

TOPIC NAME:

**“THE COMPARISON OF LOGISTICS REGRESSION, SVM,
RANDOM FOREST, XGBOOST AND CATBOOST MODELS TO
PREDICT CUSTOMERS TELECOMMUNICATIONS CHURN
AFTER CLUSTERING”**

CONTENT

CHAPTER 1: INTRODUCE.....	8
CHAPTER 2: RELATED WORK	10
2.1. Background of research	10
2.1.1. Foreign research.....	10
2.1.2. National research	14
2.2. <i>Gain relevant research results and model suggestions</i>	14
CHAPTER 3: THEORETICAL BASIC	15
3.1. Data mining.....	15
3.1.1. Definition.....	15
3.1.2. Classification techniques in Data mining	15
3.2. Churn rate definition	16
3.2.1. Churn rate definition.....	16
3.2.2. Imbalanced dataset.....	16
3.3. Customer segmentation.....	17
3.3.1. Definition	17
3.3.2. K-means clustering	18
3.3.3. Elbow method.....	20
3.4. Imbalanced dataset	20
3.4.1. Definition	20
3.4.2. Feature	21
3.5. Machine learning.....	21
3.5.1. Logistic regression.....	21
3.5.2. Random Forest.....	22
3.5.3. SVM.....	24
3.5.4. XGBoost	24

3.5.5. CatBoost.....	25
3.6. Evaluation measure	26
3.6.1. Confusion matrix	26
3.6.2. ROC Curve and AUC	27
CHAPTER 4: PROPOSED METHODOLOGY	31
CHAPTER 5: RESULTS AND DISCUSSION	32
5.1. Data understanding.....	32
5.1.1. Dataset	32
5.1.2. Independent variables	36
5.1.3. Dependent variable	36
5.2. Data Pre-processing	41
5.3. Customer segmentation with K-Means.....	46
5.4. SMOTE + SPLIT Data.....	48
5.5. Features selection - VIF	48
5.6. Machine learning method.....	50
5.6.1. Before customer segmentation with KMeans.....	50
5.6.2. After customer segmentation with KMeans	51
CHAPTER 6: CONCLUSION	56
6.1. General conclusion.....	56
6.2. Limited and future work:	56
<i>Reference</i>	57

LIST OF FIGURES

Figure 1. Example of Elbow method.....	20
Figure 2. Random Forest overfitting example.....	22
Figure 3. Random Forest model process	23
Figure 4. Comparison of CatBoost model to others model with many dataset.....	26
Figure 5. TP vs. FP rate at different classification thresholds.....	28
Figure 6. AUC (Area under the ROC Curve).....	29
Figure 7. Predictions ranked in ascending order of logistic regression score	29
Figure 8. Our proposed methodology	31
Figure 9. TotalCharges with “0” tenure value.....	35
Figure 10. Ratio Churn variable in our dataset	36
Figure 11. Gender, SeniorCitizen, Partner, Dependents variables visualization	37
Figure 12. MultipleLines, PhoneService, InternetService, OnlineSecurity variables visualization.....	38
Figure 13. OnlineBackup, DeviceProtection, TechSupport, StreamingTV variables visualization.....	39
Figure 14. StreamingMovies, PaperlessBilling, Contract, PaymentMethod variables visualization.....	40
Figure 15. Numerical variables visualization.....	40
Figure 16. Distribution of numerical variables before and after handling data transformation process.....	45
Figure 17. Distribution of the selected numerical variables.....	46
Figure 18. Number of cluster.....	47
Figure 19. The ROC curve of these 5 models before segmentation.....	50

LIST OF TABLES

Table 1. The detail information of the telecom dataset	34
Table 2. Correlation between independent variable with dependent variables	44
Table 3. Data transformation with numerical variables	45
Table 4. Number of customer after segmentation	47
Table 5. Number of customer after segmentation and SMOTE.....	48
Table 6. Selected variables for each cluster	50
Table 7. Confusion matrix (F1-Score) of these 5 models before customer segmentation	51
Table 8. Confusion matrix (actual, predicted) after customer segmentation	54
Table 9. Confusion matrix (F1-Score) after customer segmentation	55

CHAPTER 1: INTRODUCE

With the rapid development of data communication networks and the advancement in information. A rather difficult problem to be solved that every organization faces in a fiercely competitive environment is how to predict their customers are likely to leave the company. Churn Management is getting a lot of attention because it has proven that customer retention is more important than customers buying new products because the main source of profit of the business is from loyal customers so continued Approaching customers plays an important role in the existence and development of a certain field.

Especially in the field of telecommunications, Churn prediction has become one of the main tasks to be performed. The telecommunications industry is facing fierce competition due to rapid changes in technology and increasing potential competitors. Customers have enough alternatives among many telecommunications service providers and easily make the decision to switch without any difficulty. Therefore, customers will always be looking for better service providers with more suitable costs.

Realizing the importance of building Churn predictive models to identify customers who are likely to leave the service. For companies providing telecommunications services, the loss of customers can be seen as their customers switching to competitors' services, which causes huge losses to the business's revenue. Therefore, Churn prediction helps the company to come up with more effective plans and strategies suitable to the customers that the business aims to retain, helping businesses reduce the churn rate. provide a stable source of profit for the business.

For example, in Customer Relationship Management, if a business can retain 5% of old customers, its profit can increase by 25% (Peng Li, et al., 2014). Furthermore, it costs less to retain existing customers than to acquire new customers (Robert, 2000). Besides, when a satisfied customer can bring more potential customers to the business. While 1 unsatisfied customer will also affect many customers who intend to use the company's services.

Over the years, many Churn Prediction models have been launched that use different data mining techniques and algorithms to find ways to solve data imbalances to predict customer churn. The use of machine learning has become a popular idea in today's digital transformation era where there is an application to predict churn. Machine learning has powerful techniques for learning past observations to make accurate

predictions. Today, the activities related to predicting the customer's Churn have become more complex, many scopes have been extended to help find the right and effective model that contributes to the promotion and development of activities. business of the enterprise. We compared and proposed the most optimal algorithm for our dataset in the field of telecommunications.

We have implemented the project from October 22, 2022 to December 16, 2022 using a dataset of churn customers of telecommunications services companies, with 7043 customers (7043 rows) and 21 columns (21 attributes). This report aims to provide a complete solution for businesses to the problem of inversion prediction, so that they can take preventive measures to provide services that are suitable for the customers they want to target. next. In this study, the development exploits different methods to more accurately predict whether a customer is likely to leave the service in the near future. These models help businesses prioritize the right policies that have a high impact on customer choice. Help businesses make the right decisions for customer retention. The main purpose of the research is to build a reliable and accurate model that can optimize the retention of old customers.

The remaining layout of the report is organized into 5 sections. Section 2 is Related Work. Section 3 is Theoretical basic, including introduction of K-means Elbow method, SMOTE technique and 4 models Logistic Regression, Random forest, SVM, XGBoost, Catboost that the team selects for analysis. Section 4 is a proposed methodology, including data preprocessing, identifying customer segments by K-Means, Using Elbow method to find the number of phrases, SMOTE technique to solve loss problem data balancing, data file splitting process and using machine learning to analyze. Section 5 is Result and discussion. Section 6 is the conclusion.

CHAPTER 2: RELATED WORK

2.1. Background of research

2.1.1. Foreign research

2.1.1.1. *Dr.M.Balasubramanian, M.Selvarani (2014). Churn prediction in mobile telecom system using data mining techniques*

This study focuses on data mining techniques for reducing customer churn and likely reducing the error ratio. From this study, it is observed that the decision tree model surpasses the neural network model in the prediction of churn and it is also easy to construct. Selecting the right combination of attributes and fixing the proper threshold values may produce more accurate results.

2.1.1.2. *Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry*

In this study, they develop a churn prediction model tailored for the B2B e-commerce industry by testing the capability of a new model, the support vector machine (SVMauc) based on the AUC parameter-selection technique. The predictive performance of SVMauc is benchmarked to logistic regression, neural net, and classic support vector machines.

2.1.1.3. *Ammar A Ahmed, Dr. D. Maheswari linen (2017). A review and analysis of churn prediction methods for customer retention in telecom industries*

This paper focuses on analyzing the churn prediction techniques to identify churn behavior and validate the reasons for customer churn. This paper summarizes the churn prediction techniques in order to have a deeper understanding of customer churn and it shows that the most accurate churn prediction is given by the hybrid models rather than single algorithms so that telecom industries become aware of the needs of high-risk customers and enhance their services to overturn the churn decision.

2.1.1.4. Abhishek Gaur, Ratnesh Dubey (2018). Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques

In this paper, they focus on various machine learning techniques for predicting customer churn through which build classification models such as Logistic Regression, SVM, Random Forest, and Gradient boosted tree. Gradient boosting is best in all four models and the Logistic regression and Random Forest is an average and SVM is underperforming among the models.

2.1.1.5. Abhishek Gaur, Ratnesh Dubey (2021). Customer churn prediction

This project focuses on various machine learning techniques for predicting customer churn through which build the classification models such as Logistic Regression, Random Forest, and lazy learning. Evaluating the models using appropriate evaluation metrics like accuracy, precision, and recall is more important to identify churners accurately. Draw ROC graph, confusion matrix, and classification report for the Random Forest model which is predicted as the best.

2.1.1.6. K. Sandhya Rani , Shaik Thaslima , N.G.L. Prasanna , R.Vindhya , and P. Srilakshmi (2021). Analysis of Customer Churn Prediction in Telecom Industry Using Logistic Regression

This study identifies the risk factor by using the previous data. Logistic regression algorithms are used for efficacy in prediction results. The system can only predict by using limited features. This system is used to reduce the time and burden on the company by going through the historical data and news of the company to assess and react to the situations.

2.1.1.7. Lawchak Fadhil Khalid, Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree, Falah Y. H. Ahmed (2021). Customer churn prediction in telecommunications industry based on data mining

The different prediction models in this paper compared the quality measurements of prediction models such as classification algorithms and decision trees. They discovered that the accuracy achieved with the decision tree is higher (3 percent higher than the second result and 6 percent higher than the lowest-achieving algorithm) than

the other techniques used, indicating that the decision tree is an effective technique for churn prediction.

2.1.1.8. Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques

Support vector machines (SVMs) are able to predict churn in subscription services by mapping non-linear inputs into a high-dimensional feature space. In this study, SVMs outperform traditional logistic regression and random forests when the optimal parameter selection procedure is applied. The way in which the optimal parameters are selected can have significant influences on the performance of an SVM. In this study, the most important churn predictors are part of the group of variables describing the subscription. In spite of the importance of age, socio-demographics do not play an important role in explaining churn in this study. There is no complete working meta-theory to assist with the selection of the correct kernel function and SVM parameters for a specific type of problem.

2.1.1.9. Saw Thazin Khine, Win Win Myo (2019). Customer Churn Analysis in Banking Sector

In this research, the churn prediction model for classifying bank customers is built by using the hybrid model of k-means and Support Vector Machine data mining methods on the bank customer churn dataset to overcome the instability and limitations of a single prediction model and predict the churn trend of high-value users. This model also supports information about similar customer groups to consider which marketing reactions are to be provided. Thus, due to existing customers being retained, it will provide banks with increased profits and revenues. And also proposed combined model K-means-SVM reduces SVM training time by reducing support vectors using the K-means clustering algorithm.

2.1.1.10. Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019). Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry.

The use of data mining is proven to be used in predicting churn in the banking business. Support Vector Machine (SVM) with a comparison of 50:50 Class sampling

data is the best method for predicting churn customers at a private bank in Indonesia. This is in line with the research of Dolatabadi et al. which obtained SVM as modeling with the best accuracy.

2.1.1.11. Jan Mand'ák, Jana Hančlová (2019). Use of Logistic Regression for Understanding and Prediction of Customer Churn in Telecommunications

AUC, sensitivity, and accuracy are three additional metrics that were produced to evaluate the model's quality on an independent test data set. The computed model's high AUC (0.9759) and sensitivity (0.948) values confirmed its ability to predict clients who want to leave. The logistic regression model was able to accurately predict 94.8% of consumers who in reality did leave the firm, according to the value of sensitivity. In conclusion, logistic regression was effectively used in the telecom industry to identify consumers who were likely to quit the business as well as the primary causes of churn.

2.1.1.12. Xinyu Miao, Haoran Wang (2022). Customer Churn Prediction on Credit Card Services using Random Forest Method

This study uses machine learning techniques to anticipate user turnover for credit cards and, using the predictions, offers workable strategies to address the problem. Applying three models—Random Forest, Linear Regression, and K-Nearest Neighbor (KNN)—to a dataset with more than 10,000 pieces and 21 features is possible. The results of adjusting hyperparameters and comparing models using the ROC & AUC and confusion matrix show that Random Forest performs the best, with an accuracy of 96.25%.

2.1.1.13. Heng Zhao, Xumin Zuo, Yuanyuan Xie (2022). Customer Churn Prediction by Classification Models in Machine Learning

This study uses sales data from a chemical firm from 2012 to 2020 to estimate customer attrition using decision tree and random forest models. They examine the underlying risk of client attrition as well as the contributing causes. The comparison's findings show that the random forest model outperforms the decision tree model in terms of prediction accuracy when training error and generalization error are taken into account. These variables include the confusion matrix, ROC curve, AUC, precision, recall, and F1 score.

2.1.2. National research

2.1.1.14. Nguyen Thi Ha Thanh, Nguyen Thao Vy (2022) Building a proper churn prediction model for Vietnam's mobile banking service

This study aims to build a model predicting the churn rate of customers using mobile banking services in Vietnam by applying data mining techniques. The results exhibited that Gradient Boosting is the best performance in the three above classifiers with a 79.71% of accuracy rate, and an 86.23% of ROC (Receiver Operating Characteristic) curve graph.

2.2. Gain relevant research results and model suggestions

In service industries with intense market competition, losing customers is a significant problem. The analysis of consumers who are most likely to abandon the business, on the other hand, might provide a sizable new income stream if it is done early on. Numerous types of research have demonstrated that machine-learning approaches are quite effective and beneficial for forecasting both churning and non-churning occurrences by learning from past company data. The information utilized in this includes all past and present client data. In this experiment, machine learning approaches and algorithms for predicting churn in the telecom sectors are mostly focused on Logistics regression, Random Forest, SVM, XGBoost, and Catboost. Good practices must be created and built upon in order to prevent issues in the telecommunications sector. In this study, we discussed several prediction models and compared quality assessments of prediction models, such as a confusion matrix, ROC, and AUC that illustrates and summarizes a classification algorithm's performance.

CHAPTER 3: THEORETICAL BASIC

3.1. Data mining

3.1.1. *Definition*

Data mining refers to digging into or mining the data in different ways to identify patterns and get more insights into them. It involves analyzing the discovered patterns to see how they can be used effectively.

In data mining, you sort large data sets, find the required patterns and establish relationships to perform data analysis. It's one of the pivotal steps in data analytics, and without it, you can't complete a data analysis process.

In data mining, the predictive analysis task is undertaken through classification and regression techniques. Regression is a statistical method that is used to estimate relationships between dependent variables to one or more independent variables. It can also be used to assess the strength of the relationship between variables as well as to model future relationships between them, whereas classification is a predictive modeling problem where a class label is predicted for input data. They instigated classification as a procedure to find a model that demonstrates and identifies data concepts or classes. Afterward, the model has been used for predicting class labels of objects with unidentified labels.

3.1.2. *Classification techniques in Data mining*

Classification in data mining is a common technique that separates data points into different classes. It allows you to organize data sets of all sorts, including complex and large datasets as well as small and simple ones.

It primarily involves using algorithms that you can easily modify to improve the data quality. This is a big reason why supervised learning is particularly common with classification in techniques in data mining. The primary goal of classification is to connect a variable of interest with the required variables. The variable of interest should be of qualitative type.

The algorithm establishes the link between the variables for prediction. The algorithm you use for classification in data mining is called the classifier, and

observations you make through the same are called the instances. You use classification techniques in data mining when you must work with qualitative variables.

There are multiple types of classification algorithms, each with its unique functionality and application. In this, we used the data mining algorithms such as, Logistic Regression Classifier, Random Forest Classifier, SVM (Support Vector Machines, etc.).

3.2. Churn rate definition

3.2.1. Churn rate definition

The churn rate is a ratio based on the amount of customer churn and the total number of existing customers for a given period of time. To calculate the churn rate quite simply, we divide the total number of customers churn by the total number of customers at the beginning of a month/quarter/year.

- Customer churn: The total number of customers who stop contracting or using a company's services over a period of time like a month/quarter/year.
- Total customers: The total number of existing customers for that period.

Why is churn rate important: In reality, it's much more expensive to attract new customers than to just care and hold on to existing customers.

Churn rate = $\frac{\text{Number of users at the beginning} - \text{Number of user at the end}}{\text{Number of users at the beginning}} \times 100\%$.

If the churn rate is lower, the business is at a safe level. Conversely, if the churn rate is medium or high, the business is in the alarm zone.

3.2.2. Imbalanced dataset

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations.

Let's assume that XYZ is a bank that issues a credit card to its customers. Now the bank is concerned that some fraudulent transactions are going on and when the bank checks their data they found that for each 2000 transactions there are only 30 Nos of fraud recorded. So, the number of frauds per 100 transactions is less than 2%, or we can

say more than 98% transactions are “No Fraud” in nature. Here, the class “No Fraud” is called the majority class, and the much smaller in size “Fraud” class is called the minority class.

More such example of imbalanced data is:

- Disease diagnosis
- Customer churn prediction
- Fraud detection
- Natural disaster

Class imbalance is generally normal in classification problems. But, in some cases, this imbalance is quite acute where the majority class’s presence is much higher than the minority class.

3.3. Customer segmentation

3.3.1. Definition

Customer segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately.

In business-to-business marketing, a company might segment customers according to a wide range of factors, including:

- Industry
- Number of employees
- Products previously purchased from the company
- Location

In business-to-consumer marketing, companies often segment customers according to demographics that include:

- Age
- Gender
- Marital status
- Location (urban, suburban, rural)
- Life stage (single, married, divorced, empty-nester, retired, etc.)

Segmentation allows marketers to better tailor their marketing efforts to various audience subsets. Those efforts can relate to both communications and product development. Specifically, segmentation helps a company:

- Create and communicate targeted marketing messages that will resonate with specific groups of customers, but not with others (who will receive messages tailored to their needs and interests, instead).
- Select the best communication channel for the segment, which might be email, social media posts, radio advertising, or another approach, depending on the segment.
- Identify ways to improve products or new product or service opportunities.
- Establish better customer relationships.
- Test pricing options.
- Focus on the most profitable customers.
- Improve customer service.
- Upsell and cross-sell other products and services.

3.3.2. K-means clustering

K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems. It follows a simple procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand. The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached.

K-means clustering has uses in search engines, market segmentation, statistics and even astronomy.

The way k means algorithm works is as follows:

1. Specify number of clusters K.
 2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
 3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
- Compute the sum of the squared distance between data points and all centroids.

- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

The objective function is:

$$J = \sum_{i=1}^m \sum_{k=1}^K w_{ik} \|x^i - \mu_k\|^2 \quad (1)$$

where $w_{ik}=1$ for data point x_i if it belongs to cluster k ; otherwise, $w_{ik}=0$. Also, μ_k is the centroid of x_i 's cluster.

It's a minimization problem of two parts. We first minimize J w.r.t. w_{ik} and treat μ_k fixed. Then we minimize J w.r.t. μ_k and treat w_{ik} fixed. Technically speaking, we differentiate J w.r.t. w_{ik} first and update cluster assignments (E-step). Then we differentiate J w.r.t. μ_k and recompute the centroids after the cluster assignments from the previous step (M-step). Therefore, E-step is:

$$\begin{aligned} \frac{\partial J}{\partial w_{ik}} &= \sum_{i=1}^m \sum_{k=1}^K \|x^i - \mu_k\|^2 \\ \Rightarrow w_{ik} &= \begin{cases} 1 & \text{if } k = \operatorname{argmin}_j \|x^i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (2)$$

In other words, assign the data point x_i to the closest cluster judged by its sum of squared distance from cluster's centroid.

And M-step is:

$$\begin{aligned} \frac{\partial J}{\partial \mu_k} &= 2 \sum_{i=1}^m w_{ik} (x^i - \mu_k) = 0 \\ \Rightarrow \mu_k &= \frac{\sum_{i=1}^m w_{ik} x^i}{\sum_{i=1}^m w_{ik}} \end{aligned} \quad (3)$$

Which translates to recomputing the centroid of each cluster to reflect the new assignments.

3.3.3. Elbow method

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. The same method can be used to choose the number of parameters in other data-driven models, such as the number of principal components to describe a data set.

Using the "elbow" or "knee of a curve" as a cutoff point is a common heuristic in mathematical optimization to choose a point where diminishing returns are no longer worth the additional cost. In clustering, this means one should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.

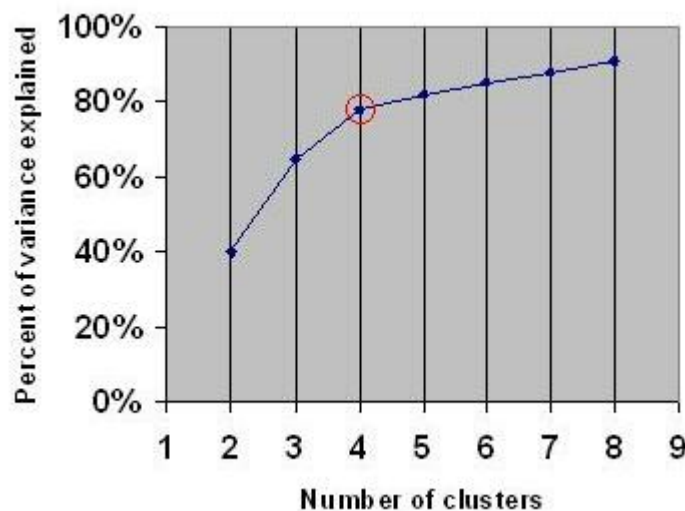


Figure 1. Example of Elbow method

The intuition is that increasing the number of clusters will naturally improve the fit (explain more of the variation), since there are more parameters (more clusters) to use, but that at some point this is over-fitting, and the elbow reflects this.

3.4. Imbalanced dataset

3.4.1. Definition

SMOTE is an algorithm that performs data augmentation by creating synthetic data points based on the original data points. SMOTE can be seen as an advanced version of oversampling, or as a specific algorithm for data augmentation. The advantage of

SMOTE is that you are not generating duplicates, but rather creating synthetic data points that are slightly different from the original data points. So, SMOTE is an improved alternative for oversampling.

3.4.2. Feature

The SMOTE algorithm works as follows:

- You draw a random sample from the minority class.
- For the observations in this sample, you will identify the k nearest neighbors.
- You will then take one of those neighbors and identify the vector between the current data point and the selected neighbor.
- You multiply the vector by a random number between 0 and 1.
- To obtain the synthetic data point, you add this to the current data point.

This operation is actually very much like slightly moving the data point in the direction of its neighbor. This way, you make sure that your synthetic data point is not an exact copy of an existing data point while making sure that it is also not too different from the known observations in your minority class.

3.5. Machine learning

3.5.1. Logistic regression

Logistic Regression is a classification algorithm used to assign objects to a discrete set of values (like 0, 1, 2, ...). A typical example is the classification of email, including work email, home email, spam email, ... Is online transaction safe or unsafe, benign or malignant tumors ...

Formula:

$$\hat{y}_i = \sigma(w_0 + w_1 * x_1^{(i)} + w_2 * x_2^{(i)}) = \frac{I}{1 + e^{-(w_0 + w_1 * x_1^{(i)} + w_2 * x_2^{(i)})}}$$

The task of Logistic Regression is to find the coefficients of the regression equation based on the input data set of independent variables and the output as the dependent variable.

Equation Logistic Regression with:

+ x_1, x_2 are input variables, based on the data set that we have

+ w_1, w_2 are the coefficients of the regression equation

+ y_i is the output, has the value from 0 to 1

- The output of logistic Regression gives 2 results True or False (0,1) instead of predicting something continuously like revenue, body size, ...

3.5.2. Random Forest

Decision Tree is a simple algorithm, easy to implement, but has the weakness that it tends to overfit during training. As shown in the figure below overfitting to the training data set will adversely affect the model.

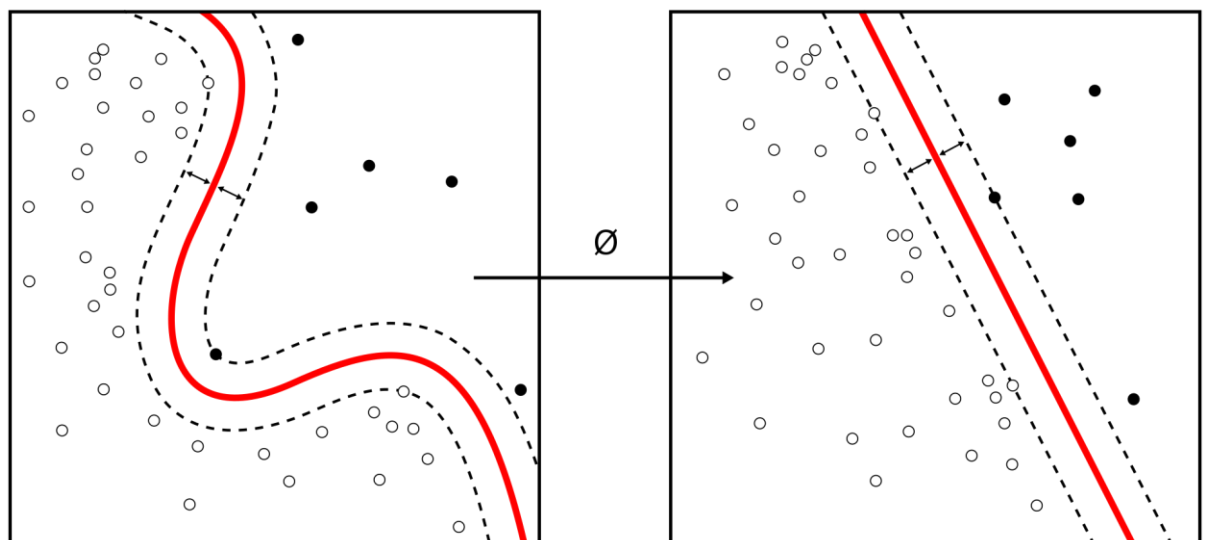


Figure 2. Random Forest overfitting example

Random forests take an ensemble approach that provides an improvement over the basic decision tree structure by combining a group of weak learners to form a stronger learner. Ensemble methods utilize a divide-and-conquer approach to improve algorithm performance. In random forests, a number of decision trees, i.e., weak learners, are built on bootstrapped training sets, and a random sample of m predictors are chosen as split candidates from the full set P predictors for each decision tree. As $m \rightarrow P$, the majority of the predictors are not considered. In this case, all of the individual trees are unlikely to be dominated by a few influential predictors. By taking the average

of these uncorrelated trees, a reduction in variance can be attained, making the final result less variable and more reliable.

To build the Random Forest algorithm, the individual trees are built on bootstrap samples rather than on the original sample. This is called bootstrap aggregating or simply bagging, and it reduces overfitting. The process is mentioned as follows:

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output. 10

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively (Sruthi E R, 2021).

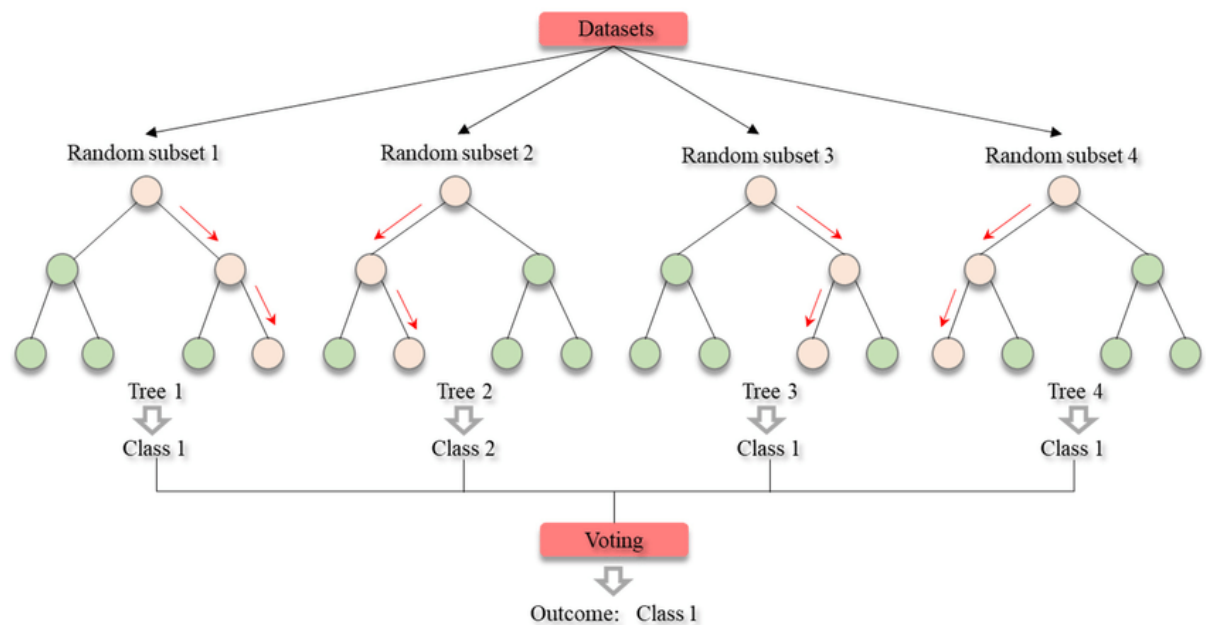


Figure 3. Random Forest model process

Workflow of the random forest algorithm

As mentioned above, random forest solves the problem of overfitting very well. Because the algorithm does not use all the training data to build the model. The final result of the algorithm is the summation of many different decision trees, so the information of one tree will complement the other, bringing good prediction results for the model.

3.5.3. SVM

Support vector machines (SVM) is a machine learning method raised by Vapnik in the early 1990s. The basic idea of SVM is to define a hyperplane which separates the n-dimensional data into two classes, wherein the hyperplane maximizes the geometric distance to the nearest data points, so-called support vectors. It is noteworthy that practical linear SVM often yields similar results as logistic regression.

In addition to performing linear classification, SVM also introduces the idea of a kernel method to efficiently perform non-linear classification. It is a feature mapping methodology which transfers the attributes into a new feature space (usually higher in dimension) where the data is separable.

In this study, we used this model and compared the prediction effect to other techniques. The SVM aims to find an optimal linearly separable SVM classification surface. The optimal classification surface requires the separating line not only can separate two class labels correctly but can also maximize the margin between the two classes.

3.5.4. XGBoost

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

The term “gradient boosting” comes from the idea of “boosting” or improving a single weak model by combining it with a number of other weak models in order to generate a collectively strong model. Gradient boosting is an extension of boosting where the process of additively generating weak models is formalized as a gradient descent algorithm over an objective function. Gradient boosting sets targeted outcomes for the next model in an effort to minimize errors. Targeted outcomes for each case are based on the gradient of the error (hence the name gradient boosting) with respect to the prediction.

GBDTs iteratively train an ensemble of shallow decision trees, with each iteration using the error residuals of the previous model to fit the next model. The final prediction is a weighted sum of all of the tree predictions. Random forest “bagging” minimizes the variance and overfitting, while GBDT “boosting” minimizes the bias and underfitting.

XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed. With XGBoost, trees are built in parallel, instead of sequentially like GBDT. It follows a level-wise strategy, scanning across gradient values and using these partial sums to evaluate the quality of splits at every possible split in the training set.

3.5.5. CatBoost

CatBoost is an open-source machine learning(gradient boosting) algorithm, with its name coined from “Category” and “Boosting.” It was developed by Yandex (Russian Google) in 2017. According to Yandex, CatBoost has been applied to a wide range of areas such as recommendation systems, search ranking, self-driving cars, forecasting, and virtual assistants. It provides a gradient boosting framework which among other features attempts to solve for Categorical features using a permutation driven alternative compared to the classical algorithm.

CatBoost has gained popularity compared to other gradient boosting algorithms primarily due to the following features:

- Native handling for categorical features
- Fast GPU training
- Visualizations and tools for model and feature analysis
- Using Oblivious Trees or Symmetric Trees for faster execution
- Ordered Boosting to overcome overfitting

CatBoost can work with diverse data types to help solve a wide range of problems that businesses face today. To top it up, it provides best-in-class accuracy.

It is especially powerful in two ways:

- It yields state-of-the-art results without extensive data training typically required by other machine learning methods, and

- Provides powerful out-of-the-box support for the more descriptive data formats that accompany many business problems.

“CatBoost” name comes from two words “Category” and “Boosting”.

Comparison to other boosting libraries:



	CatBoost		LightGBM		XGBoost		H2O	
	Tuned	Default	Tuned	Default	Tuned	Default	Tuned	Default
 Adult	0.26974	0.27298 +1.21%	0.27602 +2.33%	0.28716 +6.46%	0.27542 +2.11%	0.28009 +3.84%	0.27510 +1.99%	0.27607 +2.35%
 Amazon	0.13772	0.13811 +0.29%	0.16360 +18.80%	0.16716 +21.38%	0.16327 +18.56%	0.16536 +20.07%	0.16264 +18.10%	0.16950 +23.08%
 Click prediction	0.39090	0.39112 +0.06%	0.39633 +1.39%	0.39749 +1.69%	0.39624 +1.37%	0.39764 +1.73%	0.39759 +1.72%	0.39785 +1.78%
 KDD appetency	0.07151	0.07138 -0.19%	0.07179 +0.40%	0.07482 +4.63%	0.07176 +0.35%	0.07466 +4.41%	0.07246 +1.33%	0.07355 +2.86%
 KDD churn	0.23129	0.23193 +0.28%	0.23205 +0.33%	0.23565 +1.89%	0.23312 +0.80%	0.23369 +1.04%	0.23275 +0.64%	0.23287 +0.69%
 KDD internet	0.20875	0.22021 +5.49%	0.22315 +6.90%	0.23627 +13.19%	0.22532 +7.94%	0.23468 +12.43%	0.22209 +6.40%	0.24023 +15.09%

Figure 4. Comparison of CatBoost model to others model with many dataset

The comparison above shows the log-loss value for test data and it is lowest in the case of CatBoost in most cases. It clearly signifies that CatBoost mostly performs better for both tuned and default models.

In addition to this, CatBoost does not require conversion of data set to any specific format like XGBoost.

3.6. Evaluation measure

3.6.1. Confusion matrix

Confusion matrix is used for performance Measurement of the Classification Problem and where the output can be in two or more classes. Confusion matrix is a table with 4 different combinations of predicted and actual values.

The 4 different Combinations are:

- True Positives: Predicted positive and it's true
- True Negatives: Predicted Negative and it is true

- False Positive: Predicted positive and it's false
- False Negatives: Predicted Negative and it is false

Performance metrics of an algorithm are accuracy, precision, recall, and F1 score, which are calculated on the basis of the above-stated TP, TN, FP, and FN.

- Accuracy of an algorithm is represented as the ratio of correctly classified customers (TP+TN) to the total number of customers (TP+TN+FP+FN)

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

- Precision of an algorithm is represented as the ratio of correctly classified customers with the disease (TP) to the total customers predicted to have the disease (TP+FP).

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

- Recall metric is defined as the ratio of correctly classified diseased customers (TP) divided by total number of customers who actually have the disease.

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

The perception behind recall is how many customers have been classified as having the disease. Recall is also called sensitivity.

- F1 score is also known as the F Measure. The F1 score states the equilibrium between the precision and the recall

$$F1\ Score = \frac{2 * precision * recall}{precision + recall}$$

3.6.2. ROC Curve and AUC

ROC curve:

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

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

False Positive Rate (FPR) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

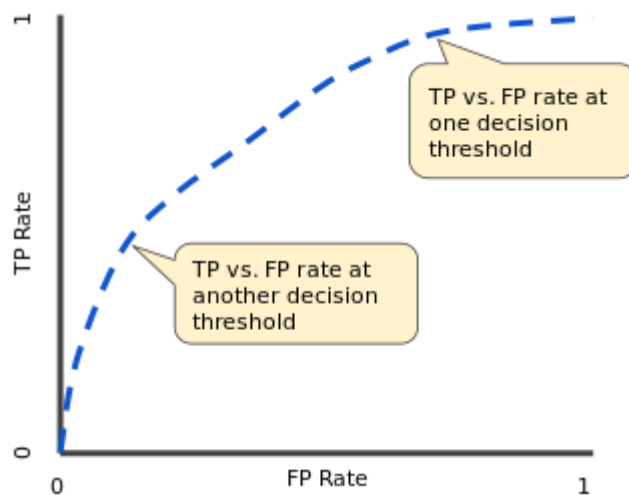


Figure 5: TP vs. FP rate at different classification thresholds.

To compute the points in an ROC curve, we could evaluate a logistic regression model many times with different classification thresholds, but this would be inefficient. Fortunately, there's an efficient, sorting-based algorithm that can provide this information for us, called AUC.

AUC: Area Under the ROC Curve

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

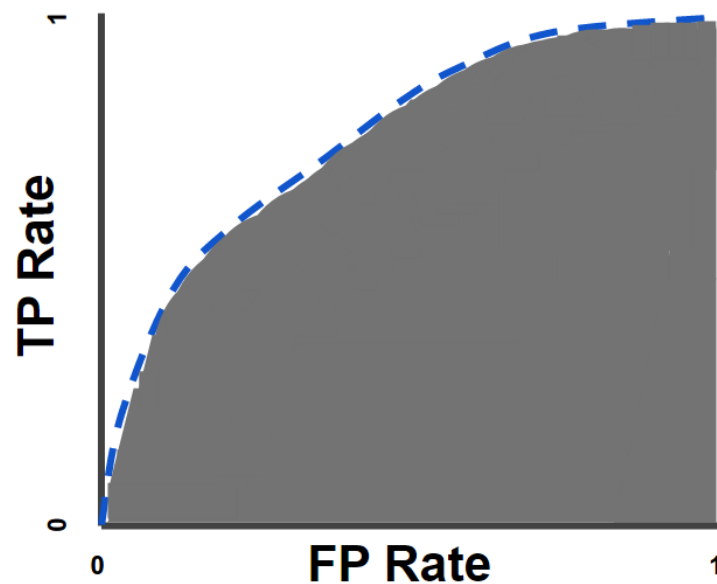


Figure 6. AUC (Area under the ROC Curve)

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. For example, given the following examples, which are arranged from left to right in ascending order of logistic regression predictions:

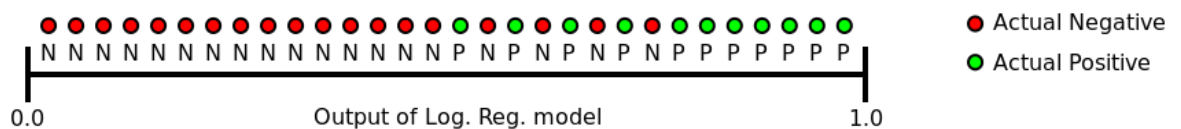


Figure 7. Predictions ranked in ascending order of logistic regression score

AUC represents the probability that a random positive (green) example is positioned to the right of a random negative (red) example.

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

AUC is desirable for the following two reasons:

- AUC is scale-invariant. It measures how well predictions are ranked, rather than their absolute values.
- AUC is classification-threshold-invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

However, both these reasons come with caveats, which may limit the usefulness of AUC in certain use cases:

- Scale invariance is not always desirable. For example, sometimes we really do need well calibrated probability outputs, and AUC won't tell us about that.
- Classification-threshold invariance is not always desirable. In cases where there are wide disparities in the cost of false negatives vs. false positives, it may be critical to minimize one type of classification error. For example, when doing email spam detection, you likely want to prioritize minimizing false positives (even if that results in a significant increase of false negatives). AUC isn't a useful metric for this type of optimization.

CHAPTER 4: PROPOSED METHODOLOGY

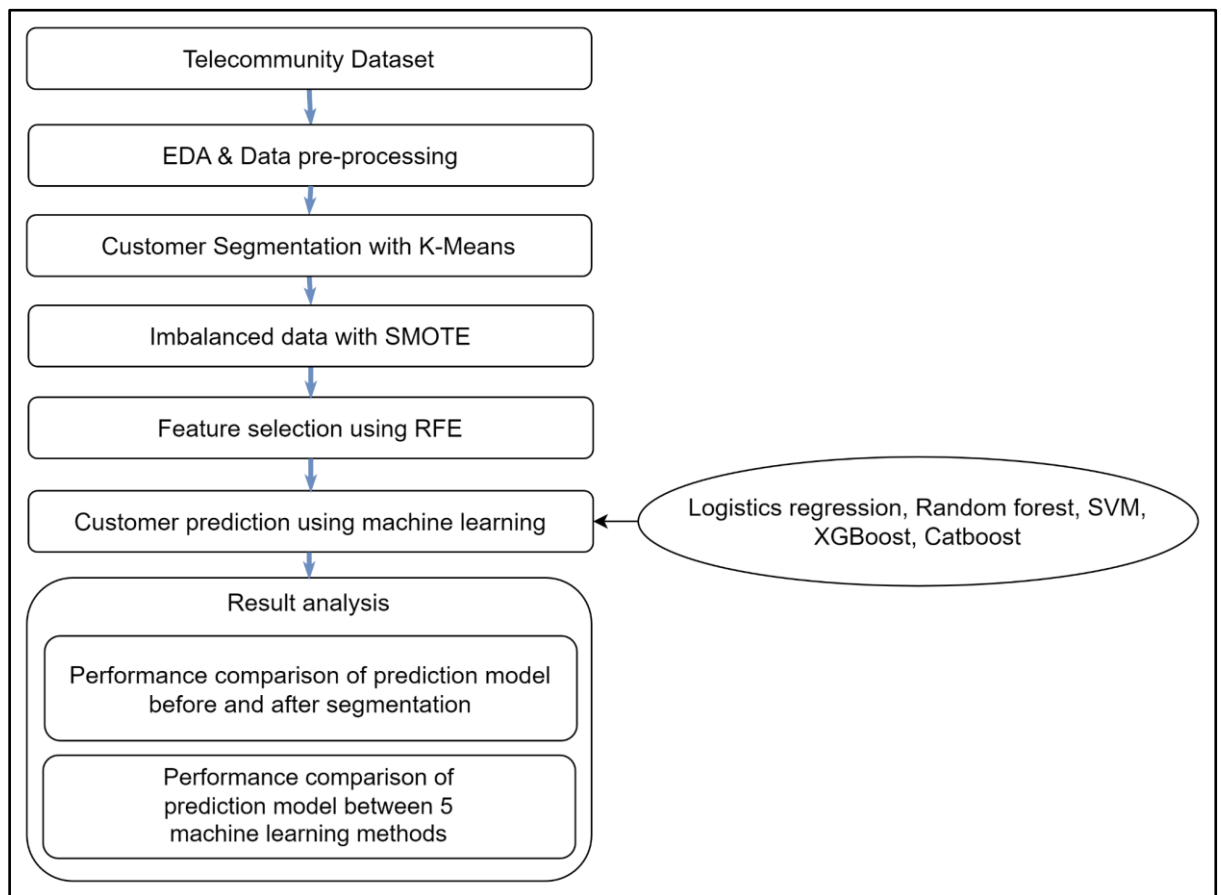


Figure 8. Our proposed methodology

Firstly selecting a dataset suitable for the topic objective. Then data mining and preprocessing is an extremely important step in AI processing in general and machine learning in particular because in the end, the processed data will be obtained and is the input for later stages. Then cluster customers using K-Means machine learning with the number of clusters selected from the Elbow method. After obtaining customer clusters, we conduct unbalanced data processing, feature selection and customer prediction on each obtained cluster. Finally, we compare and evaluate the results obtained and choose the best model suitable for the field of telecommunications based on the obtained dataset.

CHAPTER 5: RESULTS AND DISCUSSION

5.1. Data understanding

5.1.1. Dataset

Our dataset is taken from kaggle about telecommunications company. The raw data contains 7043 rows (each row represents a customer) and 21 columns (each column contains customer's attributes described on the column Metadata). The “Churn” column is our target to predict customer churn.

No	Column name	Description	Data type	Unique value
1	customerID	Customer ID	object	7043 unique values
2	gender	Whether the customer is a male or a female	object	Male, Female
3	SeniorCitizen	Whether the customer is a senior citizen or not	int64	1,0
4	Partner	Whether the customer has a partner or not	object	Yes, No
5	Dependents	Whether the customer has dependents or not	object	Yes, No
6	tenure	Number of months the customer has stayed with the company	int64	
7	PhoneService	Whether the customer has a phone service or not	object	Yes, No
8	MultipleLines	Whether the customer	object	Yes, No, No phone

		has multiple lines or not		service
9	InternetService	Customer's internet service provider	object	DSL, Fiber optic, No
10	OnlineSecurity	Whether the customer has online security or not	object	Yes, No, No internet service
11	OnlineBackup	Whether the customer has online backup or not	object	Yes, No, No internet service
12	DeviceProtection	Whether the customer has device protection or not	object	Yes, No, No internet service
13	TechSupport	Whether the customer has tech support or not	object	Yes, No, No internet service
14	StreamingTV	Whether the customer has streaming TV or not	object	Yes, No, No internet service
15	StreamingMovies	Whether the customer has streaming movies or not	object	Yes, No, No internet service
16	Contract	The contract term of the customer	object	Month-to-month, One year, Two year
17	PaperlessBilling	Whether the customer has paperless billing or not	object	Yes, No
18	PaymentMethod	The customer's payment method	object	Electronic check, Mailed check, Bank transfer (automatic), Credit

				card (automatic)
19	MonthlyCharges	The amount charged to the customer monthly	int64	
20	TotalCharges	The total amount charged to the customer	int64	
21	Churn	Whether the customer churned or not	object	Yes or No

Table 1. The detail information of the telecom dataset

The data set includes information about:

- Services has been signed up for: phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Demographic information: gender, age range, and if they have partners and dependents
- Customer account information: tenure mean how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- The last is our target, customers who left within the last month – the column is called Churn

NaN value in column TotalCharges with “0” tenure value

	tenure	TotalCharges
488	0	NaN
753	0	NaN
936	0	NaN
1082	0	NaN
1340	0	NaN
3331	0	NaN
3826	0	NaN
4380	0	NaN
5218	0	NaN
6670	0	NaN
6754	0	NaN

Figure 9. TotalCharges with “0” tenure value

After performing the data survey, the data does not have duplicate columns. In which, there are 11 cells corresponding to 11 customers in the TotalCharges column with the value NaN, and coincidentally for these 11 customers, the tenure value is 0 i.e. these are customers who have never used the service. Since there are only 11 customers, removing these customers from the dataset does not affect the quality of the data. After removing the last dataset, 7032 customers remained.

5.1.2. Independent variables

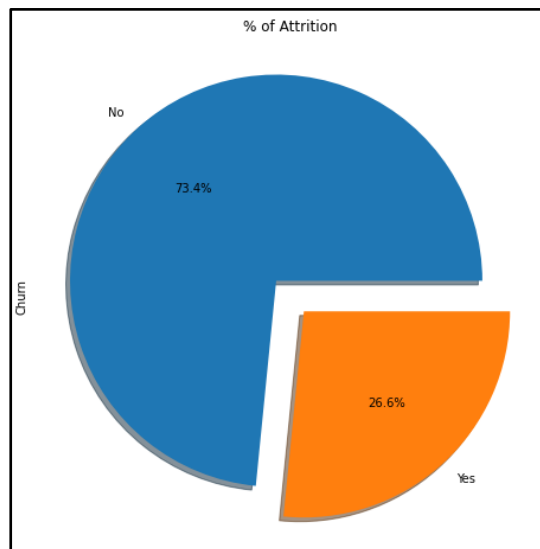


Figure 10. Ratio Churn variable in our dataset

The first thing we notice here is that our target variable is unbalanced, which means we have a much larger proportion of one specific class over another.

Only 26.6% of clients in the dataset were those who left the company.

5.1.3. Dependent variable

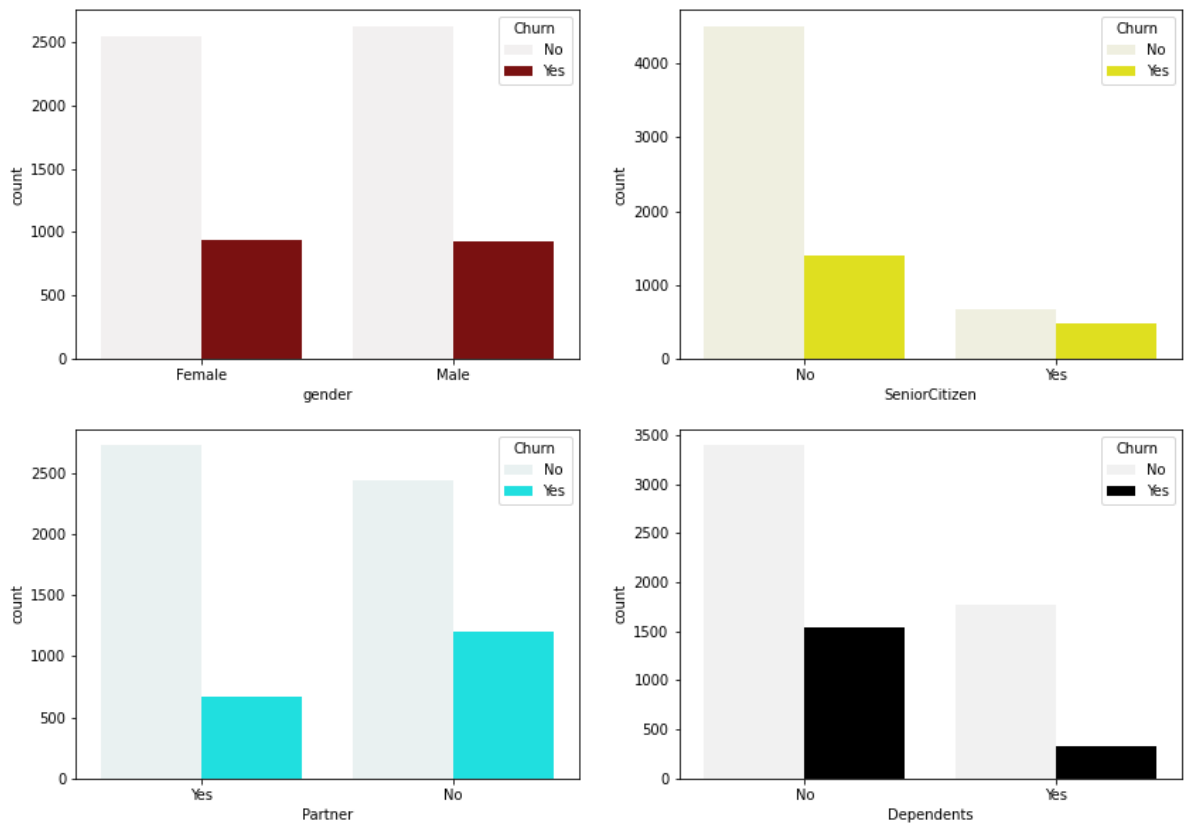


Figure 11. Gender, SeniorCitizen, Partner, Dependents variables visualization

The percentage of elderly citizens is small and most of them leave the business. People without partners or dependents tend to leave the business. There is a similarity in leaving the business between men and women.

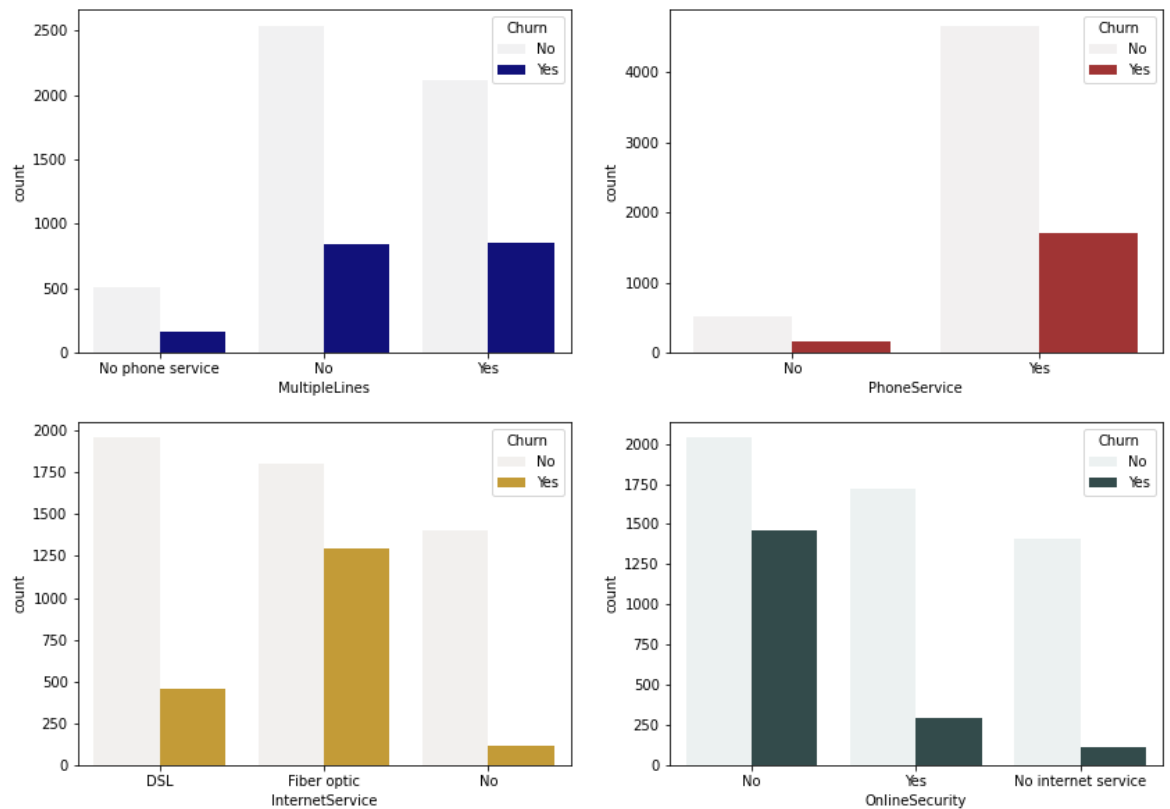


Figure 12. MultipleLines, PhoneService, InternetService, OnlineSecurity variables visualization

- Customers using Fiber optic network services tend to leave high.
- Customers who do not use online security have a high abandonment rate.
- Overall conclusion:A lot of customers choose the Fiber optic service and it's also evident that the customers who use Fiber optic have high churn rate, this might suggest a dissatisfaction with this type of internet service.

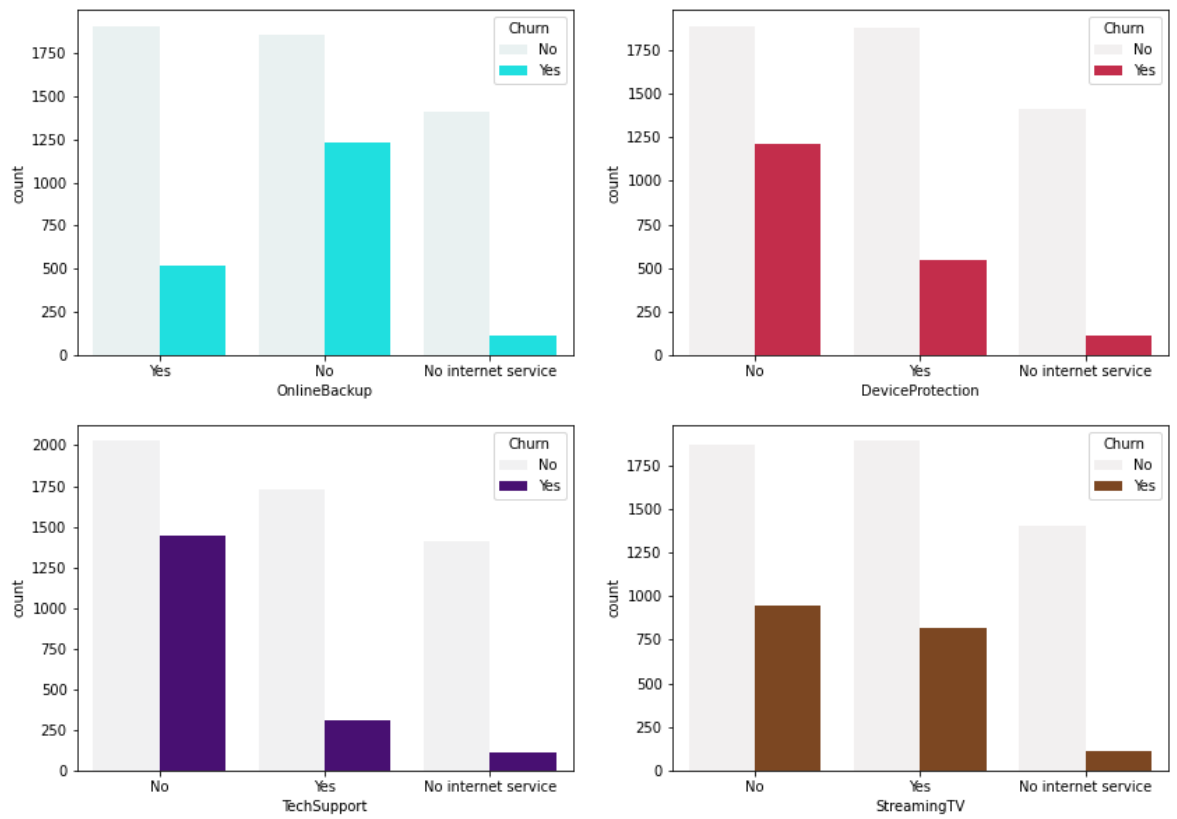


Figure 13. OnlineBackup, DeviceProtection, TechSupport, StreamingTV variables visualization

Customers who don't use tech support tend to switch to another service provider.

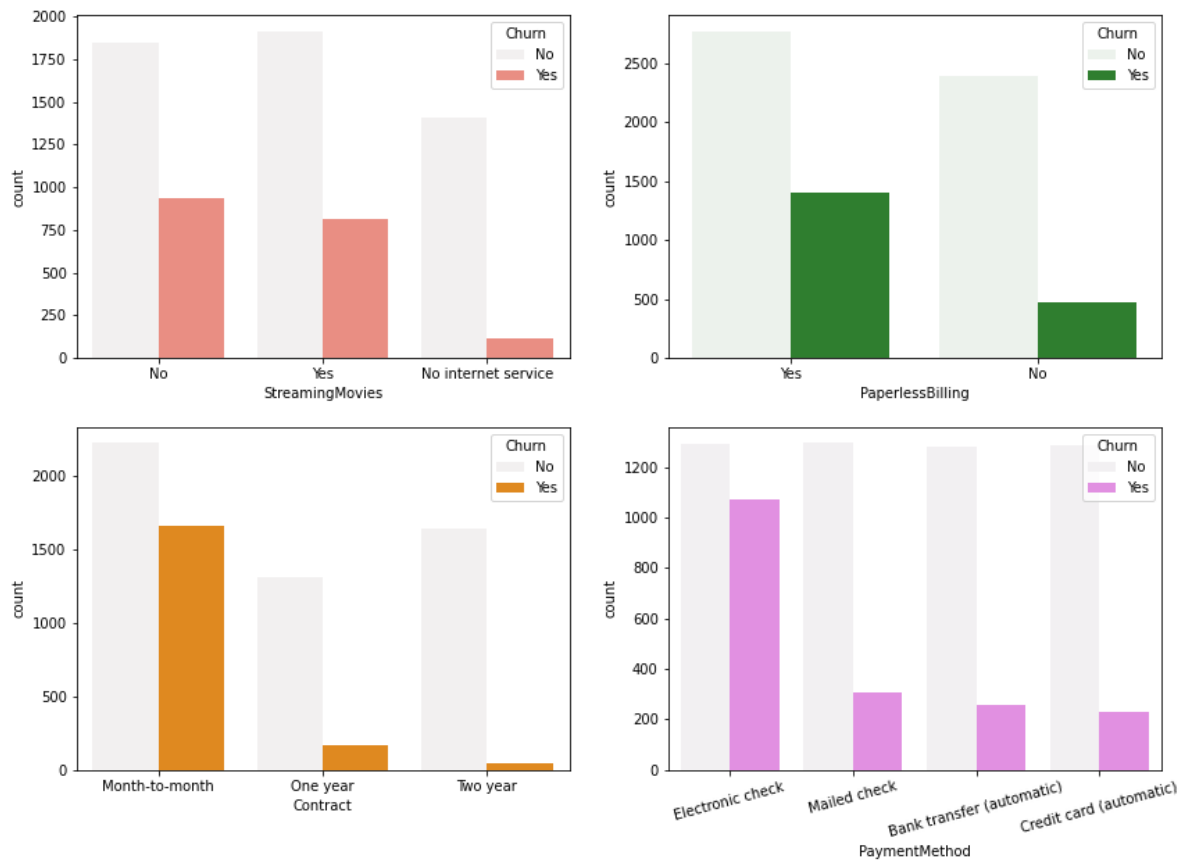


Figure 14. StreamingMovies, PaperlessBilling, Contract, PaymentMethod variables visualization

Most people who use Electronic Check payment methods or use short-term monthly contracts leave the business.

Customers who use paperless payments tend to leave the business:

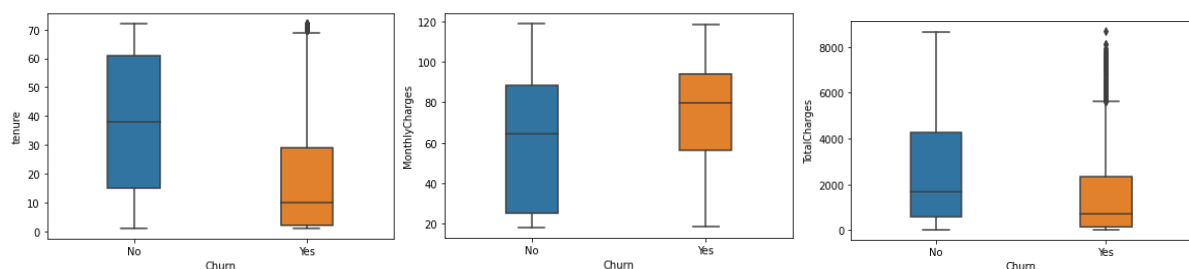


Figure 15. Numerical variables visualization

- It's possible to conclude that 75% of clients who discontinue their subscriptions do so within the first 29 months of service. Half of them leave until the 10th month, not even staying for a year as clients.
- Customers who leave are those who pay more monthly for their services!

- Averagely, the monthly cost for these clients was 21.31% higher than the average cost for clients who stayed in the company! Even among the clients who paid less, the ones who left the company still paid 124% more than those who stayed.
- Even though clients who leave the company pay more monthly, those who stay as clients end up with higher amounts of total charges. It isn't surprising, since they usually stay longer with us, so their total amount of payments will be higher than those who left the company much earlier.
- Customers who pay high monthly fees are more likely to leave the business.
- New customers are more likely to leave the business.

5.2. Data Pre-processing

Feature engineering with categories data

Some algorithms can work with categorical data directly like. For example, a decision tree can be learned directly from categorical data with no data transform required (this depends on the specific implementation).. But many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric. This is mostly a constraint of the efficient implementation of machine learning algorithms rather than hard limitations on the algorithms themselves.

This means that categorical data must be converted to a numerical form. If the categorical variable is an output variable, you may also want to convert predictions by the model back into a categorical form in order to present them or use them in some application.

In this report, all categorical variables are nominal variables. We use two common encoding methods for categorical variables, label encoding and one-hot encoding.

Onehot encoding

One hot encoding is one method of converting data to prepare it for an algorithm and get a better prediction. With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector. All the values are zero, and the index is marked with a 1.

We use the onehot encoding method for variables with 3 or more unique values such as: 'Partner', 'DeviceProtection', 'TechSupport', 'PaperlessBilling', 'PaymentMethod', 'SeniorCitizen', 'OnlineBackup', 'gender', 'StreamingMovies', 'MultipleLines', 'StreamingTV', 'Contract', 'Dependents', 'InternetService', 'PhoneService', 'Churn', 'OnlineSecurity'.

Label encoding

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. In label encoding, we replace the categorical value with a numeric value between 0 and the number of classes minus 1. If the categorical variable value contains 5 distinct classes, we use (0, 1, 2, 3, and 4). We use the label encoding method for variables with 2 unique values such as: 'Dependents', 'Partner', 'PaperlessBilling', 'SeniorCitizen', 'PhoneService', 'gender', 'Churn'.

Correlation between independent variable with dependent variables

Churn	1.000000
Contract_Month-to-month	0.404565
OnlineSecurity_No	0.342235
TechSupport_No	0.336877
InternetService_Fiber optic	0.307463
PaymentMethod_Electronic check	0.301455
OnlineBackup_No	0.267595
DeviceProtection_No	0.252056
MonthlyCharges	0.192858
PaperlessBilling	0.191454
SeniorCitizen	0.150541
StreamingMovies_No	0.130920
StreamingTV_No	0.128435
StreamingTV_Yes	0.063254
StreamingMovies_Yes	0.060860
MultipleLines_Yes	0.040033
PhoneService	0.011691
gender	-0.008545
MultipleLines_No phone service	-0.011691
MultipleLines_No	-0.032654
DeviceProtection_Yes	-0.066193
OnlineBackup_Yes	-0.082307
PaymentMethod_Mailed check	-0.090773
PaymentMethod_Bank transfer (automatic)	-0.118136
InternetService_DSL	-0.124141
PaymentMethod_Credit card (automatic)	-0.134687
Partner	-0.149982
Dependents	-0.163128
TechSupport_Yes	-0.164716
OnlineSecurity_Yes	-0.171270
Contract_One year	-0.178225
TotalCharges	-0.199484
OnlineBackup_No internet service	-0.227578
StreamingMovies_No internet service	-0.227578
StreamingTV_No internet service	-0.227578
TechSupport_No internet service	-0.227578

DeviceProtection_No internet service	-0.227578
OnlineSecurity_No internet service	-0.227578
InternetService_No	-0.227578
Contract_Two year	-0.301552
tenure	-0.354049

Table 2. Correlation between independent variable with dependent variables

In general, the correlation between the independent variables and the dependent variable ranges from -0.35 to 0.4. In which the variable with the largest correlation is the variable: 'Contract_Month-to-month' and 'tenure'.

Data transformation & Standardization

Standardization entails scaling data to fit a standard normal distribution. A standard normal distribution is defined as a distribution with a mean of 0 and a standard deviation of 1. To better understand standardization, it would help to visualize its effects on some data.

We use 3 data transformation methods to transform the transactional dataset into nearly normal distribution data such as: logarit, sqrt and boxcox stat.

A Box Cox transformation is a transformation of non-normal dependent variables into a normal shape. The Box Cox transformation is named after statisticians George Box and Sir David Roxbee Cox who collaborated on a 1964 paper and developed the technique. At the core of the Box Cox transformation is an exponent, lambda (λ), which varies from -5 to 5. All values of λ are considered and the optimal value for your data is selected; The “optimal value” is the one which results in the best approximation of a normal distribution curve. The transformation of Y has the form:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

This test only works for positive data. However, Box and Cox did propose a second formula that can be used for negative y-values:

$$y(\lambda) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \text{if } \lambda_1 \neq 0; \\ \log(y + \lambda_2), & \text{if } \lambda_1 = 0. \end{cases}$$

We can see the result below:

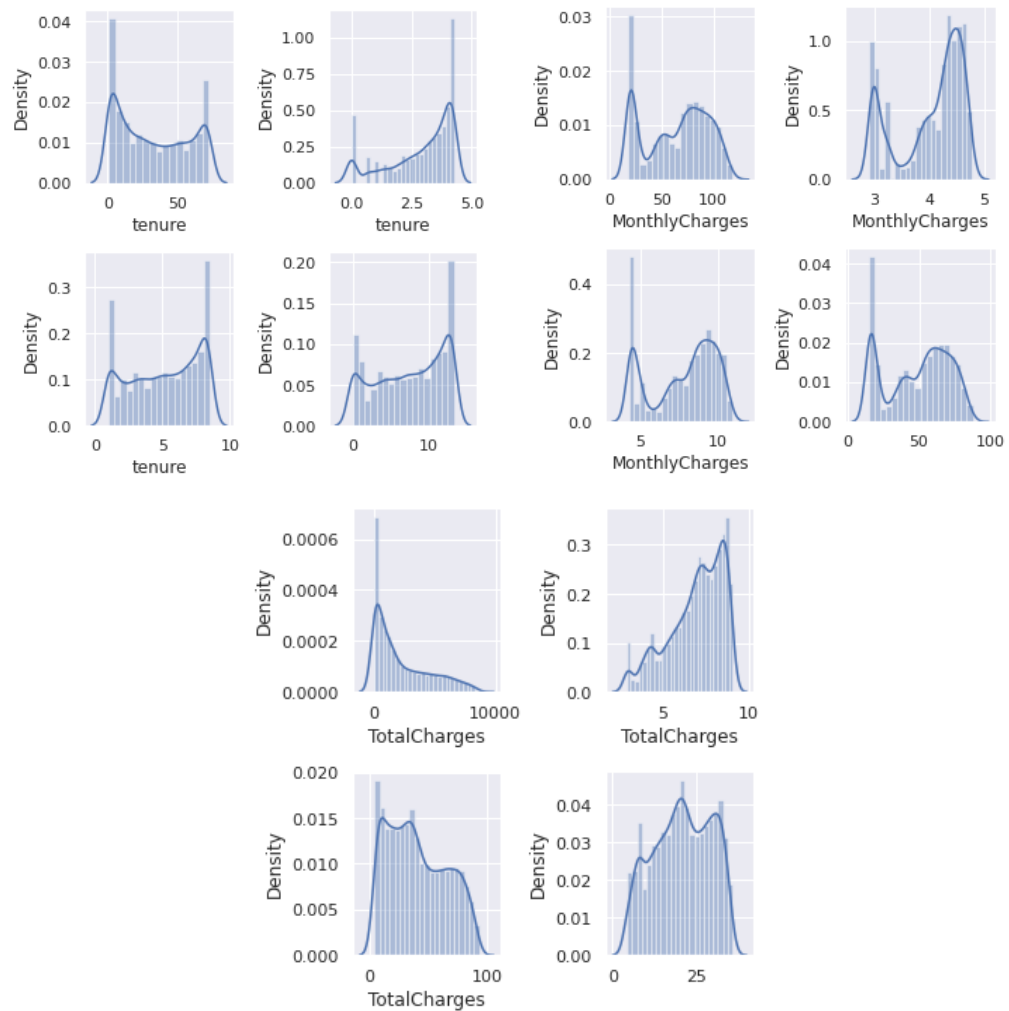


Figure 16. Distribution of numerical variables before and after handling data transformation process

	Origin data	log	sqrt	boxcox
Tenure	0.24	-0.96	-0.23	-0.29
MonthlyCharges	-0.22	-0.73	-0.49	-0.26
TotalCharges	0.96	-0.75	0.31	-0.15

Table 3. Data transformation with numerical variables

As the table shown above we choose sqrt scaler for tenure variable, box cox stata for TotalCharges variable and with MonthlyCharge variable, origin data has the best distribution.

And then we use StandardScaler to scale the 3 variables shown below.

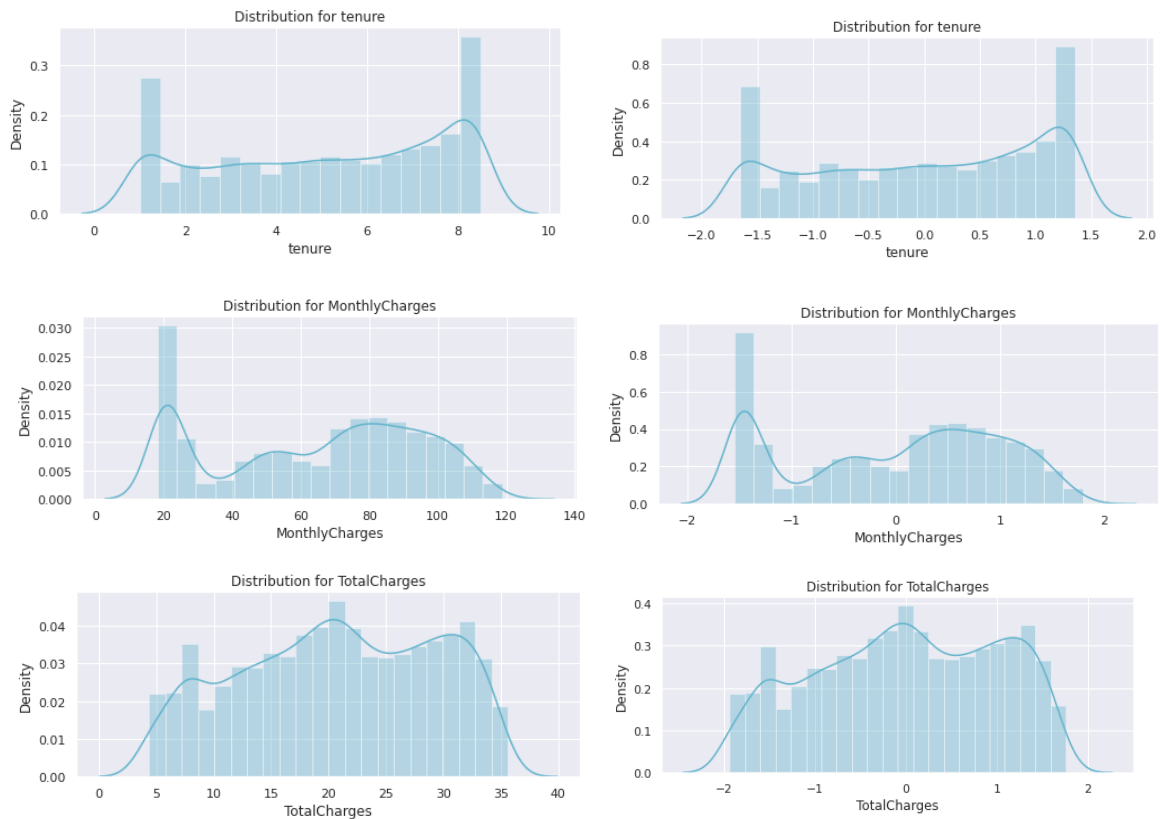


Figure 17. Distribution of the selected numerical variables

5.3. Customer segmentation with K-Means

Elbow method

We finally get the good input dataset after data pre-processing. Firstly, we use elbow method to get the fit number of clusters for KMeans algorithm.

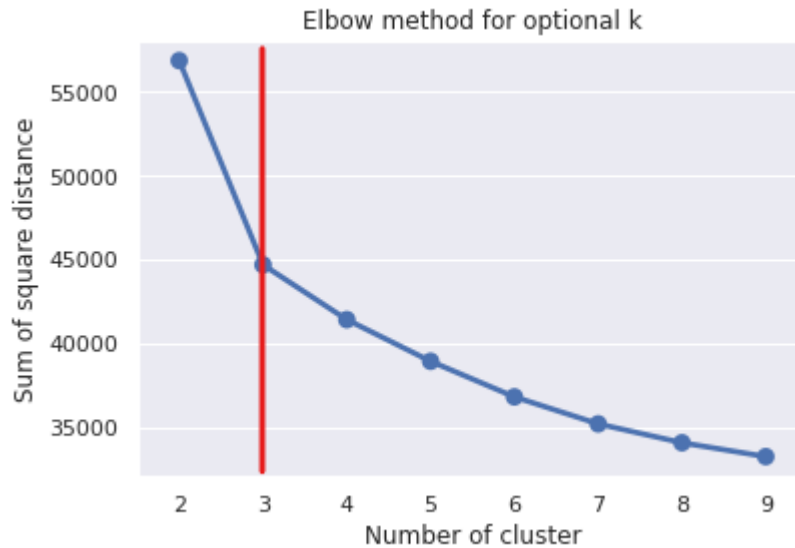


Figure 18. Number of cluster

As the figure shown above, to determine the optimal number of clusters, we have to select the value of k at the “elbow” i.e. the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is 3. The clustering results are shown in Table X.

Cluster	Churn	Number of customers
0	0	2401
	1	497
1	0	1355
	1	1259
2	0	1407
	1	113

Table 4. Number of customer after segmentation

As can be seen in Table 4, there were 2898 customers in Cluster I, among which churn customers were 497, accounting for 17.15% of the total number of Cluster I customers. There were 2614 customers in Cluster II, of which 1259 were churn customers, accounting for 48.16% of the total number of Cluster II customers. There

were 1520 customers in Cluster III customers, of which 113 were churn customers, accounting for 7.43% of the total number of Cluster III customers.

5.4. SMOTE + SPLIT Data

The customer churn data set contained many customer features, and not all variables were conducive to churn prediction performance. Excessive redundancy and irrelevant variables in the dataset may hinder the predictive performance of the model. Therefore, feature selection was carried out next.

Number of customer of each cluster after using SMOTE

Cluster	Churn	Number of customers
0	0	2401
	1	2401
1	0	1355
	1	1355
2	0	1407
	1	1407

Table 5. Number of customer after segmentation and SMOTE

5.5. Features selection - VIF

Recursive Feature Elimination

Recursive Feature Elimination (RFE) is based on the idea to repeatedly construct a model and choose either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features.

VIF

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

After using RFE in combination with VIF processing, we get the last variables that fit each cluster shown in table X.

Cluster	Selected variables using RFE
0	'MultipleLines_No', 'MultipleLines_Yes', 'InternetService_DSL', 'OnlineSecurity_No', 'OnlineSecurity_Yes', 'OnlineBackup_No', 'OnlineBackup_Yes', 'DeviceProtection_No', 'DeviceProtection_Yes', 'TechSupport_No', 'TechSupport_Yes', 'StreamingTV_No', 'StreamingTV_Yes', 'StreamingMovies_No', 'PaymentMethod_Bank transfer (automatic)', 'PaymentMethod_Credit card (automatic)', 'PaymentMethod_Electronic check', 'PaymentMethod_Mailed check', 'Contract_One year', 'Contract_Two year'
1	'PaperlessBilling', 'TotalCharges', 'MultipleLines_No', 'InternetService_DSL', 'OnlineSecurity_Yes', 'TechSupport_Yes', 'StreamingTV_Yes', 'StreamingMovies_No', 'PaymentMethod_Bank transfer (automatic)', 'PaymentMethod_Credit card (automatic)', 'PaymentMethod_Mailed check', 'Contract_Month-to-month', 'Contract_Two year'
2	'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'PaperlessBilling', 'TotalCharges', 'MultipleLines_No', 'MultipleLines_Yes', 'OnlineSecurity_No internet service', 'TechSupport_No internet service', 'StreamingMovies_No internet service', 'PaymentMethod_Bank transfer (automatic)', 'PaymentMethod_Credit card (automatic)', 'PaymentMethod_Electronic check', 'PaymentMethod_Mailed check', 'Contract_Month-to-month', 'Contract_One year', 'Contract_Two year'

Table 6. Selected variables for each cluster

5.6. Machine learning method

5.6.1. Before customer segmentation with KMeans

After the feature selection processing was completed, the data were input into the proposed machine learning models for prediction such as: Logistics regression model, Random Forest, SVM, XGBoost and CatBoost.

These models were applied to the test set data to obtain the confusion matrix. Below show the ROC curves predicted by these models before and after subdivision, and show the confusion matrices before and after subdivision.

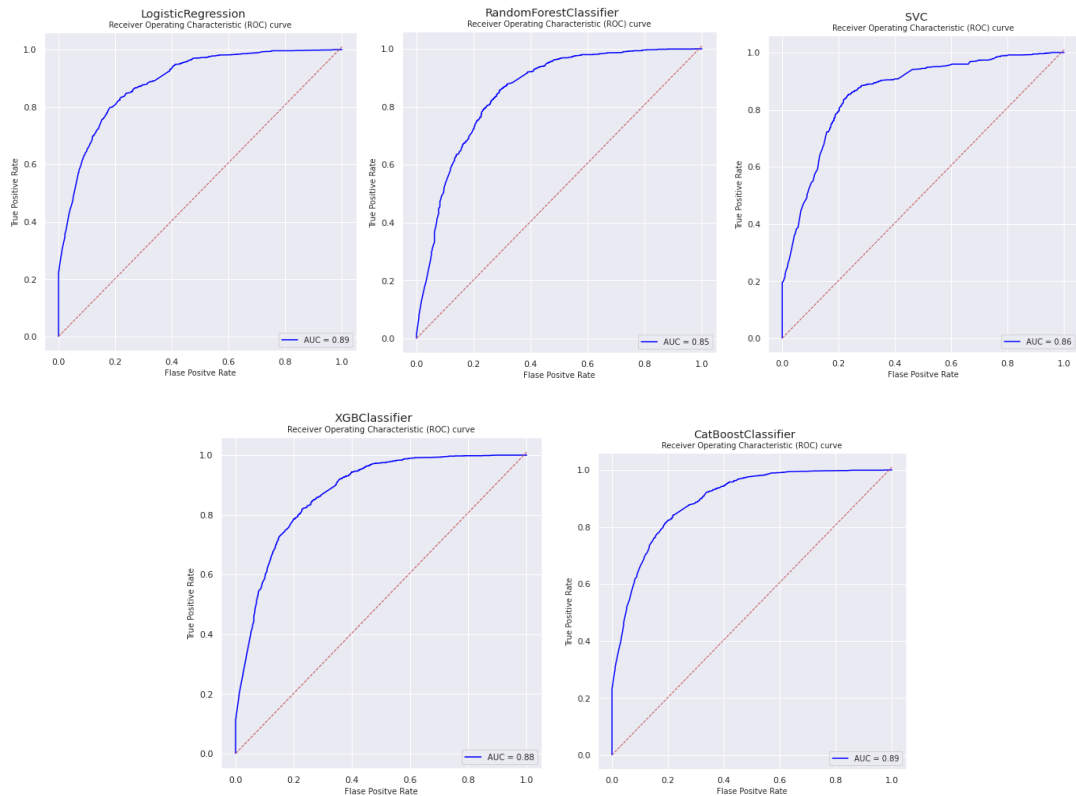


Figure 19. The ROC curve of these 5 models before segmentation

Machine learning	Precision	Recall	F-Score	Accuracy	AUC
Logistics regression model	0.77	0.85	0.81	0.80	0.89

Random Forest	0.77	0.79	0.78	0.78	0.85
SVM	0.79	0.83	0.81	0.81	0.86
XGBoost	0.76	0.84	0.80	0.79	0.88
CatBoost	0.80	0.83	0.81	0.81	0.89

Table 7. Confusion matrix (F1-Score) of these 5 models before customer segmentation

5.6.2. After customer segmentation with KMeans

We conducted comparative experiments on these 5 models. Accuracy, recall, precision and F1-Score values for each category were calculated according to the confusion matrix to evaluate the performance of the three categories in the dataset. It can be seen that the prediction accuracy of the CatBoost model for the three types of customers was higher than other models, and the prediction effect of the CatBoost model was good. However, to confirm the prediction quality, accuracy alone is sometimes misleading. Therefore, when evaluating the prediction performance of the model, not only accuracy but also recall and precision should be observed, and the performance of the prediction model should be comprehensively determined according to the four performance indicators: accuracy, precision, recall and F1-Score. In the experimental results of this paper, on average, the four indicators of the CatBoost model after customer segmentation were 0.81, 0.87, 0.84 and 0.83 is the highest. Therefore, the prediction performance of the CatBoost model was better than others.

Machine learning	Cluster			Predicted	
				Predicted Positive (0)	Predicted Negative (1)
Logistics regression model	0	Actual	Actual positive (0)	405	76
			Actual	75	405

			negative (1)		
	1	Actual	Actual positive (0)	198	73
			Actual negative (1)	82	189
	2	Actual	Actual positive (0)	221	61
			Actual negative (1)	17	264
Random Forest	0	Actual	Actual positive (0)	378	103
			Actual negative (1)	79	401
	1	Actual	Actual positive (0)	186	85
			Actual negative (1)	72	199
	2	Actual	Actual positive (0)	229	53
			Actual negative (1)	15	256
SVM	0	Actual	Actual positive (0)	420	61
			Actual negative (1)	84	396

	1	Actual	Actual positive (0)	191	80
			Actual negative (1)	76	195
	2	Actual	Actual positive (0)	226	56
			Actual negative (1)	13	268
XGBoost	0	Actual	Actual positive (0)	420	61
			Actual negative (1)	84	396
	1	Actual	Actual positive (0)	187	84
			Actual negative (1)	85	186
	2	Actual	Actual positive (0)	229	53
			Actual negative (1)	15	266
CatBoost	0	Actual	Actual positive (0)	421	60
			Actual negative (1)	68	412
	1	Actual	Actual	174	97

			positive (0)		
			Actual negative (1)	53	218
	2	Actual	Actual positive (0)	241	41
			Actual negative (1)	13	268

Table 8. Confusion matrix (actual, predicted) after customer segmentation

Machine learning	Cluster	Precision	Recall	F1-Score	Accuracy	AUC
Logistics regression model	0	0.84	0.84	0.84	0.84	0.93
	1	0.72	0.70	0.71	0.71	0.79
	2	0.81	0.94	0.87	0.86	0.91
	Ave.	0.79	0.83	0.81	0.8	0.88
Random Forest	0	0.80	0.84	0.82	0.81	0.88
	1	0.70	0.73	0.72	0.71	0.78
	2	0.83	0.95	0.89	0.88	0.92
	Ave.	0.78	0.84	0.81	0.8	0.86
SVM	0	0.87	0.82	0.85	0.85	0.93
	1	0.71	0.72	0.71	0.71	0.77
	2	0.83	0.95	0.89	0.88	0.93
	Ave.	0.8	0.83	0.82	0.81	0.88

XGBoost	0	0.83	0.86	0.84	0.84	0.92
	1	0.69	0.69	0.69	0.69	0.78
	2	0.83	0.95	0.89	0.88	0.94
	Ave.	0.78	0.83	0.81	0.8	0.88
CatBoost	0	0.87	0.86	0.87	0.87	0.94
	1	0.69	0.80	0.74	0.72	0.79
	2	0.87	0.95	0.91	0.90	0.96
	Ave.	0.81	0.87	0.84	0.83	0.90

Table 9. .Confusion matrix (F1-Score) after customer segmentation

The generalization ability of a prediction model is an important aspect of whether the prediction performance of the model is satisfactory. Therefore, we used the ROC and AUC to evaluate the generalization ability of the model. ROC can easily detect the influence of arbitrary thresholds on the generalization performance of the learner. We can see the AUC from each cluster of CatBoost models is 0.94, 0.79 and 0.96. They are the highest AUC values. These experimental data prove that the CatBoost model has good generalization ability and good prediction performance. We recommend using the CatBoost model for customer churn predictions in telecommunication companies.

CHAPTER 6: CONCLUSION

6.1. General conclusion

We can see that predicting change is very important in e-business. To be able to stay competitive in the market, businesses should leverage machine learning to predict the likelihood of customer loss and devise appropriate marketing strategies based on the predicted results. The study uses a telecommunity company dataset on business customer churn to test the predictive ability of 5 models: Logistics regression, SVM, Random Forest, XGBoost and CatBoost. To evaluate the goodness of the model, the Accuracy, recall, precision, F1-Score and AUC metrics were calculated. In this study, we've got two worthwhile conclusions for our dataset. Firstly, we used K-Means to cluster the dataset into 3 datasets and each cluster has similar characteristics. Then we used our method to analyze and evaluate each cluster, based on the results showing that the application of the K-Means algorithm to predict is essential because the evaluation metrics have improved significantly. Second conclusion, compare the performance of these 5 prediction models and in the Results section, the Catboost prediction model is proven to be the best predictor model with higher accuracy than the rest. Businesses should regularly use machine learning to predict and evaluate customer abandonment rates to offer customer retention policies to limit costs for businesses.

6.2. Limited and future work:

In the future, we can use deep learning methods for comparison and evaluation. Secondly, we can dig deeper into the input indexes of algorithms, especially CatBoost to turn the model for better efficiency and accuracy prediction. Finally, our dataset is quite small so to have a more objective view we will need a larger dataset about the number of customers and more attributes.

Reference

- [1] BALASUBRAMANIAN, M. (n.d.). CHURN PREDICTION IN MOBILE TELECOM SYSTEM USING DATA MINING TECHNIQUES. International Journal of Scientific and Research Publications. Retrieved December 11, 2022, from <https://www.ijsrp.org/research-paper-0414/ijsrp-p2848.pdf>
- [2] Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management*, 62, 100-107.
- [3] Ahmed, A., & Linen, D.M. (2017). A review and analysis of churn prediction methods for customer retention in telecom industries. 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), 1-7.
- [4] Gaur, A.K., & Dubey, R.K. (2018). Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques. 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), 1-5.
- [5] Senthilnayagi, B. & M, Swetha & D, Nivedha. (2021). CUSTOMER CHURN PREDICTION. IARJSET. 8. 527-531. 10.17148/IARJSET.2021.8692.
- [6] Analysis of Customer Churn Prediction in Telecom Industry Using Logistic Regression | Source Details. (2021, July 4). Scope Database. Retrieved December 11, 2022, from <https://sdbindex.com/Documents/index/00000266/00001-00560>
- [7] Customer Churn Prediction in Telecommunications Industry Based on Data Mining. (2022). Retrieved 11 December 2022, from <https://ieeexplore.ieee.org/document/9509>
- [8] Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems With Applications*, 34(1), 313-327. doi: 10.1016/j.eswa.2006.09.038
- [9] Khine, T., & Myo, W.W. (2020). Customer Churn Analysis in Banking Sector.
- [10] Karvana, K., Yazid, S., Syalim, A., & Mursanto, P. (2019). Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry. 2019 International Workshop On Big Data And Information Security (IWBIS). doi: 10.1109/iwbis.2019.8935884

[11] Mand'ák, J., & Hančlová, J. (2019). Use of Logistic Regression for Understanding and Prediction of Customer Churn in Telecommunications. *Statistika: Statistics And Economy Journal*, 99(2), 129-141. Retrieved from <https://doaj.org/article/1b430587cd884d43aea1f6ef8adfe2d0?>

[12] Miao, X., & Wang, H. (2022). Customer Churn Prediction on Credit Card Services using Random Forest Method. *Proceedings Of The 2022 7Th International Conference On Financial Innovation And Economic Development (ICFIED 2022)*. doi: 10.2991/aebmr.k.220307.104.

[13] Customer Churn Prediction by Classification Models in Machine Learning. (2022). Retrieved 11 December 2022, from <https://ieeexplore.ieee.org/document/9772553>.

[14] Thanh, N., & Vy, N. (2022). Building a proper churn prediction model for Vietnam's mobile banking service. *International Journal Of ADVANCED AND APPLIED SCIENCES*, 9(7), 139-149. doi: 10.21833/ijaas.2022.07.014

[15] machinelearningcoban(2017). Bài 10: Logistic Regression. Retrieved December 16, 2022. from: <https://machinelearningcoban.com/2017/01/27/logisticregression/>.

[16] Peng Li, Siben Li, Tingting Bi, Yang Liu (2014). Telecom customer churn prediction method based on cluster stratified sampling logistic regression.

[17] Roberts, J. H. (2000). Developing new rules for new markets. *Journal of the Academy of Marketing Science*, 28(1), 31–44.

[18] Hemlata Jain, Ajay Khunteta, Sumit Srivastava (2019). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. Retrieved December 16, 2022. from <https://www.sciencedirect.com/science/article/pii/S1877050920306529>.

---END!---