PROPOSAL

Rohan Narayan Koli
MSc. in Data Analytics 2020-21
Student Id : x19224842

November 20, 2020

1 Motivation

1.1 Credit Card Clients

In recent times, plastic money has replaced currency notes to a large extent for day-to-day transactions. One of the routes of daily transactions is via *credit cards*. Credit cards are mainly used for benefits such as security against fraudulent transactions, reward points, cashback schemes, sky miles etc. While many banks issuing credit cards earn by the route of annual fees, interest on delayed payments and other miscellaneous charges, they are at a risk if their customers default on bill payments. Hence, *customer profiling* and *monitoring the credit* limits are highly important tasks for the business.

To understand these factors, I found a dataset of around 30,000 customers using credit cards provided by a Taiwanese bank in the year 2005.

1.2 Climate Forecast

Accurate weather forecasts play a very important role in the fields of agriculture, prediction and assessment of natural hazards such as floods, wildfires, energy consumption of households and industries, airplane travel, tourism, in assessing global warming and sea levels.

To understand and get more insights of how many weather providers such as weather channels, weather websites predict weather accurately, I have chosen the dataset of the weather in Dublin, Ireland.

1.3 Cardiovascular Disease

According to WHO in 2016, heart diseases were the number one cause of deaths worldwide. Many activities related to our lifestyle such as exercise, smoking, alcohol consumption, food habits, etc. affect the well-being of the heart. Early, swift and automated diagnosis through machine learning algorithms can be the key to saving many lives worldwide. I believe, such methods if implemented would decrease consultation costs and hence increase affordability.

I came across such dataset on *Kaggle* containing the medical records of 70,000 people which can be used for training and testing different machine learning algorithms.

2 Research Question

2.1 From Credit Card Clients dataset:

Can we use machine learning algorithms to predict defaulting customers before time, and hence restrain further credits to the respective customers and ultimately minimize losses?

2.2 From Weather dataset:

Can we estimate the temperature using historical hourly data? Is it possible to predict rainfall when parameters such as wind speed, temperature, vapour pressure are know by using machine learning algorithms?

2.3 From Cardiovascular Disease dataset:

Can machine learning be used for accurate and automated diagnosis of Cardiovascular Diseases?

3 Initial Literature Review

3.1 Credit Card Clients

”I. Cheng Yeh and Che-hui Lien” used six different machine learning algorithms to predict the defaulting clients and noted that artificial neural networks performed the best among them with R square of 0.9647. [1]. In another study, the author compared the performance of three different machine learning algorithms and found Random forest performed better compared to Logistic Regression and Decision Trees on the same dataset [2]. Furthermore, ”Liyu Gui’s” research indicates AdaBoosting and neural networks were the most competing algorithms that predicted defaulting clients correctly [3].

3.2 Climate Forecast

A K-NN classifier to predict the weather in Sicily,Italy and the conditions of dry weather were predicted accurately, whereas while predicting the conditions of heavy rain, only 17 percent accuracy was achieved [4]. Another study suggests using Polynomial Regression or SVR to predict temperature instead of Isotonic Regression as it performs poorly with predicted data [5]. Furthermore, ”P. Lohani” suggests using ARIMA with LSTM and concludes that LSTM is better at predicting the amount of precipitation in Punjab, India [6].

3.3 Cardiovascular Disease

The research paper by ”A. Maram, N. Mahalakshmi, and N. Niriksha” compares various machine learning algorithms and ranks Multi-linear perceptron neural network with the highest accuracy followed by Random forest and SVM [7] . In a similar study, Decision tree classifiers performed better in diagnosing CDV compared to other algorithms [8]. Furthermore, a study found that logistic regression yields better results compared to other machine learning algorithms apart from

implementing deep learning techniques (which yielded a 7.52 percent higher accuracy over logistic regression) [9].

4 Data Sources

4.1 Credit Card Clients

The dataset is a credit card bill payment of a Taiwanese bank in the year 2005 of 30,000 customers with 23 variables. The variables include the customer's basic demographic profile description along with their past payment and billing records from April to September 2015. The data is available in an excel file format with no missing values.
Source : https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

4.2 Climate Forecast

The dataset contains an hourly record of the climate parameters over 30 years collected from the *Dublin Airport* weather station of Ireland since January 01, 1990. There are 270,289 records along with 20 variable climate parameters. The data is available in a Comma-separated values file format with no missing values.
Source : https://www.met.ie/climate/available-data/historical-data

4.3 Cardiovascular Disease

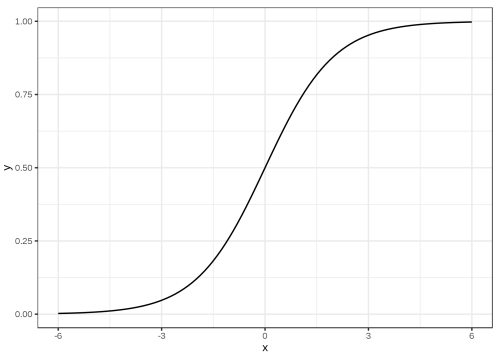
The dataset contains medical records of 70,000 patients containing basic information along with habitual and lifestyle information over 12 variables or features. The target variable i.e the presence or absence of cardiovascular disease is binary coded. The data is available in a Comma-separated values file format with no missing values.
Source : https://www.kaggle.com/sulianova/cardiovascular-disease-dataset?select=cardio_train.csv

5 Identification of Machine Learning Methods

5.1 Credit Card Clients

5.1.1 Logistic regression

The dependent variable of our dataset (Default of payment) is a dichotomous variable and a problem of classification. Logistic regression incorporates a *logit* which is a natural log of the odds ratio and models the probabilities with two possible outcomes [10]. It has a 'S' or 'Sigmoid' shaped graph. Unlike linear regression, here the coefficients are estimated using maximum likelihood estimation (MLE).



$$P = \frac{1}{1 + e^{-(\beta_0 + \Sigma\beta_i X_i)}}$$

5.1.2 Random Forest

Random forest is a step ahead of *Decision tree* model which when used selects bootstrapped data randomly from the dataset [11]. Next, the model selects parameters randomly as the root node forming a variety of individual decision trees. The output is generated by the *bagging* algorithm by aggregating the votes from each decision tree [12]. I chose Random forest as our problem question is of a *classification* type and the model handles both classification and regression tasks. The mechanism of the algorithm introduces randomness which generates more accurate results through cross-validation.

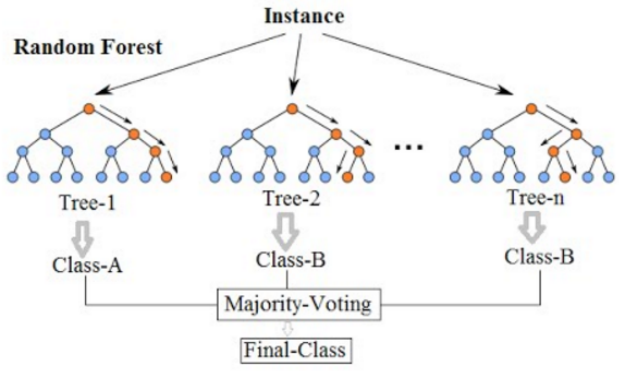


Figure 2: Random Forest

5.2 Climate Forecast

5.2.1 ARIMA : Auto-Regressive Integrated Moving Average Model

Auto-regressive (AR) models predict values based on the past models whereas Moving Average models (MA) use only the error terms for forecasting. Combining both the techniques, we get the ARMA model. An ARMA model applied to d differenced data is called ARIMA. The model is based on the principle of *stationarity* which means that the mean and variance at any point in time should be stationary [13]. To check for stationarity, Augmented Dickey-Fuller (ADF) test is used. To make the data stationary, the model needs to be differenced. The lags are determined by Auto-corelation function (ACF) and Partial Auto-correlation function (PACF). Other than using multi-linear regression algorithm, I would also use the ARIMA algorithm to forecast temperature and compare the two models for accuracy.

5.3 Cardiovascular Disease

5.3.1 k-NN algorithm

K-NN algorithm is useful in the classification problems as it uses the nearest input in the train-set to determine the class of this nearest input [10]. The nearest neighbours are determined by the *Euclidean distance* or the shortest direct route method to determine the class [11]. Choosing the best value for 'K' is similar to maintaining the balance between bias and variance. Although this algorithm is categorized as *lazy* learning algorithm (since abstraction and generalization do not occur), it can be useful in predicting the probability of a person being diagnosed of Cardio vascular disease (CDV) from the dataset.

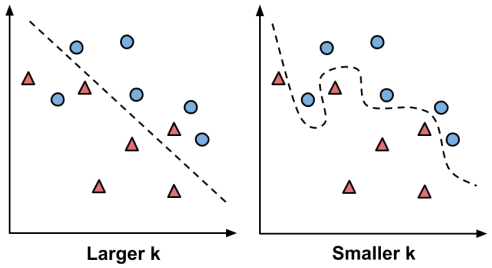


Figure 3: Random Forest

5.3.2 Naive Bayes

Naive Bayes is one of the popular and widely used classification algorithm that utilizes Bayes' conditional theorem as a base to classify the dependent variable. The algorithm assumes that all features in the dataset are equally important and independent; hence the word *naive* [12]. Although these assumptions are rarely true in the real world, this algorithm fetches fast and accurate results. This algorithm is therefore used in most real-time calculations such as in forecasting weather, medical science and stock markets. Being a classification problem to predict CDV, I am eager to test the effectiveness with the selected database.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where, P(A|B) is the posterior probability of class given predictors; P(A) is the prior probability of class; P(B|A) is the the probability of predictor given class; P(B) is the prior probability of predictor.

6 Identification of Evaluation Methods

6.1 Classification Model

6.1.1 Confusion Matrix

The confusion matrix is a table which categorizes the predictions with their actual values.It is a square matrix whose dimension are equal to the predictors used [10]. The matrix classifies pre-

dictions into true positive (TP), true negative(TN), false positive (FP), false negative(FN). Using these categories, the model's accuracy and error rate can be measured as below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \text{ and } Error\ rate = 1 - Accuracy = \frac{FP + FN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN} \text{ and } Specificity = \frac{TN}{TN + FP}$$

Sensitivity and Specificity measures the positive and negative instances respectively that were correctly classified.

$$Precision = \frac{TP}{TP + FP} \text{ and } Recall = \frac{TP}{TP + FN}$$

Precision measures how frequently, when a model makes a positive prediction, it ends up being right while Recall measures how certain we can be that all the occurrences with the positive predictions have been found by the model.

$$F1\ measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1 measure is the harmonic mean of precision and recall which assumes values between 0 and 1; higher values indicate better performance.

6.1.2 ROC curve

The Receiver operating characteristic (ROC) depicts the performance of the classifier by plotting the True positive rate (TP Rate = 100 X TP/(TP+FN)) on vertical axis against the False positive rate (FP Rate = 100 X FP/(FP+TN)) on the horizontal axis [10].To create the curves, the predictions are sorted by the estimated probability of the positive class, in ascending order. Beginning at the origin, each prediction's impact on the TP rate and FP rate will result in a curve tracing vertically (for a correct prediction) or horizontally (for an incorrect prediction). The closer the curve is to the perfect classifier, the better is the model at predicting values which is measured with the help of area under the ROC curve (AUC).

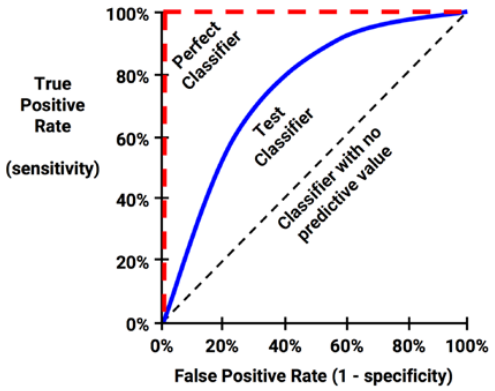


Figure 4: The ROC curve depicts classifier shapes relative to perfect and useless classifiers

6.2 Regression Model

6.2.1 RMSE

The root mean squared error is derived from the residual errors by taking the square root of the mean squared error between the predicted and actual values [14]. RMSE penalizes large errors.

$$RMSE = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

6.2.2 R squared error

R squared error is the measure of how closely the linear regression line slope matches the independent variables. It represents the proportion of variance between the dependent and independent variables. The R square value is in the range between 0 and 1. Higher value symbolizes better predicting model.

$$R^2 = 1 - \frac{error\ \ sum\ \ of\ \ squares}{total\ \ sum\ \ of\ \ squares}$$

References

[1] I.-C. Yeh and C. hui Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications*, vol. 36, no. 2, Part 1, pp. 2473 – 2480, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417407006719

[2] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi, and K. A. Kasmiran, "Credit card default prediction using machine learning techniques," in *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*. IEEE, oct 2018. [Online]. Available: https://doi.org/10.1109%2Ficaccaf.2018.8776802

[3] L. Gui, "Application of machine learning algorithms in predicting credit card default payment," 2019. [Online]. Available: https://escholarship.org/uc/item/9zg7157q#author

[4] M. Raffaele, M. T. Caccamo, G. Castorina, G. Munaò, and S. Magazù, "A didactic approach to the machine learning application to weather forecast," 2020. [Online]. Available: https://arxiv.org/abs/2006.16162

[5] A. Shafin, "Machine learning approach to forecast average weather temperature of bangladesh global journal of computer science and technology: D neural artificial intelligence," 08 2019. [Online]. Available: https://computerresearch.org/index.php/computer/article/view/1858/1842

[6] P. Lohani, "Daily precipitation forecasting using neural network - a case study of punjab,india," 07 2019. [Online]. Available: http://norma.ncirl.ie/id/eprint/4324

[7] A. Maram, N. Mahalakshmi, and N. Niriksha, "Prediction of heart disease and diabetes using machine learning," 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222105058

[8] A. K. Sonam Nikhar, "Prediction of heart disease using machine learning algorithms," *International Journal of Advanced Engineering, Management and Science(ISSN: 2454-1311)*, 2016.

[9] N. Pereira, "Using machine learning classification methods to detect the presence of heart disease," 2019.

[10] B. Lantz, *Machine learning with R: learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*, 5th ed. Birmingham: Packt Publ., 2015.

[11] T. Hastie, R. Tibshirani, J. Friedman, R. Tibshirani, and J. Friedman, *Elements of Statistical Learning, The*.

[12] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of machine learning for predictive data analytics*.

[13] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications*, 2017.

[14] M. Harrison, *Machine learning pocket reference*.