Department of Computer Science
Practical Business Analytics Project (COMM053)

Group: D-Walk

Group Members:
Mohamed Ali (6675010)
Ranjeet Singh Yadav (6656616)
Pavan Siddihally Venugopala Reddy (6644984)
Evans Gichuki (6677196)
Arun Chawdary (6679047)
Hitesh Marisetty (6673633)

Will our customers have the intention of leaving us in the future?

Table of Contents

2 Objectives

2.1 Problem Definition

With growing competition in many different industries, we find that maintaining customer interest and satisfaction is a matter of great importance. Whether the customer remains interested in the service being provided depends on a number of factors and variables; from which patterns can be analysed in order to provide a future prediction for a business. The importance of this cannot be understated as the revenue received by the business will be directly proportional to the customers interest and satisfaction.

2.2 Outlined Aims

This report therefore aims to explore these factors and variables to determine whether the Telecommunications Industry will be able to maintain the customers interest in the future. It also aims to determine which factors and variables are the most paramount in ascertaining this. It seems that in previous studies of this nature, it has been found that there is no real consensus with regards to the most efficient classifiers in order to determine voluntary customer churn. (Telco Customer Churn, 2021) Therefore, in this study we will be employing a number of classification algorithms to weigh up the patterns that we are able to detect through their collective analyses. Through following a CRISP-DM approach, we will be able to underline the main process and principles with which the data will be analysed in an efficient and concise manner – figure 1.1 shows the outlined approach.

2.3 CRISP-DM Roadmap

The 6 CRISP-DM phases our approach followed are:

1. Business Understanding

Our project focuses on the telecommunications industry, in order to gain some business understanding we conducted a short literature review to see what has been done previously in this study as shown in the outlined aims.

2. Data Understanding

In terms of understanding the data, we collected the data first as a dataset. Following this we examined and described the data in terms of the number of records and field identities to determine how we must prepare the data. We verified that the data we have is relatively clean through creating box plots and determining the number of outliers.

3. Data Preparation

We prepared the data through selection and cleaning. This was done by handling null values, level encoding processes, class balancing, creating dummy derivative columns and other methods of normalization. We also integrated and formatted the values of variables as required to be analysed efficiently and concisely.

4. Modeling

We modelled the data through four main models: naïve bayes, logistic regression (using cross validation), support vector machine, and k-nearest neighbours algorithm.

The efficiency of these models were assessed often using confusion matrices and other graphs to compare the accuracy of each model.

5. Evaluation

The models were analysed to determine the outcomes and the aims that were met. Patterns were evaluated throughout the different pieces of data, to determine conclusions in conjunction with the business criteria.

6. Deployment

This report and study is a product of the above steps and evaluation, to demonstrate the final presentation of our findings.

The CRISP-DM approach is illustrated in the figure 1.1 below.

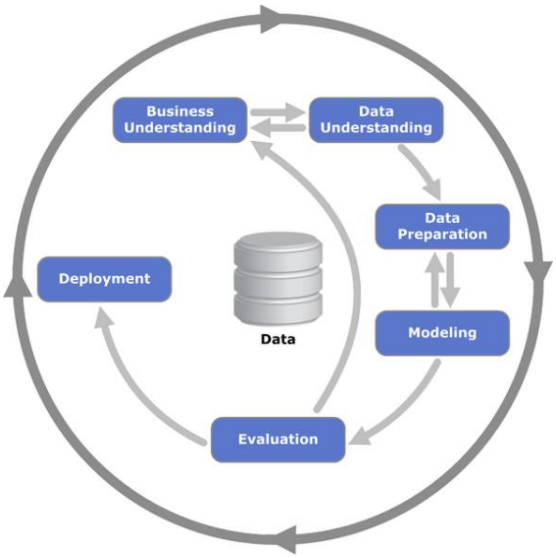


Figure 1.1: Illustration of the CRISP-DM approach (Kaur and Kaur, 2020)

3 Pre-processing

In this project we have used telco-customer-churn dataset which is used to retain the customer by predicting their behavior. The dataset is available at https://www.kaggle.com/blastchar/telco-customer-churn. The dataset contains 21 features and 7032 observations. Churn is the output variable Y which needs to be predicted. R is the programming that is used to perform the task. As we explore the dataset, we have found that dataset is not clean and needs to be pre-processed. The dataset had missing values, categorical variables, scaling the data and randomly shuffling the dataset for randomization. We have dropped the missing value columns and removed unwanted columns like customerID from our dataset. The categorical variables such as 'SeniorCitizen', 'tenure', 'MonthlyCharges', 'TotalCharges' are encoded with labelencoder. The dataset is then scaled between (-1 to 1) so that it becomes normalized, and algorithm does not show any bias. After preprocessing of dataset, we need to train the model.

3.1 Handling Null Values

With regards to our dataset, we find that all our null values exist in the total charges column. There are a total number of 11 values that are null values. In order to handle these values, our approach was that rather than omitting them we could replace them through using the other variables that the total charges are dependent on.

Our approach was to multiply the total tenure with the monthly charges, effectively producing the values for the total charges. We believed this was a more efficient way of dealing with our particular dataset, as we could therefore render all the data as useful rather than omitting values that may prove necessary in our analysis.

$$total\ tenure\ \times monthly\ charges = total\ charges \quad (1.1)$$

3.2 Normalisation and Class Balancing

Using a for loop – we translated all the ‘yes/no’ variables into a compact and optimized binary code prior to executing our model. This is necessary because the instructions would then be treated as numeric datum, and therefore normalized. Another problem that was encountered in the dataset was that some of the columns contained more than 2 variables – as an example in the internet service, we find that the variables included answers ‘yes’, ‘no’, ‘no internet service’. In this case we normalized the variables that entail the same outcome, by changing ‘no internet service’ variables to ‘no’ for example.

Following this, in the case that the variables were multiple in number and did not entail the same outcome, we created dummy columns to level encoding for those multiple variables. Some examples of of the dummy columns we created with regards to the payment method of the customers are ‘PaymentMethod.xCredit.card..automatic.” to represent the amount of people paying with credit card (automatic), and ‘PaymentMethod.xMailed.check’ to represent those who paid through mail.

Another issue that we encountered was some large distances in data within some columns. We understood that due to the features not having the same scales, there would be a chance of the higher values being considered as weightier in the analysis; somewhat affecting the performance of the algorithm and making it biased towards higher magnitude values. Also in

consideration that we were going to employ a KNN approach on our data, we felt it was necessary to use feature scaling in order to try and normalise the data, so that the values are all between 0-1 and efficiently analysed.

An example of the result of this analysis is illustrated in the figure 1.2 below, whereby the first table shows the original data and the second shows it after it has been feature scaled.

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

	Student	CGPA	Salary '000
0	1	-1.184341	1.520013
1	2	-1.184341	-1.100699
2	3	0.416120	-1.100699
3	4	1.216350	0.209657
4	5	0.736212	0.471728

Figure 1.2: An illustration showing feature scaling of a set given of data (Bhandari et al. 2020)

3.3 Data Sampling

Opting for a conventional approach, we separated our data into training and testing data into 70% and 30% respectively. We did this using the 'createDataPartition' function which creates balanced splits of the data.

We also wanted to determine the correlation between specific variables prior to the analysis so that we can project outcomes and use correlations to further analyses. We find that in our figure 1.3 shown below, that we are able to correlate between monthly and total charges as their respective values of 0.65 and 0.83 are relatively close to one another.

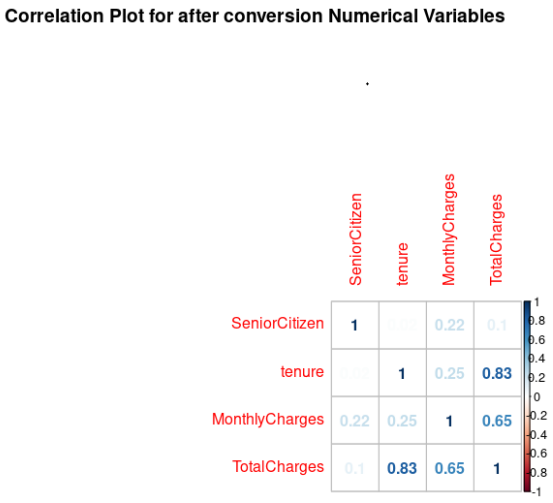


Figure 1.3: Illustration showing the correlation between specific variables

4 Data Modelling

4.1 Naïve Bayes approach

The basis that the Naïve Bayes approach operates on is where joint probability is considered, i.e. the likelihood of event X to occur alongside event Y. This probability is illustrated in the equation 1.2, showing the basic equation with regards to joint probability.

$$p(X|Y) = \frac{p(X)\cdot p(Y|X)}{p(Y)} \quad (1.2)$$

Employing this method, the Naïve Bayes Classifier approach is where this basic principle is utilized in a method of multiplicity to find a class of observation. This is to calculate the probability of a class given a set of feature values. This idea is illustrated in equation 1.3 below.

$$p(y_i \mid x_1, x_2, \dots, x_n) = \frac{p(x_1,x_2,\dots,x_n \mid y_i)\cdot p(y_i)}{p(x_1,x_2,\dots,x_n)} \quad (1.3)$$

The Naïve Bayes approach was utilized for a specific reason in our model. The first being that this approach assumes that all features are independent of one another, this would give us the conditional probability of multiple variables affecting our prediction outcome. Whilst acknowledging that often, rather the variables in our case may not be completely independent of one another; this would be utilized for a restricted conclusion with regards to our aim to determine how paramount each variable is in affecting our final predicted outcome. (Turhan and Bener, 2009)

4.2 Logistic Regression (Using Cross Validation)

Logistic regression is a method which is used to model the probability of an event occurring, in terms of odds ratio in the presence of one or more explanatory variables. It builds upon the theory of linear regression whereby a straight line is fit to a binary response as in equation 1.4. (Kumar Bhowmik, 2015)

$$p(X) = \beta_0 + \beta_1 X \quad (1.4)$$

Logistic regression however is modelled to ensure outputs between 0 and 1 for all values of X, by using the logistic function to model p(X) shown in equation 1.4, and then taking the logarithm of each side to provide the logistic regression model equation as shown in equation 1.5 below. The specific type of logistic regression we employed is binomial whereby the variables represent two possible values. (Attanayake, Jayasundara and Peiris, 2016)

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \quad (1.5)$$

We implemented the logistic regression approach through using the glm() function, which is usually generally used to fit linear models. (Bhandari et al. 2020) It was modified through setting the ‘family’ argument to ‘binomial’; this specifies the logistic regression model through R.

Following the data being split into training and testing data, we also implemented the k-fold cross validation method on our model to increase accuracy and efficiency. This works by dividing the samples in the training data into *k* number of groups. These are known as folds,

and they are then validated to obtain a *k*-fold value. Following this, the model can be built in order to train our dataset in a more efficient and accurate manner.

The logistic regression approach was employed in our model because our model is a binary classification model, in which this approach is proven to work well and in many cases is the go-to for this type of model due to it being less inclined to overfitting in low dimensional datasets. The main advantages with regards to employing this model in our case are that it is simple and fast to implement with unknown records, along with indicating the feature importance of specific variables as outlined in our aims.

In order to the measure the performance of this classification model, a confusion matrix was also applied along with the other models as shown below in figure 1.4.

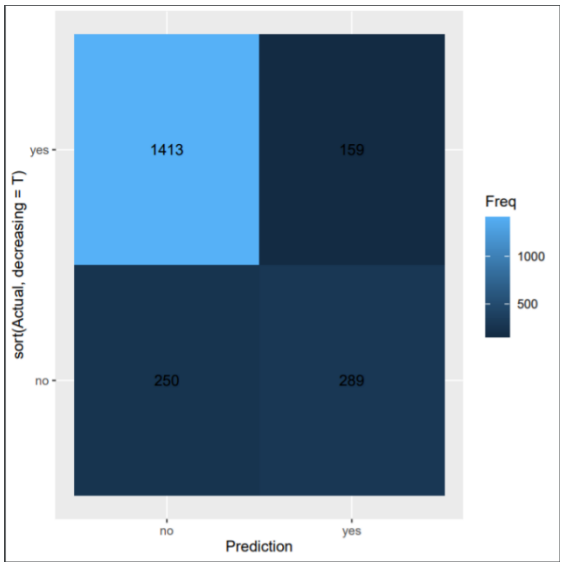


Figure 1.4: An illustration of the confusion matrix applied on the logistic regression approach

4.3 Support Vector Machine (SVM)

We also employed a Support Vector Machine which is a supervised machine learning algorithm commonly used in classification problems. This is done by plotting each data item as a point in *n*-dimensional space (where *n* is the number of features) holding a value of a co-ordinate on the plot which is then kernelled. Following this, the hyper-plane is determined which should differentiate between the classes in a manner which is efficient and accurate. (Shah and Gopal, 2012)

We employed this method in our model firstly due to the size of our dataset being relatively small and clean, which Support Vector Machines are more accurate in determining. We find that with larger noisier datasets, this is not usually the case.

4.4 K-Nearest Neighbours Algorithm (KNN)

K-Nearest Neighbours Algorithm was also implemented as a classification method. This method works by assuming that similar things are close to each other and exist in close proximity. This uses the distance factor, to determine which factors and variables are similar to one another. There are different types of distance that are considered in this method such as Manhattan distance, cosine similarity, Makowski, Euclidean distance and others. (Attanayake, Jayasundara and Peiris, 2016) For this project we are using Euclidean distance.

In order to choose the right value of K, we can use a trial and error method to see which value of K is most reasonable. As the value of K is decreased, the predictions become less stable whereas when they are increased too high then overfitting may occur. (Kaur and Kaur, 2020) Following this model, we created a graph to show the accuracy of the model whilst changing the k-value as illustrated in the figure 1.5 below. We chose a value of K at 8 because it gave us a decent accuracy of 95.7%.

One reason why we used this method of classification is because it is likely to show inherent relationships between variables prior to any changes in parameters or assumptions being made. Other reasons we used this method was because it was easy to implement, and is well known to be employed when dealing with a binary classification problem.

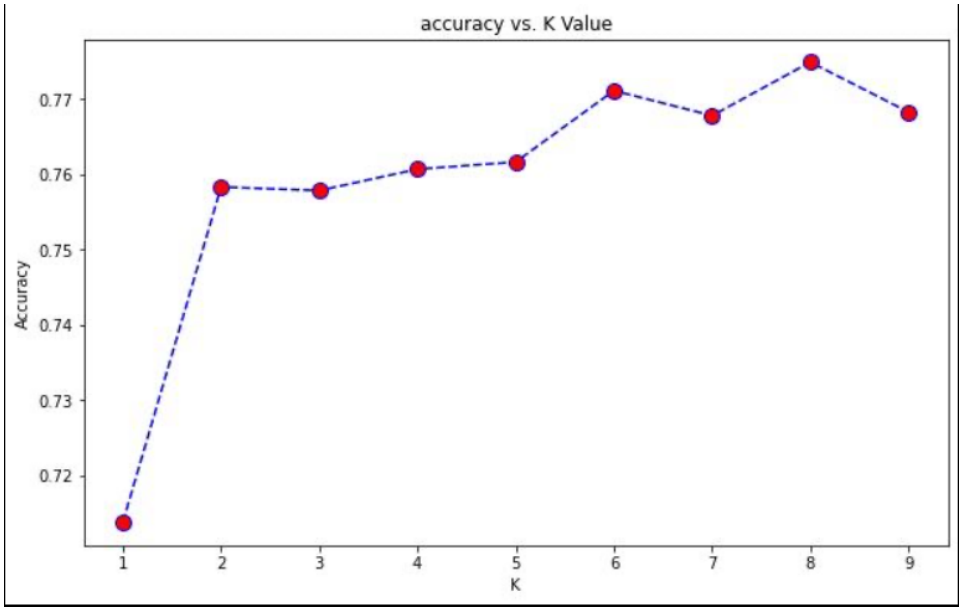


Figure 1.5: An illustration of the relationship between accuracy and the K-value variable

5 Results and Discussion

5.1 Individual Model Analysis

In our model, we plotted two graphs showing the correlation between the charges that a customer was incurring and the churn rates. The two graphs were to show the charges incurred per month, and total charges incurred as opposed to the churn rate respectively. We find that generally speaking, the relationship between the amount of customers churning and the monthly charges has an unclear relationship, although if we consider the early values to be an outlier considering the count of those who do not churn being exponentially higher - it seems to be a somewhat positive relationship.

From the second illustration showing the relationship between churn rates and total charges, we find that there is a negative relationship whereby those who incur less charges are more likely to churn. This is an interesting observation, considering that the first illustration showing the relationship between monthly charges and the churn rate has a somewhat positive relationship. The described relationship is shown in the figure 1.6 below.

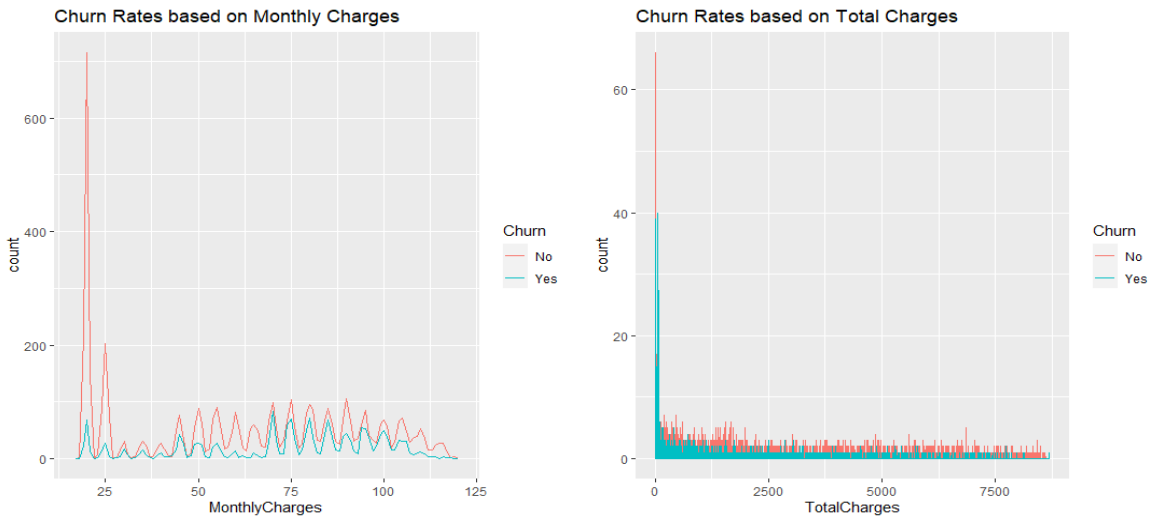
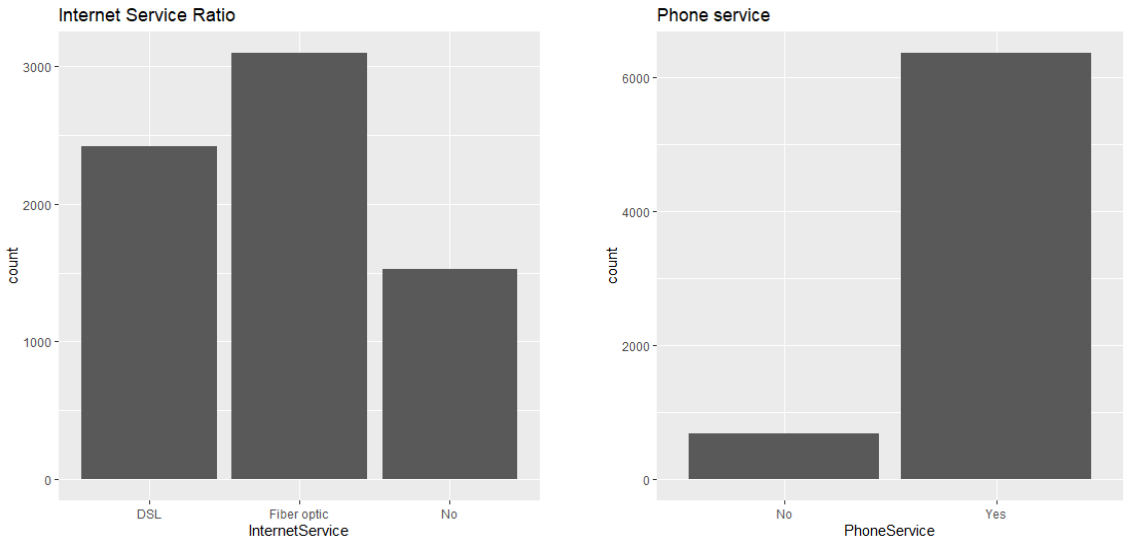


Figure 1.6: Illustrations showing the relationship of the churn rates and the charges (monthly and total)

Two graphs were plotted to compare the services being used by the customers in the industry. The two overall services plotted were the broadband service ratio, alongside the telephone service ratio. The broadband service was split into DSL and fibre optic, as there is a difference in the prices that are paid for these services.

We find in figure 1.7 illustrated below that although we separate the broadband services into two, all the broadband users together summate to approximately 5500. On the other hand, when looking at the telephone service, the customers are greater in number at over 6000. We find therefore, that the more commonly used service in this industry is the telephone service. Also what we find is that a higher proportion of customers do not use the internet service than the telephone service.

Figure 1.7: Illustrations showing the services used by the customers (broadband and telephone services)



Another relationship that we explored was the relationship between the genders and the services that they bought through the figure 1.8 below, alongside the relationship between the total charges and the internet service that the customer chooses. We find that the disparity between the male and female genders in choosing an internet service is very minimal, however it is notable that the DSL service and those who choose not to use the internet service, both have a greater number of females.

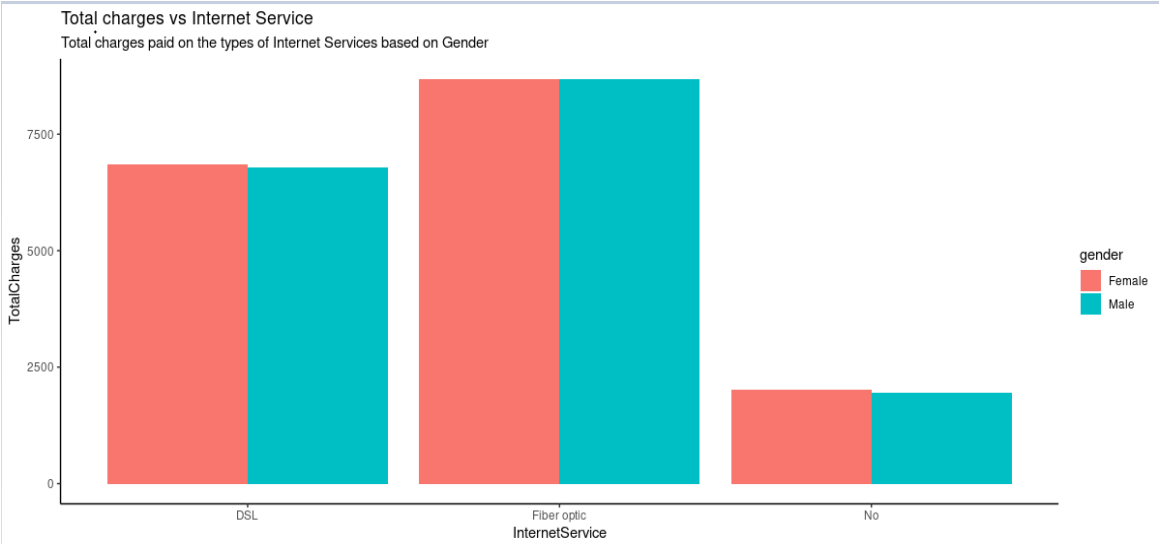


Figure 1.8: An illustration showing the relationship between gender, total charges and internet service.

We further explored the gender relationship by projecting a model with regards to the correlation between the churn rate and the two genders. We had three models in this case to show the proportion of males and females who churned respectively, along with the gender churn rate all illustrated in figure 1.9 below.

We find that in terms of the two genders, visibly there is no major disparity between them in which is more likely to churn from the first two illustrations. In the figure showing titled 'Gender Churn Rate', we can see that there is a 0.1% greater chance of a female churning than a male. From this, it seems that there is no significant disparity between the two genders. However, there is almost a 1% difference in the males and females in the percentages showing those who did not churn. It is important to note that in a larger dataset, this disparity could be a lot more significant and could affect the prediction by a great amount.

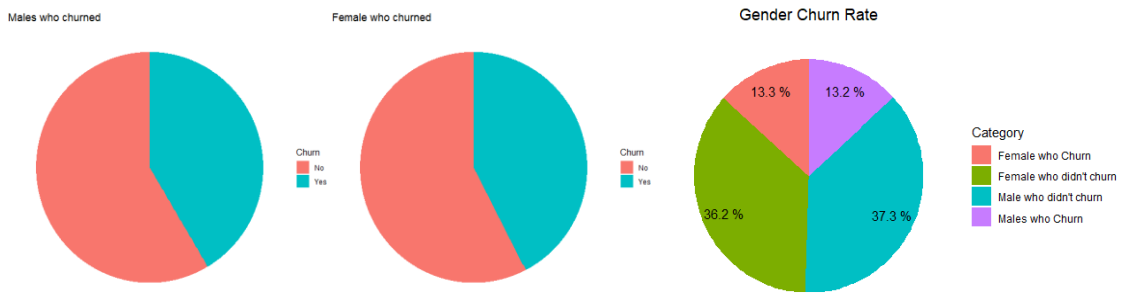


Figure 1.9: Illustrations showing the correlation between gender and the churn rate

5.2 Pattern Recognition / Evaluation

In evaluating the accuracy of our individual models, we created a bar graph in which the different methods were assessed alongside each other. We find that the highest level of accuracy was found in the KNN method, followed by logistic regression, support vector machine and lastly the naïve-bayes method. Although the highest accuracy was obtained through the KNN method, we found that there was a significant amount of overfitting indicating that this would not necessarily be the best form of analysis.

We therefore concluded that the logistic regression method was the best out of the remaining methods we employed, due to it giving us the highest accuracy. The SVM method was similar in its accuracy, with both methods giving accuracies between 80-81%. The Naïve-Bayes approach gave the lowest accuracy at 76.2%, and this was expected due to the high level of assumptions the approach is based on. The chart is shown below on the figure 2.0.

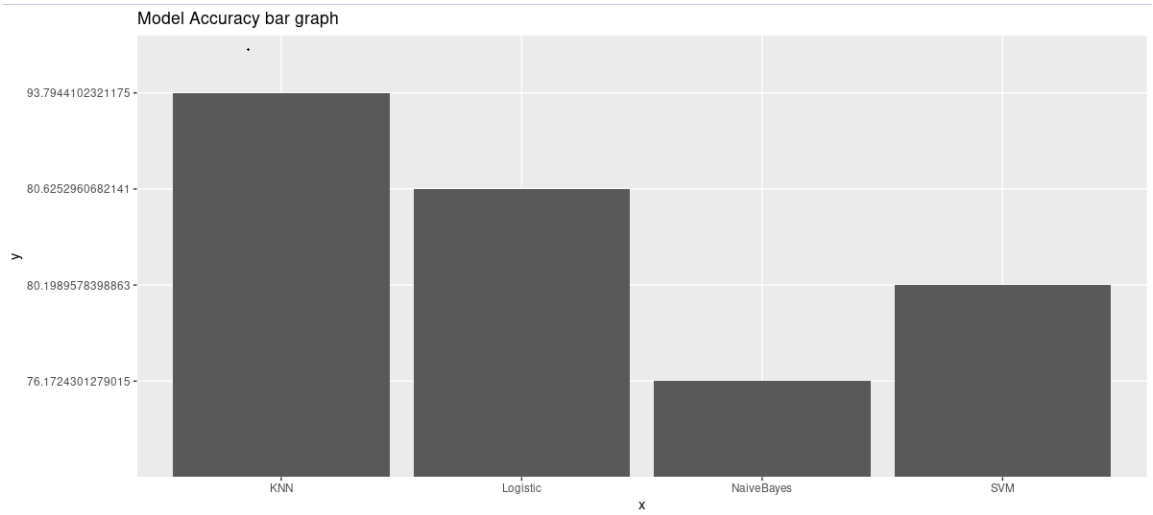


Figure 2.0: An illustration showing the accuracy of each approach that was employed

Alongside the previous figure, we also have a table showing the accuracies of the models as shown below in figure 2.1, whereby the upper and lower quartiles of each model are also considered. We also have the Kappa values that evaluate the classifiers, amongst themselves. We find that all three Kappa values are between 0.38 and 0.46 – which are described as fair or moderate values in determining the accuracy of the observed results.

	Naive Base	Logistic Regression	SVM
Accuracy	0.7617243	8.06E-01	8.02E-01
Kappa	0.38207143	4.61E-01	4.45E-01
AccuracyLower	0.74295765	7.89E-01	7.84E-01
AccuracyUpper	0.77976144	8.23E-01	8.19E-01
AccuracyNull	0.74467077	7.45E-01	7.45E-01
AccuracyPValue	0.03745414	1.42E-11	3.26E-10
McnemarPValue	0.32662651	8.58E-06	7.81E-07

Figure 2.1: Table showing the accuracies of the remaining chosen models

Another main aim of our project was to determine the most significant feature in affecting the churn rate, whereby we have the figure 2.2 below comparing between the features. We find that the tenure is the most significant in affecting the churn rate by a large measure. The lengths of the contract that the clients have are also significant, whereby we find that the longer the client has a contract for; the more significant this is to the churn rate.

We also find that those who use the services: internet, technical support, online security, paperless billing, electronic check all have a very similar level of significance in their churn rates. This is important to the business, as they will be able to predict the rate of churn through examining these key factors and variables that the clients are described through. We found that the least significant factors in affecting the churn rates were whether the clients paid monthly, and used online backup as a service. This is illustrated in the figure 2.2 below whereby the chart demonstrates the importance of these factors.

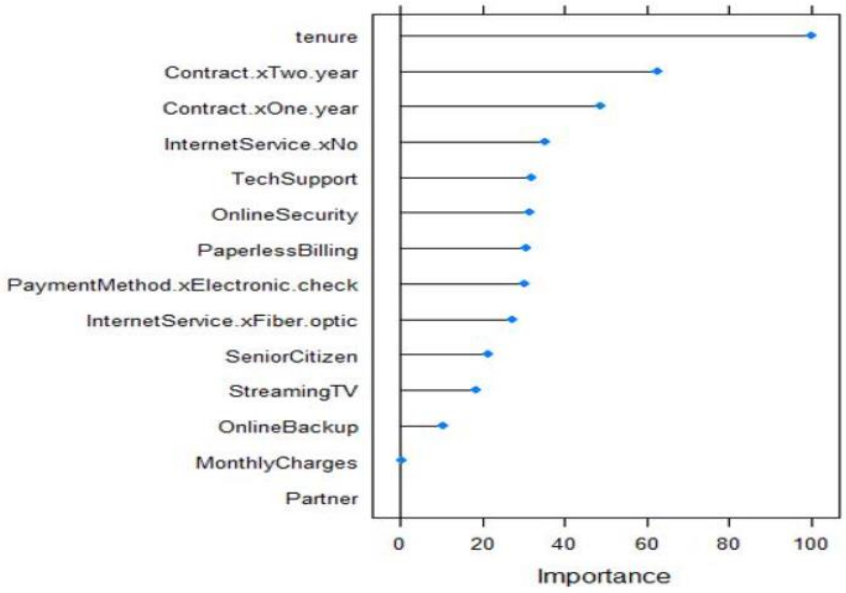


Figure 2.2: An illustration of the relationship between the features and significance in affecting churn rates

6 Conclusions

To answer our aims, we found that the most effective approach in analysing and drawing conclusions from the chosen approaches was the logistic regression approach. This gave us an accuracy of approximately 81%, and was better overall than the other approaches as the KNN method contained some overfitting.

With regards to the most significant features that need to be concentrated in order to achieve customer retention, we concluded from the models and visualisations that the customers who utilize the phone and internet services are less likely to churn as their tenures tend to be longer when using these services. From the analysis we obtained, we saw that there was negligible differences in correlation with regards to gender and churn rate.

In order to ensure that customer retention occurs at an increasing rate, the business would be required to focus on the services through which retention has tended to stay high i.e. the internet and phone services. With regards to the services which do not have this effect, it would be useful to have further perks and subscription offers to gain the trust of the customers.

7 Bibliography

Attanayake, A., Jayasundara, D. and Peiris, T., 2016. AN APPLICATION OF 5-FOLD CROSS VALIDATION ON A BINARY LOGISTIC REGRESSION MODEL. *Advances and Applications in Statistics*, 49(6), pp.443-451.

Bhandari, 2021. *Feature Scaling for Machine Learning: Normalization vs. Standardization*. [online] Medium. Available at: <https://medium.com/analytics-vidhya/feature-scaling-for-machine-learning-normalization-vs-standardization-34daa2d4f707> [Accessed May 2021].

Kaggle.com. 2021. *Telco Customer Churn*. [online] Available at: <https://www.kaggle.com/blastchar/telco-customer-churn> [Accessed May 2021].

Kaur, D. and kaur, S., 2020. Machine Learning Approach for Credit Card Fraud Detection (KNN & Naïve Bayes). *SSRN Electronic Journal*,.

Kumar Bhowmik, T., 2015. Naive Bayes vs Logistic Regression: Theory, Implementation and Experimental Validation. *INTELIGENCIA ARTIFICIAL*, 18(56), pp.14-30.

Shah, H. and Gopal, M., 2012. Reinforcement Learning with Kernel Recursive Least-Squares Support Vector Machine. *International Journal of Machine Learning and Computing*, pp.618-622.

Turhan, B. and Bener, A., 2009. Analysis of Naive Bayes' assumptions on software fault data: An empirical study. *Data & Knowledge Engineering*, 68(2), pp.278-290.

Yancey, R., Xin, B. and Matloff, N., 2020. Modernizing k-Nearest Neighbors. *Stat*,.

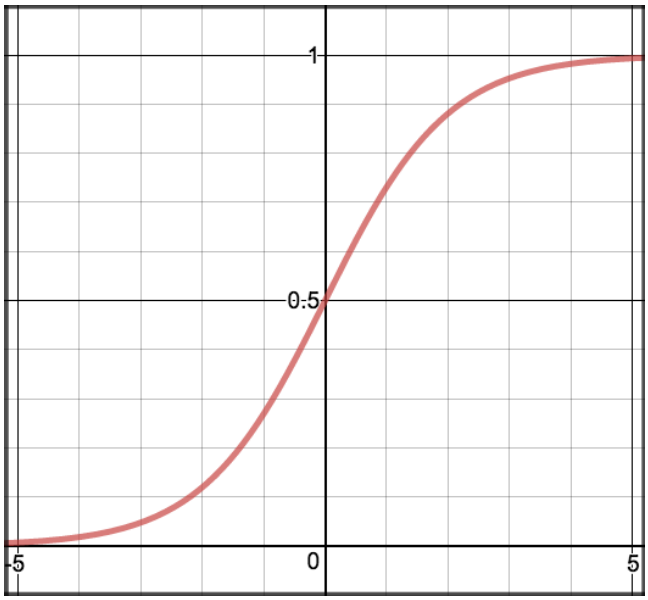
8 Appendix

Sigmoid activation

$$S(z) = \frac{1}{1 + e^{-z}}$$

where:

- s(z) = output between 0 and 1 (probability estimate)
- z = input to the function
- c = base of natural log



Cross-Entropy

$$j(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta\left(x^{(i)}\right), y^{(i)})$$

$$Cost(h_\theta(x), y) = -\log\left(h_\theta(x)\right)$$

if y = 1

$$Cost(h_\theta(x), y) = -\log\left(1 - h_\theta(x)\right)$$

if y = 0

$$j(\theta) = \frac{1}{m}\sum_{i=1}^{m}[y^{(i)} \log\left(h_\theta\left(x^{(i)}\right)\right) + (1 - y^{(i)})\log\left(1 - h_\theta(x)^{(i)}\right)]$$