|$. In this study, the Euclidean distance was used because it is common to use as the distance metric.

The implementation of the k-NN algorithm applied to a test sample vector ***x*** firstly calculates the distance $\delta_i = d(x_i, x), for\ i = 1, \ldots, n$. Let $\{a_1, \ldots, a_k\}$ represent the k smallest values of δ_i. Define k_j to be the number of these k nearest training data whose class is ω_j, and $I(za_i = j) \in \{0, 1\}$ indicates whether a_i belongs to class ω_j. Thus, $k_j = \sum_{i=1}^{k} I(za_i = j)$ can be obtained. The k-nearest neighbor decision rule will assign the test sample ***x*** into class ω_m if $k_m \geq k_j\ for\ all\ j$. In this algorithm, the unlabeled test sample is classified to the class that labeled most frequently among the k nearest training samples.

4.2 Logistic Regression

The Logistic Regression (LR) classifier is well suited to solve a variety of data sets. In addition, it is well suited to describe and test assumptions about the relationship between classification result variables. The LR uses maximum probability estimation rather than the least squares estimation. In this project, we use L2 regularization to LR, also called ridge Logistic Regression. The loss function of LR is:

$$L = \left[\sum_{i=1}^{m} \ln(1 + e^{-y_i x_i^\top w})\right] + \frac{\lambda}{2}\sum_{j=2}^{n} w_j^2$$

In order to optimize this function, we would use both Newton's method and gradient descent to train LR model. However, Newton's method is relatively straight-forward, except when it doesn't work. For that reason, we need to check the value of the loss function after each iteration. Each time the Newton step fails, we should try to take a gradient step. As soon as a step succeeds, we should return to Newton steps. Repeat to do this until convergence.

4.3 Neural Networks

Neural Networks (NN) is a data mining model that imitates the structure of the human neural system. It is a model consists of one input layer, one output layer, and multiple hidden layers. Each layer corresponding to an activation function, and there are weights applied to neurons between two layers. Commonly used activation functions include sigmoid, tanh, rectified linear unit (relu), softmax and so on. Those activation functions can introduce non-linearity to the model, which helps to fit complicated dataset better. But this non-linearity also prohibits us to directly train NN, same as logistic regression, gradient descent is used to tune weight iteratively. Each iteration can be divided into two steps, feedforward, and backpropagation. Feedforwarding computes the output with the present model and the loss. After that, backpropagation updates the weights based on the derivative of the loss function and an assigned learning rate. Beyond always keep using the same learning rate for every weights factor, some optimizer functions can be applied to choose learning rate more wisely, such as stochastic, Adagrad, Adadelta, adam [7] and so on. To build the NN, we use Keras library [8]. Here in our experiments, we define two hidden layers, first one with relu as activation function and another one with sigmoid. We also use adam optimizer, which set a different learning rate for each weights factor. This design is recommended in the Keras documentation for binary classifications. We also tried several other designs but the results are similar.

5 Experimental Results and Evaluation

5.1 Evaluation Measurements

The performance of each classification model is evaluated using three statistical measures; classification accuracy, precision and recall. We use true positive (TP), true

negative (TN), false positive (FP) and false negative (FN). True positive (TP) is the number of correct predictions for which the instance is true. True negative (TN) presents the number of correct predictions for which the instance is false. These measures can be calculated from a confusion matrix. The False Positive (FP) is the number of incorrect predictions that an instance is true. Finally, False Negative (FN) is the number of incorrect predictions that an instance is false. Table 4 shows the confusion matrix in binary classification.

		Predicted Value	
		Negative (no)	Positive (yes)
Truth Value	Negative (no)	TN	FP
	Positive (yes)	FN	TP

Table 4: Confusion Matrix example

The following equations defined the measurements used to evaluate the performance in this project:

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP}$$

$$Precision@no = \frac{TN}{TN+FN}$$

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

$$Recall@no = \frac{TN}{TN+FP}$$

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

Since this project aims to provide a prediction of if the client will subscribe a term deposit hence bank can be more efficient while trying to market the term deposit. It is more important to have more positive entries be predicted correctly rather than having a higher accuracy. Therefore, in our evaluation, the recall(yes) is considered to be the key measurement of the performance of methods.

5.2 Classification Results

5.2.1 Analysis of k Nearest Neighbors

In order to pick k for our kNN model, we applied kNN on the oversampling data with all features. The results as shown in Figure 3 demonstrated what the value k we should pick. Although k=1 will achieve the highest accuracy in the testing data, we should not choose k=1. First, if we choose k=1, the kNN model may become overfitting since it only observes one neighbor. Second, since we applied the kNN model on the oversampled dataset, if we choose k=1, this cannot reflect the benefit of oversampling. As a result, oversampling will become useless. Based on these reasons, we decide to choose k=17 with the testing accuracy 83.37 % for our kNN model.

5.2.2 Analysis of Logistic Regression

The regularized term λ decides how the LR model complex. The larger the λ is, the simpler the LR model is. Therefore, we should also decide the value of λ, this is similar to k value in kNN. The following Figure 4 shows the testing accuracy versus different λ in ridge LR model. Based on the results, we set the value of λ to 0.76 with the testing accuracy 86.67 % for our LR model.

5.2.3 Analysis of Confusion Matrix

Figure 5, 6 and 7 shows the confusion matrix for kNN, LR, and NN. From them we can find out that all three methods have high TN and low FP, but LR and NN works much better than kNN in terms of FN and TP. Considering that our goal is to predict if the client are willing to subscribe a term deposit hence strengthen marketing, we prefer those methods that results in higher TP and lower FN. Therefore we would say that NN works better than LR though NN results in a lower TN.

5.2.4 Statistics Measurement Comparison

Table 5 shows the statistics measurement for each method working with original dataset and oversampling dataset. We can see that all methods have considerable accuracy, as well as precision(no) and recall(no) especially for those without oversampling due to the extremely high portion of negative entries in original dataset. The highest accuracy achieved by NN in original dataset. Although after oversampling the accuracy decreases a little, the recall(yes) for each method doubles, which is a great improvement. As our previous analysis indicating, recall(yes) is the key measurement that we want to optimize, we will say that our data mining methods performs better in oversampling datasets, indicating the effectiveness of oversampling for imbalanced dataset.

Table 6 shows the statistics measurement using oversampling dataset with only features selected by PCA and forward selection. From the results we found out that PCA performs much worse than forward selection in both accuracy and recall(yes) for each method. Combining table 5 and table 6 we can see that PCA doesn't even defeat the performance of using all features. This may due to the complexity of dataset or the existence of some unrelated features, which can have bigger effects on the outputs when the number of features decrease. However, forward selection does improve the results. It increases both accuracy and recall(yes). Among all methods we tried in our

experiments, NN using only features selected by forward selection in oversampling dataset performs the best, it achieved 85.11% accuracy and 96% recall(yes).

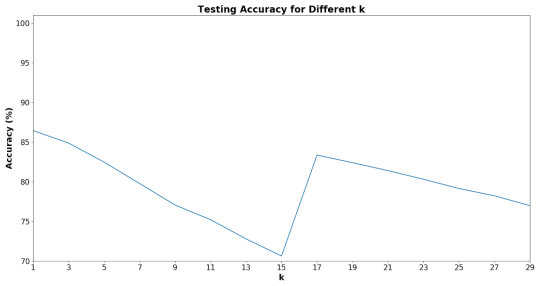

Figure 3: Testing Accuracy versus k for kNN

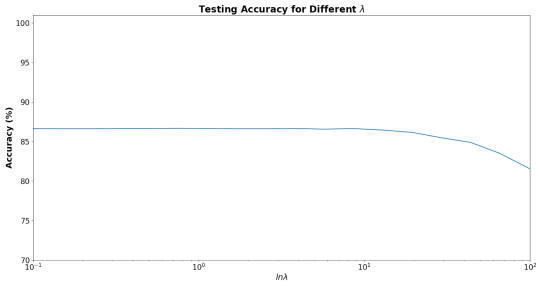

Figure 4: Testing Accuracy versus λ in Logistic Regression

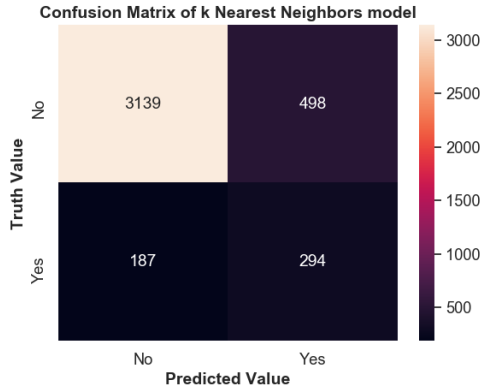

Figure 5: Confusion Matrix for kNN with all features and after oversampling

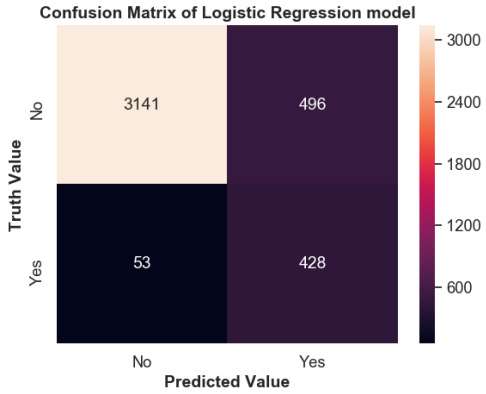

Figure 6: Confusion Matrix of Logistic Regression with all features and after oversampling

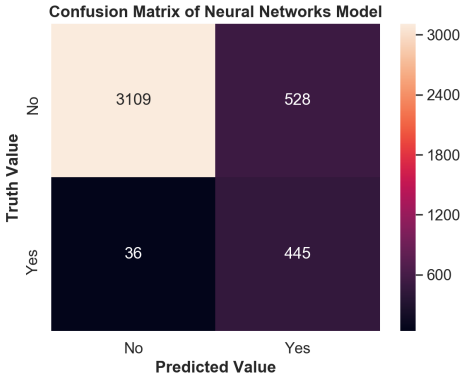

Figure 7: Confusion Matrix of Neural Networks with all features and after oversampling

	Accuracy	Precision (no)	Precision (yes)	Recall (no)	Recall (yes)
kNN	89.39 %	0.91	0.62	0.98	0.23
kNN (Oversampling)	83.37 %	0.94	0.37	0.86	0.61
LR	91.19 %	0.93	0.71	0.98	0.41
LR (Oversampling)	86.67 %	0.98	0.46	0.86	0.89
NN	91.31 %	0.93	0.71	0.97	0.43
NN (Oversampling)	84.82 %	0.99	0.43	0.83	0.96

Table 5: statistics measurements of kNN, LR, NN in dataset with and without being oversampled

Oversampling	Accuracy	Precision (no)	Precision (yes)	Recall (no)	Recall (yes)
kNN (PCA)	73.80 %	0.92	0.22	0.77	0.50
kNN (Forward Selection)	85.65 %	0.96	0.43	0.88	0.71
LR (PCA)	63.09 %	0.94	0.19	0.62	0.68
LR (Forward Selection)	86.30 %	0.99	0.46	0.86	0.91
NN (PCA)	70.37 %	0.93	0.22	0.72	0.62
NN (Forward Selection)	85.11 %	0.99	0.44	0.94	0.96

Table 6: statistics measurements of kNN, LR, NN in oversampling dataset with features selected by PCA and forward selection

6 Conclusions

In this project, we are trying to build models for predicting if a bank client will subscribe to a term deposit. We compare the performance of kNN, LR, and NN. Also, we want to see the impacts of various feature selection, namely PCA and forward selection. Besides, since the datasets we used is imbalanced, we apply the oversampling method to get a better prediction result.

Due to the imbalance problem, classifiers generated from original dataset have similar and high accuracy, precision and recall for negative entries but relatively low precision and recall for positive entries. But we found out that oversampling is helpful to handle our imbalanced dataset, it maintains a considerable accuracy meanwhile raise the recall (yes) for each method. The forward selection also helps optimize results while PCA doesn't show to be useful in our case. Among the three classifying methods, NN works best in terms of recall (Yes), which we considered as the key performance evaluation.

ACKNOWLEDGMENTS

We would like to express sincere gratitude for our professor, Vagelis Papalexakis. He gave a wonderful lecture this quarter about data mining and provided plenty of advice about how we design and implement the project.

REFERENCES

[1] S. Moro, R. Laureano and P. Cortez, "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology," Proceedings of the European Simulation and Modelling Conference, pp. 117-121, Guimaraes, Portugal, October, 2011.
[2] J.S. Malik, P. Goyal, M.K. Sharma, "A Comprehensive Approach Towards Data Preprocessing Techniques & Association Rules", *IES-IPS Academy*, Rajendra Nagar Indore, 2010.
[3] S. Moro, P. Cortez and P. Rita, "A Data-Driven Approach to Predict the Success of Bank Telemarketing," Decision Support Systems, Elsevier, vol. 62, pp. 22-31, June 2014.
[4] A. Khashman, "Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes," Expert Systems with Applications, vol. 37, no. 9, pp. 6233-6239, 2010.
[5] C. S. T. Koumétio, W. Cherif and S. Hassan, "Optimizing the prediction of telemarketing target calls by a classification technique," International Conference on Wireless Networks and Mobile Communications (WINCOM), Marrakesh, Morocco, pp. 1-6, 2018.
[6] H. A. Elsalamony, "Bank Direct Marketing Analysis of Data Mining Techniques," International Journal of Computer Applications, vol. 85, no. 7, pp. 12-22, 2014.
[7] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.
[8] Chollet, Fran., et al. "Keras" https://keras.io. 2015
[9] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16. 321-357, 2002.