Machine learning based churn prediction in telecom

Xuhao Dai *

College of Computer and Information Science College of Software, Software Engineering, Southwest University, ChongQing, 401100, China

* Corresponding Author Email: dxh20020113@email.swu.edu.cn

Abstract. In the new era of continuous social development, the level of information technology service construction is also increasing, among which the telecommunications service industry has also developed into a relatively large business cluster under this trend. Due to fierce commercial competition, telecommunication companies also face the problem of losing customers. Asking telecom companies to determine the trend of customer churn from the huge amount of customer information data has become a challenge for all telecom companies to solve. This paper takes a telecom company as a reference, using a large amount of subscriber data from it as a research sample, and uses data mining machine learning and other techniques to make churn predictions on customer data, using three classification methods to evaluate the prediction results and help telecom service providers analyze the causes of customer churn. Firstly, the situation faced by the telecom service company is discussed in terms of the relevant background, then by presenting and analyzing the customer data, and finally by using Logical Regression, Mission Tree, and SVM classification models for classification and evaluation, the causes associated with customer churn are summarized, then the methods of customer churn retention are outlined, and the factors associated with customer churn are listed. The factors associated with churn are listed. This provides a reference for an in-depth study of the theoretical approach and application of customer churn..

Keywords: Churn prediction, Data Analysis, Machine Learning.

1. Introduction

1.1. Background

Along with the increasingly sophisticated development of the public service sector, more and more service industries are facing the problem of losing their customers, especially the communication service industry. Customer churn, which is a more concrete definition of the probability of churn, is when a customer leaves a service company and uses the services of another company in the same sector. In service industries with highly competitive services, customer churn is a major worry. On the other hand, if done early on, identifying consumers who are likely to quit the business could offer a sizable new revenue source [1]. The primary reasons for churn include dissatisfaction with customer service, outrageous pricing, unappealing plans, and inadequate assistance [2]. For a large consumer service formula, it is important to analyse company's customer churn rate. For these industries, customers are accurately linked to their own product services and what they say about the company determines the survival of a service company. Once a customer has churned, the product service chosen by the customer will follow. It is very difficult for Telecoms business to develop a new user into a loyal regular user. The Company need to spend money on publicity and promotion. The following is a series of predictions and determinations based on data on the relevant customer characteristics of a telecommunications company. By serving 90% of the global population, telecommunications have risen to the top of the list of the fastest-growing businesses. It is one of the industries where the customer base is crucial to sustaining income. Customer turnover poses a serious danger to the telecom industry [3].

For a telecom company, all together as a consumer service industry compared to other telecom companies, a telecom company needs to provide better services and better value for money to consumers so that they will be more willing to use the company's product services, firstly to build up a customer base and then to develop customer habits. Get users used to using the company's products

and services rather than other companies. It takes a lot of time and resources to develop an established user, and once that established user is lost, the company must spend the same resources to develop other loyal users. This is very unhealthy for a company's business in terms of return on investment and can have a significant negative impact on the growth of the business as a whole.

Competition in the telecommunications industry is a very widespread issue throughout the world, take for example the telecommunications services industry in India. There are more than 15 mobile operators in India and the entire Indian telecom industry is a highly competitive, prepayment-based market with approximately 96% of subscribers constantly switching between service providers to achieve lower prices for services across the market. Such a phenomenon is largely a reflection of subscriber loyalty, which is directly correlated with the willingness to churn by different operators [4].

Therefore, it becomes crucial for a telecoms company to maximize its profits and position itself in the market. the cost of acquiring a new customer is five times greater than the cost of retaining an existing one, thus increasing client retention by 0.5 can increase profits by 25-95%. Furthermore, sales success rates for existing customers are between 60-70%, while those of new customers are 5-20 [5].

1.2. Purpose

The following key points can first be summarized from the above background analysis.

Retaining as many old users as possible will give us a better basis for competition with other companies and a better edge.

Retaining regular users will allow us to reduce company’s costs significantly.

Retaining old users helps us to develop new ones. Through word of mouth, it is easier to cultivate new users and develop more new users.

It will be easier for the company to expand into new business. This is because old users will be more loyal and more willing to buy and experience new products and services.

The aim of this project is to help companies to characterise their users and analyse the reasons for churn. It will help operators to identify and improve the user experience and identify better solutions to reduce the churn rate.

From the research conducted over the past few years it is apparent that data mining techniques are being used more and more extensively and that machine learning techniques in data mining are becoming more effective in predicting customer churn. Building an effective churn model from data, finding and identifying the most representative variables in customer data from the large amount of available customer data, and mining the data through feature set prediction, involves a very large amount of research [6]. Especially Predictive modeling techniques are often
found to be more accurate in churn prediction [7].

2. Method

2.1. Dataset

The dataset used to analyse customer data contains various indicators, of which 'state', 'account_length' and 'area_code ' indicate the area code of the customer, the location of the service provider used by the customer and the area code of the customer respectively. 'international_plan' indicates whether the user subscribes to the service provider's international service plan and 'voice_mail_plan' indicates whether the customer subscribes to a voice mail plan. In both service plans there is also data on the actual usage of the customer on the basis of the subscription to these plan packages. Among the data indicating 'international_plan' are 'minutes', 'calls', 'charges' and 'charge' represent the minutes used, the number of times used and the total cost of using the plan respectively; the only data representing customer usage of 'voice_mail' is 'number_ customer_service_calls' represents the number of times a customer has used the plan. The next customer data has the customer's usage during the day, indicated by 'day', 'eve' and 'night' respectively,

and then This is combined with the number of minutes, number of calls and cost of calls in a day to give a total of 9 usage figures. The last indicator in the dataset is the churn rate. The three indicators 'state', 'account_length' and 'area_code' are used to identify the area of subscriber and the length of subscription to company's service provider. information and the length of subscription to this service provider. The second indicators 'international_plan', 'voice_mail_plan' and 'number_vmail_messages' are used to identify whether company's subscribers have subscribed to an international service package and a messaging package in addition to this service, both of which may allow customers to adhere to the service. The next data indicators are divided into three main categories, namely the number of calls, the length of the call and the charge for the call, which are "minutes", "calls" and "charge" respectively. ". These are divided into three categories: "_day_", "_eve_" and "_night_". A special indicator remains to identify the length of international calls, international call charges and the number of international calls made by the customer. As this is an international call indicator, the international related call indicator is specifically identified. Rather than being confused with the normal one-day call indicator. The three indicators are "total_intl_minutes", "total_intl_calls" and "total_intl_charge ". The "number_customer_service_calls" identifies the total number of customer service calls, while the final "churn" directly identifies whether a customer has been churned. The initial presentation of the data indicators in the data set shows that the churn analysis is a direct indicator of whether or not a customer has churned based on the specific usage of the customer.

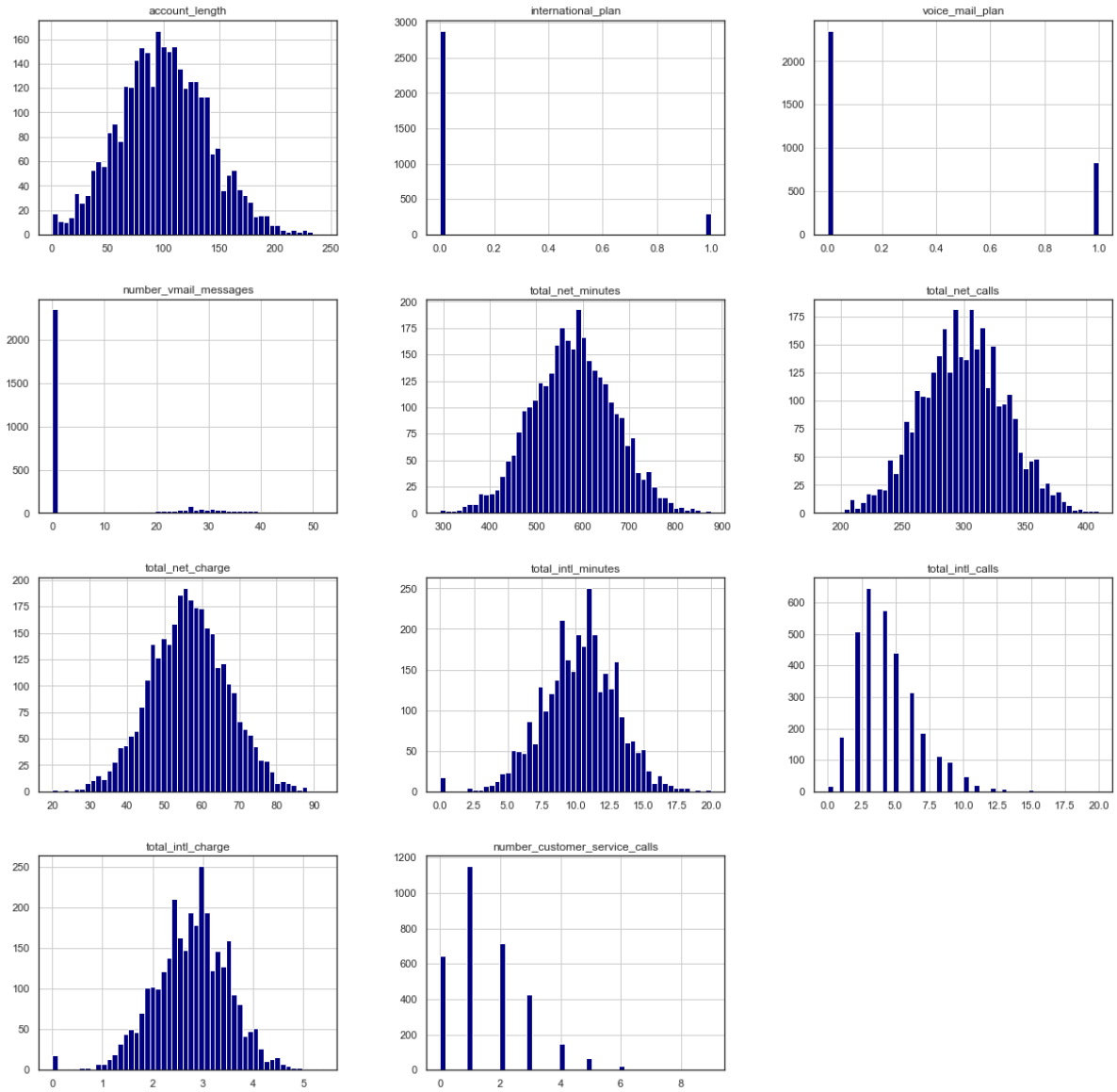


Figure 1. Data visualisation of all data indicators in the dataset.

Figure 1 shows a visualization of the data for each indicator in the dataset, with the full data represented as a bar chart. Data involving bool types can be treated as 1 and 0, with 1 representing true values and 0 representing negative values. All customer usage data can be visualized in a visual bar chart to facilitate the analysis of churn factors.

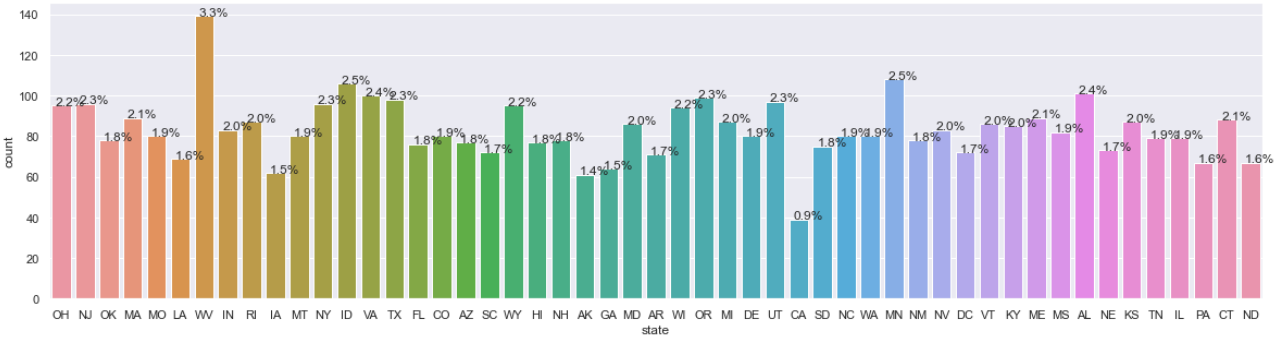
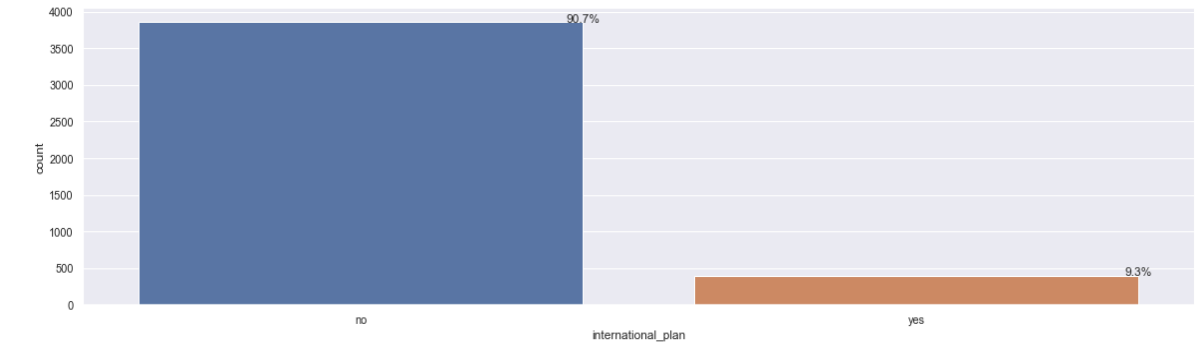
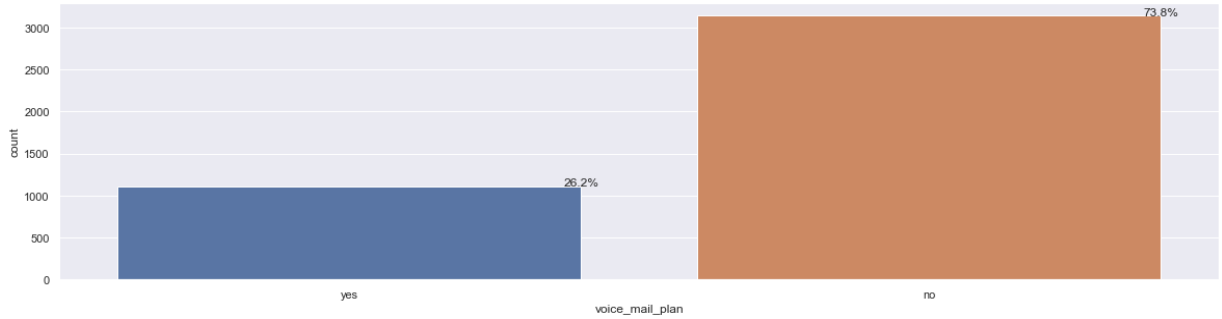


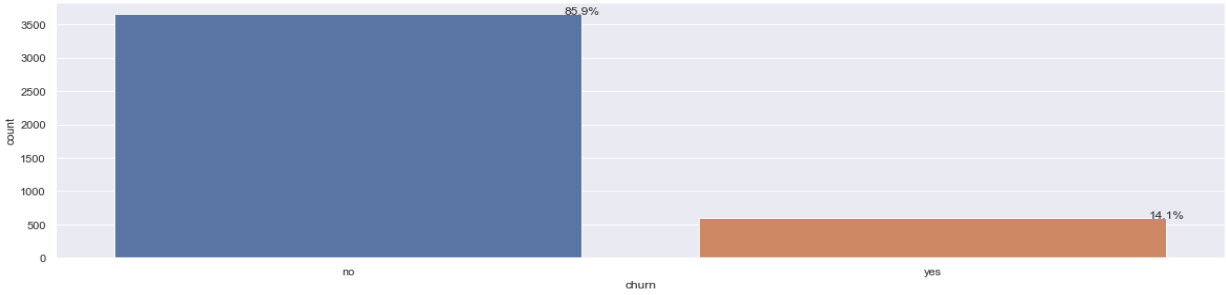
Figure 2. Display of the percentage of place numbers



(a)



(b)



(c)

Figure 3. (a) Display of the international plan subscription percentage; (b) Display of the voice mail plan subscription percentage; (c) Display of customer churn percentage

In Figure 2 it can be concluded that the largest number of customers are from "WV" and the smallest number from "CA". As shown in figure3,90.7 per cent of people subscribe to the international plan. Also 73.8 per cent of people do not subscribe to the voice_mail_plan.In total, only 14.1 per cent of customers were lost across the telecom companies.

The next step is to verify whether the churn is related to three specific indicators of customer usage of the service, shown in the graph below:

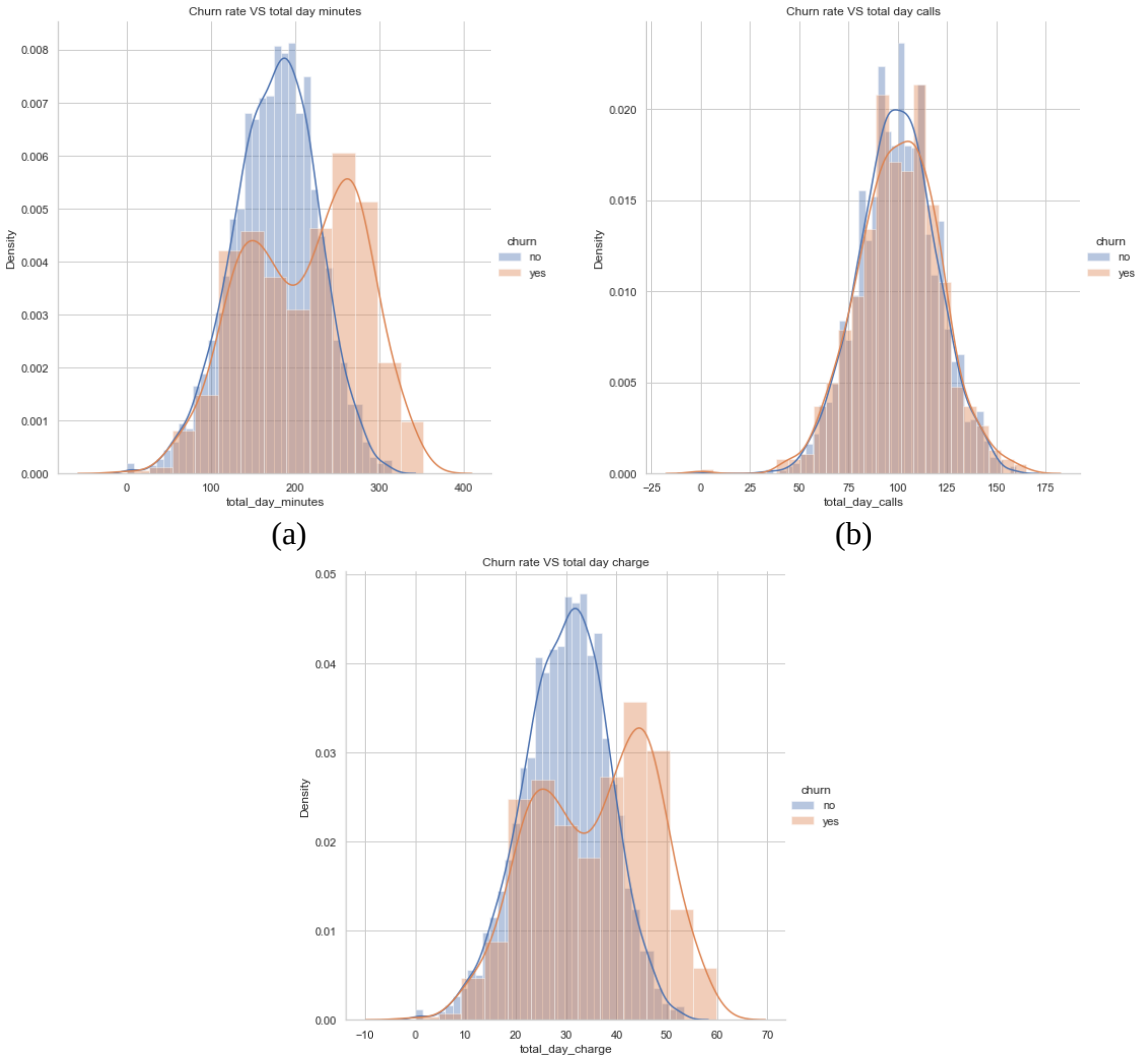


Figure 4. (a) Display of call minutes and customer churn; (b)Display of number of calls and customer churn; (c)Display of call charges and customer churn

As can be seen from the Figure 4 that Churn rate is high when the total_day_minutes lies between 210 min to 300 min. churn rate is high when the total_day_calls lie between 85 to 115. Churn rate is high when total_day_charge lies between 40 to 50. From these charts, one can probably conclude that the more frequently customers use the service, the more time they spend on the service due to the increased frequency and the more money they end up spending on the service.

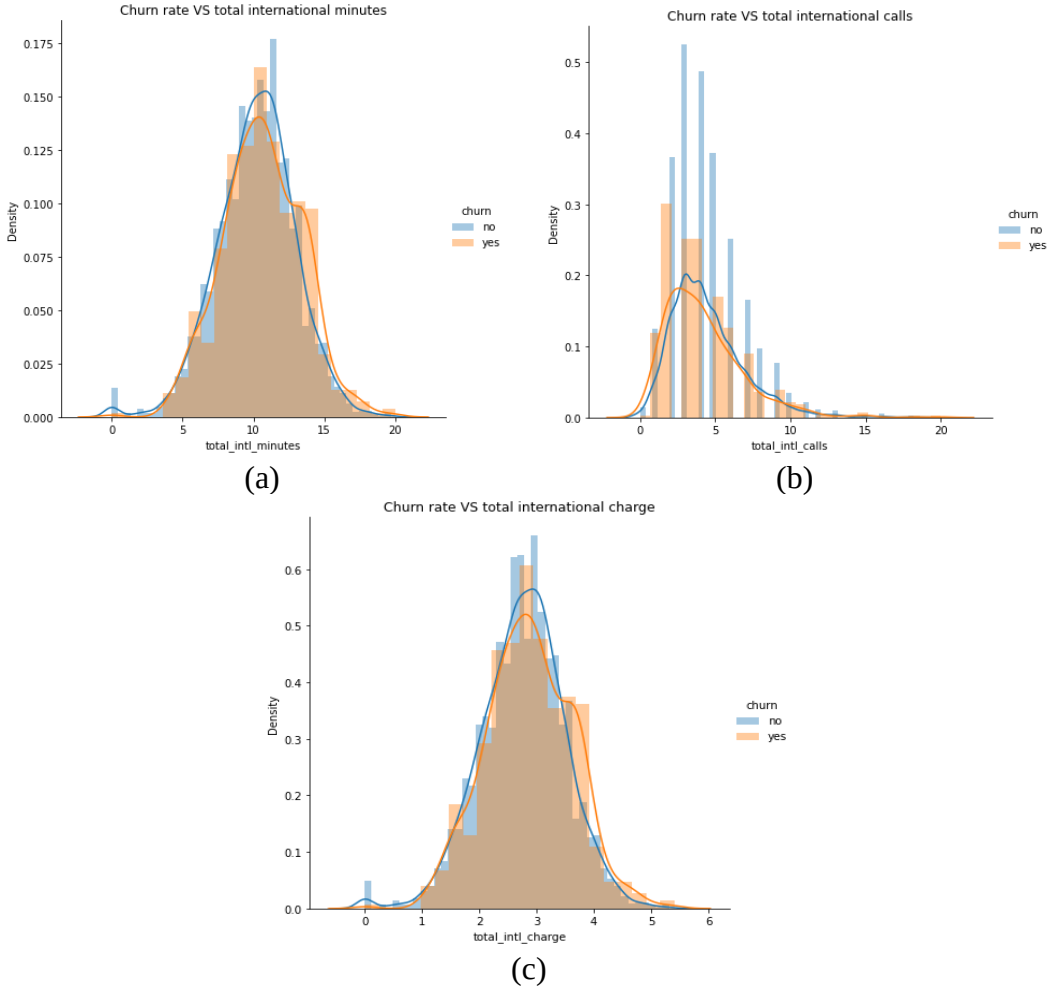


Figure 5. (a) Display of international call minutes and customer churn; (b) Display of international number of calls and customer churn; (c) Display of international call charges and customer churn

Figure 5 shows the impact of minutes, number of calls and call charges relative to total customer churn for customers subscribed to an international plan. Customers who subscribed to the international plan were concentrated around 12 minutes, the number of calls was concentrated at 3 and the call charges were concentrated at 3. Under the premise that all these subscribers could generate commonality, a relatively high level of customer retention was obtained, which also resulted in the highest indicator of customer churn. This illustrates the existence of a form of polarisation in the scheme. Customers who subscribed to the international plan were concentrated around 12 minutes, the number of calls was concentrated at 3 and the call charges were concentrated at 3. Under the premise that all these subscribers could generate commonality, a relatively high level of customer retention was obtained, which also resulted in the highest indicator of customer churn. This illustrates the existence of a form of polarisation in the scheme.

2.2. Algorithm

2.2.1 Logistic Regression

Although it uses a binomial response variable, logistic regression functions very similarly to linear regression [8]. The expression of the logistic function is as follows:

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

$$z = w^T x + w_0 \tag{2}$$

Logistic regression is a linear classifier that maps data features to a probability value (the likelihood of a sample belonging to a positive example) in the interval 0 to 1 by means of a logistic function (Sigmoid function) and derives the classification to which the data belongs by comparing it with 0.5. The expression in Equation 1 is the simplest one, which is an expression that tests whether the linear relationship between the dependent and independent variables is significant.

In the logical equation shown above x is the feature vector of the data$(x_0, x_1 \dots \dots x_n)$,$w$ is a parameter vector (requires data fitting),w_0 is offset.

Suppose there are a number of data points, and a line is used to fit these points (the line is called a best-fit line), this fitting process is called regression. The main idea behind user use of logistic regression for classification is to create a regression approach to the boundaries of the classification based on the available data, and in this way to classify.

The process of logistic regression is as follows:

1. Collecting data: Obtain task-related data by arbitrary means.
2. Preparing the data: Since distance calculations need to be performed, the data type required is generally of numeric type.
3. Analyse the data: the data is analysed using any method.
4. Training the algorithm: most of the resources will be used for training, the aim of which is to find the best coefficients for the classification regression.
5. Testing algorithm
6. Using the algorithm: first a series of data are input and converted into corresponding structured values; then a simple regression calculation is performed on these values for the trained regression coefficients to determine which category they belong to; after this some specific analysis can be done on the output category.

2.2.2 Decision Tree

A decision tree is a supervised learning technique used to summarize decision rules from a collection of categorised and labelled data and display these rules in a tree diagram structure to address classification and regression issues. A graph structure is the basis of a decision tree. The development of prediction algorithms for a target variable or the establishment of classification systems based on numerous variables are two prominent applications of the decision tree technique [9].

Which characteristic on top of the existing data set does the most important work in classifying the data? is the first issue to be solved while building a decision tree. Each characteristic must be assessed in order to identify the key feature and categorize the best outcome. The original dataset is split into numerous data subsets when the test is finished. The branches of the initial decision point will each receive a different distribution of these data subsets. There is currently no need for further dataset partitioning if the data under a branch is of the same kind. If the data within a data subset does not belong to the same type, a process of repartitioning the data subset is required. This is done until all data of the same type are in the same data subset.

The goal of dividing data is to provide greater order to disorganized data. Information gain refers to the difference between the data set before and after splitting. The ideal option is to obtain the feature with the maximum information gain. Once the information gain has been calculated, it is possible to determine the information gain achieved by splitting the data for each feature value.

Entropy, which is referred to as the measure of information in a set, is characterized as the anticipated value of the set. If the transaction k may be classified in more than one way, the information might be characterized as follows:

$$I(x_j) = -\log_{2^P}(x_i) \tag{3}$$

where $p(x_i)$ is the probability of selecting that classification.

The following formula calculates the expectation of information contained in all possible values of the category:

$$H = -\sum_{i=1}^{n} P(x_i) \log_2 p(x_i) \tag{4}$$

where n is the number of categories.

Now that entropy is available as a measure of the set of training samples, it is now possible to define a metric for the effectiveness of attribute classification training data. This criterion is the "information gain". Now imagine that attribute A is used to partition the training tuple D. The following results are anticipated after dividing A into D:

$$I_A(D) = -\Sigma_{j=1}^{v} \frac{D\vec{i}}{D} I(D_j) \tag{5}$$

The difference between the two values is the information gain.

2.2.3 SVM

Support Vector Machine is one of the classical machine learning techniques that can still help solve big data classification problems. Especially, it can help the multidomain applications in a big data environment [10]. To understand SVM it is necessary to first know that the lines separating the data sets are called separating hyperplane. Using this approach to construct the classifier, if the data points are further away from the decision boundary, then the more reliable the final prediction will be. Therefore, it is crucial to identify the points that are closest to the separating plane and make sure they are as far away from it as possible.

The separating hyperplane is closest to the support vector. The next step is to try to maximise the distance from the support vector to the separating plane, and to find an optimal solution to this problem. A support vector machine classifier searches for the best of all possible linear classifiers based on the distribution of the training samples, i. E. finding an optimal hyperplane.

It is challenging to completely prevent dimensional disasters since the computational complexity of SVM depends more on the quantity of support vectors than the dimensionality of the sample space. Additionally, since the classification of an SVM is mostly based on the support vectors, the absence of support vectors in a sample set has little impact on the final classification.

SVM is solved using quadratic programming, so it requires more storage space during the solving process. This leads to the disadvantage that SVM is suitable for small samples of data, while it will not work very well for larger scale data.

3. Results

All the results of the situation and the actual situation are mixed with each other, which results in 4 situations that make up the confusion matrix. A confusion matrix is a table of similar truth tables applied to a binary classification model. The actual indicator of this model is in fact whether the customer is churned or not. To judge whether the conditions are correct, it is also possible to make a mixed statistic of customer churn and customer not churned with the real value, so that can easily visualise the accuracy of the classification model.

Table 1. All scoring indicators for the three models.

Item	Logistic Regression	Decision Tree	SVM
Accuracy Score	0.858	0.945	0.848
Precision Score	0.858	0.945	0.848
Recall Score	0.178	0.837	0.0
f1 Score	0.275	0.821	0.0
cross val score	0.861	0.954	0.859

Accuracy is the percentage of the total sample that is predicted to be correct, but as prediction when the sample is unbalanced can easily result in classifying the data uniformly, resulting in a large negative sample for the high accuracy species, other metrics are needed to further aid prediction. The sample probability of the predicted outcome being positive among all the predicted outcomes at the time of precision. This largely solves the problem of uniformly large numbers of classifications in accuracy rates causing results to be inaccurate. Recall, on the other hand, is relative to the original sample, meaning the sample that happens to be predicted to be positive out of the sample that is actually positive. The F1 score is the result of considering both accuracy and completeness, by finding an equilibrium point in the prediction curve with the highest threshold to achieve both. The F1 score provides a harmonic mean of recall and precision, which is the average of them. As all of the precision and recall are rates, it makes logical sense to use the harmonic mean, meaning it will give equal weight to precision and recall. If both precision and recall are high, the F1 score will also be high, and if the F1 score is low, it means that both precision and recall are low [11].

It can be seen in table 1 that Decision Tree scores very high in all five-performance metrics, with precision and accuracy reaching over 90%, recall and f1 also scoring above 80%; Logistic Regression scores higher in precision and accuracy, reaching 80%, but recall and f1 are relatively low, both scoring over 20. SVM also scored over 80% for both exact and accurate, but only 0 for recall and f1, indicating that the SVM prediction model has very low or no coverage of the positive examples, resulting in a score of the two scores are zero.

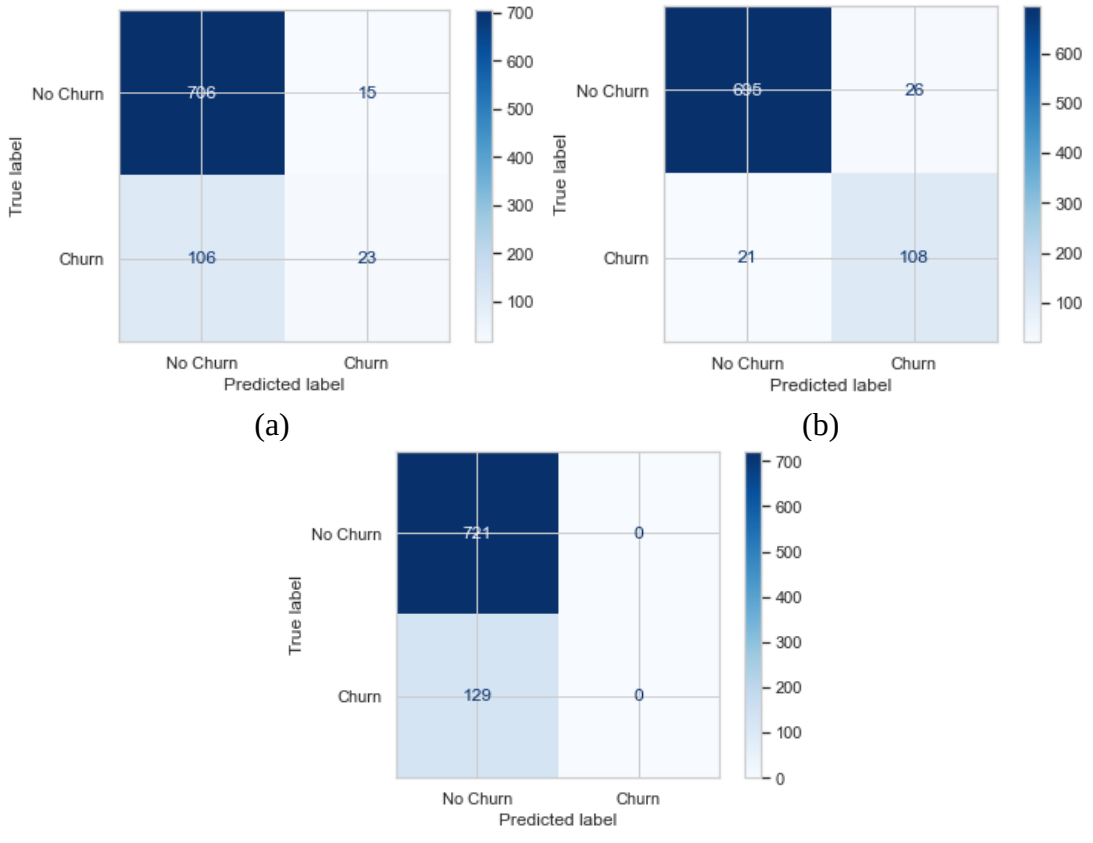


Figure 6. (a) Confusion matrix diagram for logistic regression; (b) Confusion matrix diagram for Decision Tree; (c) Confusion matrix diagram for SVM

In addition to confusion matrices, it is also possible to use actual numerical metrics for relevant reviews, using indicators such as accuracy, precision, recall, F1-score. The confusion matrix of the three classification models can be seen in figure 6 that the logistic regression is relatively accurate in predicting churn, with the exception of recall, cross val and f1, which have relatively low scores, with all scores above 80%.

Compared to decision trees, which are more accurate for predicting churn and non-churn of customers combined, they are more accurate for predicting churn than a classification model like logistic regression. The combined score between the two is also very high, with all metrics scoring above 80 per cent, with accuracy and cross val reaching 90 per cent.

SVM is definitely the most accurate in predicting one indicator of churn or not, predicting the data all as no churn and not accurately predicting customers relative to those who have already churned, which is a more extreme method of classification. The accuracy problem is exposed in the SVM algorithm, which predicts all outcomes as no churn, which in turn results in a large proportion of the predicted outcomes actually being in fact churned but will divide to outcomes where customers are not churned. This prediction can actually be very fatal for a company. The clear difference between SVM and the other two classification methods can be clearly seen by the recall and accuracy scores. It may have a high accuracy rate but the data in it can easily confuse data analysts. This problem can also be identified by the results of the confusion matrix.

Combining the results of these three classification models, logistic regression has been exposed as a problem in dealing with two indicators of customer churn, it does not handle a large number of multi-class features or variables well, here the churn prediction only involves two indicators, but the graph shows that for the churn prediction results, the correct data is not as good as the correct rate of its prediction of a single indicator. The decision tree, with its intuitive decision rules and its ability to handle many non-linear features, gives excellent results in the comparison of prediction results. The addition of the random forest also overcomes the problem of tedious pruning of decision trees in data fitting and takes full account of the interactions between variables. The SVM classification method is far inferior to the previous two classification methods because of the reliance on data boundaries in the prediction of customer churn which results in a large amount of data being confounded.

4. Conclusion

Through data analysis of the entire project using a number of different methods, classification models forecasting and other steps, it became intuitively clear that the most important indicator of churn is whether or not a user has subscribed to the company's international or voice messaging plans, and that once a user has subscribed to company’s product service, the customer is less likely to abandon it. But a deeper analysis of these two consumer plan services can also reveal many questions about how to solve the problem of subscribing to a plan that also causes the greatest churn rate above the point of greatest commonality of indicators. This will require the telcos to strengthen the quality of service of the consumer service plans and strive to retain as many of the top subscribers as possible, which will slowly solve the problem of very high customer churn. In addition to subscription product service plans, the impact on churn is also related to the amount of money a subscriber spends, which is a very simple fact, as is the case with subscription service plans. The more money a consumer spends on the services offered by this provider, the less likely the consumer is to churn. So, for a telecoms company that wants to retain the majority of its customers, to solve the problem of churn, the telecoms company can keep proper track of customer spending and get them to pay for this company's services whenever possible, and once they have spent the most enough money in them, they are much less likely to churn. The more indicators there are, the more likely it is that the company will be able to analyse what is wrong with the services it provides, and also to collect research on the needs of its customers, so that it can continue to provide services and make progress, so that it will lose fewer customers and bring greater benefits to the company.

In many contemporary situations, many of the dilemmas faced by the telecommunications industry are related to customer churn, the essence of the problem being the loyalty of subscribers and the loss of resources of the companies behind them. Telecom companies have to acknowledge the problem of customer churn because it is a certainty for a large industry that is not monopolised, and in the communications sector, it has to exist in order to generate pressure for the industry as a whole to collectively focus on building its own services. Churn cannot be solved only avoided as far as possible.

The solution is to retain the existing customer base through predictive models of customer churn. Data analytics technology helps the telecoms industry to solve this problem by using visual displays and classification models to help provide additional ideas and solutions to the telecoms industry.

References

[1] Qureshi, S. A., Rehman, A. S., Qamar, A. M., Kamal, A., & Rehman, A., "Telecommunication subscribers' churn prediction model using machine learning.", Eighth international conference on digital information management (ICDIM 2013). IEEE, (2013).

[2] Gopal, P., & MohdNawi, N. B., "A survey on customer churn prediction using machine learning techniques.", International Journal of Computer Applications 975 (2016).

[3] Hashmi, N., Butt, N. A., & Iqbal, M., "Customer churn prediction in telecommunication a decade review and classification." *International Journal of Computer Science Issues (IJCSI)* 10.5 (2013).

[4] Rajeswari, P. S., and P. Ravilochanan. "Churn analytics on Indian prepaid mobile services." *Asian Social Science* 10.13 (2014): 169.

[5] Landis T. and Philips S., "Customer retention marketing vs. customer acquisition marketing", Apr (2022). [Online]. Available: https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing/

[6] Ahmad, A. K., Jafar, A., & Aljoumaa, K., "Customer churn prediction in telecom using machine learning in big data platform", Journal of Big Data 6.1 (2019).

[7] Umayaparvathi, V., and K. Iyakutti. "A survey on customer churn prediction in telecom industry: Datasets, methods and metrics." *International Research Journal of Engineering and Technology (IRJET)* 3.04 (2016).

[8] Sperandei, S., "Understanding logistic regression analysis." *Biochemia medica* 24.1 (2014).

[9] Song, Y. Y., & Ying, L. U., "Decision tree methods: applications for classification and prediction.", Shanghai archives of psychiatry 27.2 (2015)

[10] Suthaharan, S. "Machine learning models and algorithms for big data classification." Integr. Ser. Inf. Syst 36 (2016).

[11] Aldalan, A. M., & Almaleh, A., "Customer Churn Prediction Using Four Machine Learning Algorithms Integrating Feature Selection and Normalization in the Telecom Sector." International Journal of Electronics and Communication Engineering 17.3 (2023).