International Journal of Accounting and Management Sciences
Vol.3 No.1, January 2024
Print ISSN: 2834-8923 Online ISSN: 2832-8175
DOI: https://www.doi.org/10.56830/IJAMS012024

Proposed Combined Technique of Statistical Filter and Machine Learning for Exploratory Data Analytics and Features Selecting of Telecommunication Customer Churn

https://www.doi.org/10.56830/IJAMS01202404

Noha Nabawy
Assistant Lecturer in Statistics, Mathematics, and Insurance Department, Faculty of Commerce, Benha University, Egypt
Correspondent author noha.bahi@fcom.bu.edu.eg

Zohdy Nofal
Professor and Head Department of Statistics, Mathematics, and Insurance Department. Faculty of Commerce, Benha University, Egypt.

Eman Mahmoud
Lecturer in Statistics, Mathematics, and Insurance Department. Faculty of Commerce, Benha University, Egypt

Received: 27 November 2023. Accepted: 10 January 2024. Published: 19 January 2024

Abstract

This study can determine a customer's churn based on his historical data and behavior. It indicates that an efficient churn prediction model should employ a significant volume of historical data to identify churners. However, existing models have several limitations that make it difficult to do churn prediction reasonably and accurately. To solve this issue this study proposed new combined technique of statistical filter and machine learning preprocessing is used. Furthermore, statistical methods are utilized to generate models, resulting in poor prediction performance. Also, benchmark datasets are not employed in the literature for model evaluation, resulting in a poor representation of the actual visual representation of data. Without benchmark datasets, it is impossible to compare different models fairly. An intelligent model can be utilized to relieve current issues and deliver more accurate churn prediction.

Keywords: Customer churn, Data Acquisition, Exploratory Data Analytics, Feature selecting technique, Filter statistical technique, Machine learning.

1. Introduction

By increasing number of telecom users, corporations are now offering a variety of services to retain customers. Customer churn can occur for a variety of reasons. The most significant of these are call or package rates that are unsuitable for the customer (Tiwari, Sam, & Shaikh, 2017); (Petkovski, Stojkoska, Trivodaliev, & Kalajdziski, 2016). It occurs when a client moves service providers to acquire better services and advantages. When a customer transfers from one service provider to another, the company's revenue suffers. In the telecom industry an enormous volume of data with missing values is generated.

To avoid this issue, the operator must understand the reason for the customer's choice of changing to another telecom company. Prediction remains the most effective method for analyzing churn behavior. Because of the enormous amount and challenges nature of the data set, predicting customer attrition in the telecommunications business has traditionally been a difficult challenge. Attrition prediction is used to discover which consumers are most likely to churn. Churn prediction and analysis can assist a company in improving the sustainability of their customer satisfaction strategy.

2. Literature review

(Ahmad, Jafar, & Aljoumaa, 2019) Data pruning and cleansing are done during the pre-processing stage. Based on previous behaviors and historical customer data, it is possible to identify users or customers who are subject to churning. The Synthetic Minority Over-Sampling Technique (SMOTE) is used to optimize the model. However, local optimal solution problems in feature selection strategies of this kind require large-scale data feature extraction with high accuracy in feature classification.

(Dwiyanti & Ardiyanti, 2017) In general, the dataset distribution for churn prediction is skewed, counting many cases in a single class relative to other classes. The class with more instances is called the majority class, and the class with fewer samples is called the minority class. The distribution of imbalanced instances within the dataset is shown by the imbalance ratio between classes. The number of non-churners in the churn-prediction model was more than the number of churners.

(Gui, 2017) Pre-processing is used to balance the imbalanced dataset, and RUS (Random Under sampling) and SMOTE sampling techniques are used to extract the features.

(Nguyen & Duong, 2021) There are two types of methodologies commonly used in dealing with unbalanced data inside the churn prediction model: resampling approaches and cost-sensitive learning techniques. During the model-training phase, cost-sensitive learning approaches alter the relative mistake costs. Resampling data performs effectively in managing unbalanced data and sorting out with balanced information before the model's training stages.

(Nurhidayat & Anggraini, 2023) Customer analysis using exploratory data analysis (EDA) for visualizing data and the use of machine learning for the classification of customer churn are often used by past analysts. The Synthetic Minority Over-Sampling Technique (SMOTE) method is a popular method applied to deal with class imbalances in datasets.

3. Methodology

- The proposed model's first phase is to pre-process the uploaded dataset and merge Benchmarked Data Sets.
- This process assists in cleaning the input data by removing any redundant items, removing, or replacing String elements, removing non-processing data entries, and so on.
- Feature Selection assists in improved data comprehension. The feature selection-based method reduces computing time and storage requirements.
- The feature selection combination of statistical techniques (Filters) and machine learning which is wrapper technique chooses relevant features.
- the real data set has more than two variables, it is still possible to glean useful information from analyzing every potential bivariate heat map matrix between all pairs of variables of the two variables given by the row and column. The squares maintaining the variable names additionally contain the variable's minimum and maximum values. Although, we lose some information about the distribution, it is important to build bivariate statistical indices that further summaries the frequency distribution, increasing data interpretation. These indexes allow us to summarize the distribution of each data variable in the bivariate situation, and more broadly in the multivariate case, as well as learn about the relationship between the variables (corresponding to the columns of the data matrix).
- Concordance is the tendency to observe high (low) values of one variable with high (low) values of another. Discordance is the tendency to observe low (high) values of one variable with high (low) values of another. The most frequent summary measure for measuring concordance is covariance, which is defined as

$$Cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu(x))(y_i - \mu(y)) \qquad (1)$$

where, μ(X) is the mean of variable X and μ(Y) is the mean of variable Y. The covariance takes positive values if the variables are concordant and negative values if they are discordant.

- Independence between two variables, x and y , holds when

$$n_{ij} = \frac{n_i + n_j}{n} \qquad (2)$$

Where, ∀i = 1,2,...,I; ∀j = 1,2,...,J

for all joint frequencies of the contingency data frequencies actually observed n_{ij} and those expected in the hypothesis of independence between the two variables $\frac{n_i+n_j}{n}$. Karl Pearson is the most widely used measure for verifying the hypothesis of independence between x and y . It is defined as following

$$z^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*} \qquad (3)$$

Note that z^2 = 0 if the variables x and yare independent. This reveals a serious inconvenience the value of z^2 is an increasing function of the sample size n. To overcome such inconvenience, some alternative measures have been pro- posed, all functions of the previous statistic.

The Cramer index is equal to

$$w^2 = \frac{z^2}{n\,\min\big((I-1),(J-1)\big)} \qquad (4)$$

If 0 ≤w^2 ≤1for any I×J contingency data, and w^2=0 if and only if x and y are independent.

w^2 = 1 for maximum dependency between the two variables. Then three cases can be distinguished, referring without loss of generality to data as following:

1.

Furthermore, w^2 has an asymptotic probabilistic (theoretical) distribution, so it can also be used to assess an inferential threshold to evaluate inductively whether the examined variables are significantly dependent.

- High Correlation Filter which plotted the correlation between each variable and another and dropped the ones that had a very high correlation with each other, indicating that we had redundant features. For categorical data Label encoding has been applied for binary categories data and one hot encoding has been applied for multi categories data.

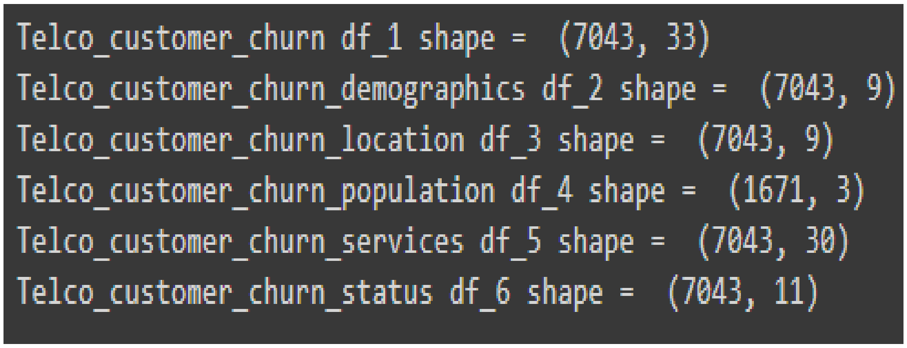
- Wrapper approaches the topic as a search problem, where numerous combinations are tested and assessed to find the best possibilities and eliminate the remainder. This is like the recursive feature removal algorithm.
- This proposed technique could reduce attribute size while also reducing misclassification errors.
- The study also verified the importance of using the feature-selection procedure to extract important characteristics for training the prediction model.

4. Results

The raw data were turned into a final dataset during the data preparation step so that we could feed it into the modelling algorithms and create models. At this phase, many tasks were completed, including data cleansing, handling missing values, selecting features, and data transformation.

This section discusses the churn analysis results, and the study work has created a script in the PYTHON programming language.

i. Data acquisition in six Data frames from Benchmarked IBM datasets.



```
Telco_customer_churn df_1 shape =  (7043, 33)
Telco_customer_churn_demographics df_2 shape =  (7043, 9)
Telco_customer_churn_location df_3 shape =  (7043, 9)
Telco_customer_churn_population df_4 shape =  (1671, 3)
Telco_customer_churn_services df_5 shape =  (7043, 30)
Telco_customer_churn_status df_6 shape =  (7043, 11)
```

ii. Data Merging: Merge all data frames df1, df2, df3, df5, df6 using Customer ID and merge all data frames with df4 population using Zip Code.

iii. Data Cleaning: Drop duplicated columns, drop noisy data (null elements) and drop missing data and Impute others.

iv. Feature selecting firstly through correlation method as follows.

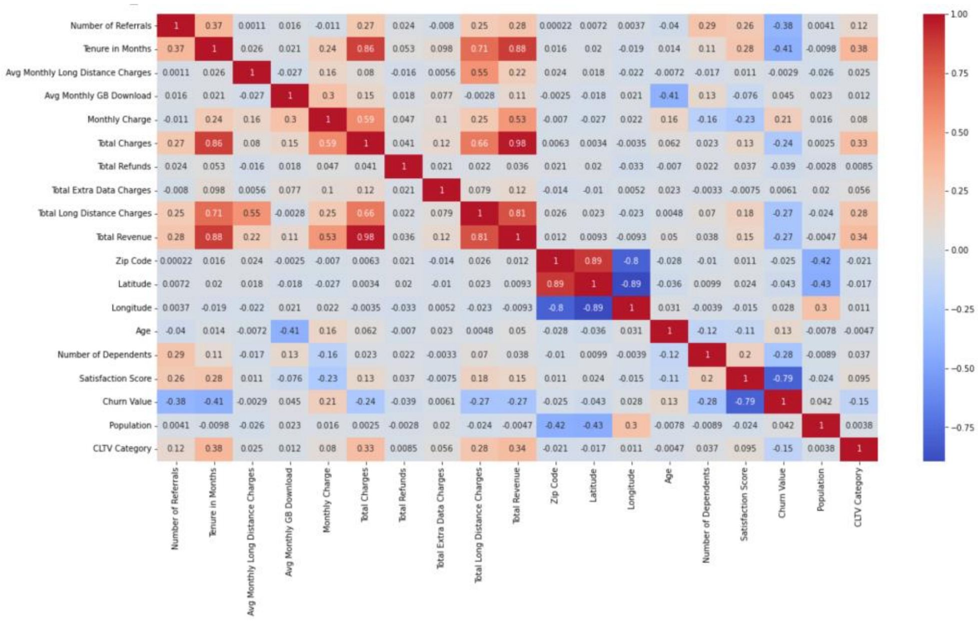


Figure1. Heat map correlation of features

v. Proposed selection feature by splitting Data into numerical and categorized data.

A. Numerical columns selected by applying (Pearson and ANOVA)

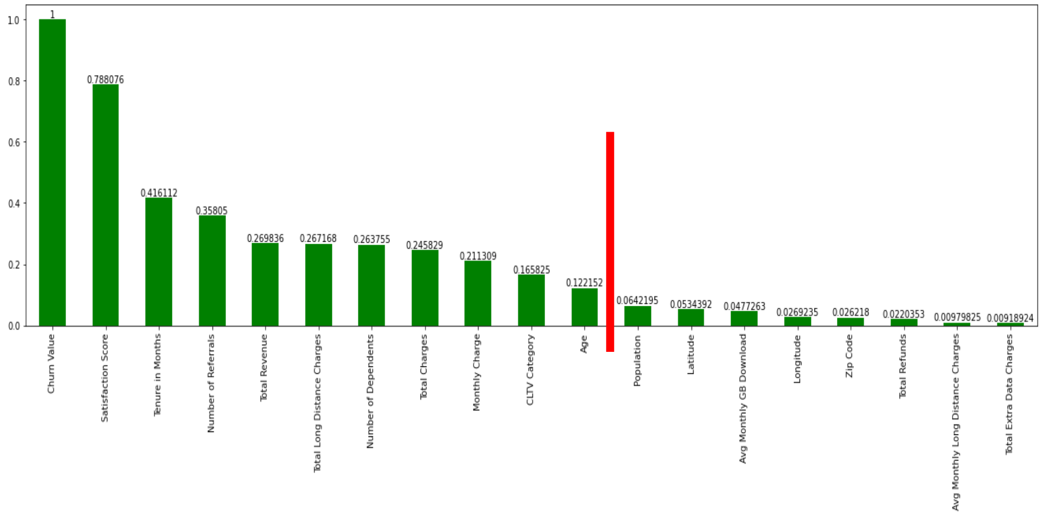


Figure2. persons with threshold 0.1

Pearson selected column (10 columns) and ANOVA feature selection (11 columns). Add ANOVA and persons selected column to Data frame then select common between two techniques.

Selected_numerical_columns = ['Satisfaction Score', 'Tenure in Months', 'Number of Dependents', 'Number of Referrals', 'Age', 'Monthly Charge', 'Total Revenue', 'CLTV Category', 'Total Charges', 'Total Long-Distance Charges'].

B. Categorical columns selected by applying three feature selection techniques (Chi-Square test, ANOVA, and Mutual_info_classif).

Split categorical data binary and multi category. Select common columns from 3 techniques. Remain 23 columns after hyper tuning use ANOVA and chi-Square.Encoding is the process of transforming category data into numerical data. most common type of data in the sample was categorical data, and the values for these variables were typically kept as text. However, because machine learning algorithms are built on mathematical equations, they can only be used with numerical data. As a result, leaving the categorical variables in their current state was impossible, and they had to be transformed into a numerical format. Therefore, apply label encoding on binary category columns and apply one hot encoding on Multi category columns then category columns became 35 columns.

vi. Merge selected numerical and categorical Columns by wrapper.

Became 33 column and apply wrapper technique (backward and forward). Set K = 28 to wrapper models then select common features between backward and forward became 23 columns as follows.

1. Customer ID: A unique ID that identifies each customer.
2. Gender: The customer’s gender; Male, Female.
3. Age: The customer’s current age, in years, at the time the fiscal quarter ended.
4. Senior Citizen: Indicates if the customer is 65 or older: Yes, No.
5. Married (Partner): Indicates if the customer is married; Yes, No.
6. Dependents: Indicates if the customer lives with any dependents; Yes, No. Dependents could be children, parents, grandparents, etc.
7. Number of Dependents: Indicates the number of dependents that live with customer.
8. Phone Service: Indicates if the customer subscribes to home phone service with company; Yes, No.
9. Multiple Lines: Indicates if the customer subscribes to multiple telephone lines with the company; Yes, No.

10. Internet Service: Indicates if the customer subscribes to internet with the company; No, DSL, Fiber Optic, Cable.
11. Online Security: Indicates if the customer subscribes to an additional online security service provided by the company; Yes, No.
12. Online Backup: Indicates if the customer subscribes to an additional online backup service provided by the company; Yes, No.
13. Device Protection Plan: Indicates if the customer subscribes to an additional device protection plan for their internet equipment provided by the company; Yes, No.
14. Premium Tech Support: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times; Yes, No.
15. Streaming TV: Indicates if the customer uses their internet service to stream television programming from a third-party provider; Yes, No. The company doesn't charge an additional fee for this service.
16. Streaming Movies: Indicates if the customer uses their internet service to stream movies programming from a third-party provider; Yes, No. The company doesn't charge an additional fee for this service.
17. Contract: Indicates the customer's current contract type; Month-to-Month, One Year, Two years.
18. Paperless Billing: Indicates if the customer has chosen paperless billing; Yes, No.
19. Payment Method: Indicates how the customer pays their bill; bank withdrawal, credit card, mailed check.
20. Monthly Charge: Indicates the customer's current total monthly charge for all their services from the company.
21. Total Charges: Indicates the customer's total charges, calculated to the end of the quarter specified above.
22. Tenure: Indicates the total amount of months that the customer has been with the company.
23. Churn: Yes= the customer left the company this quarter. No=the customer remained with the company. Directly related to Churn value.

vii. By deeply analytics for the final features from Data found the following Churn Reasons

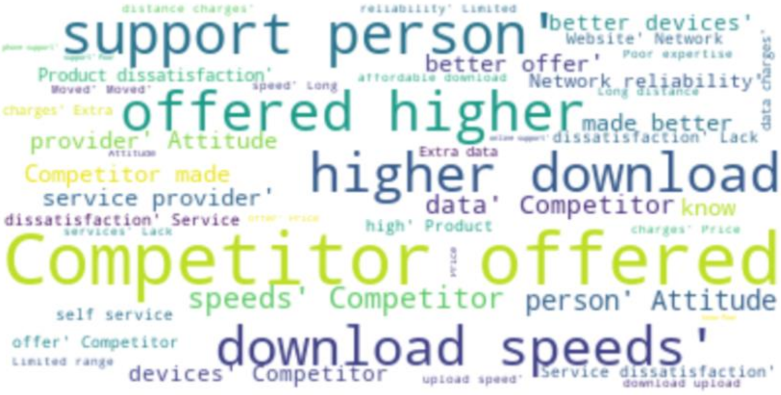


Figure 3. Word cloud of churn reasons

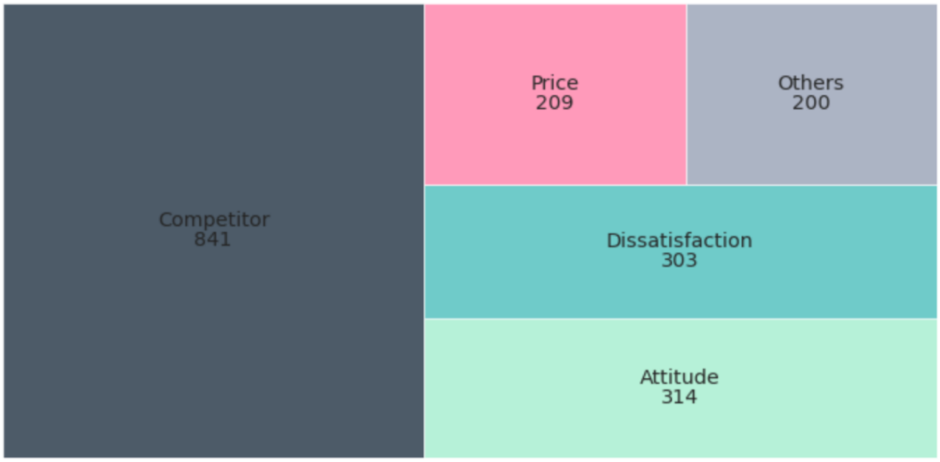


Figure 4. Churn reasons categories

International Journal of Accounting and Management Sciences
Vol.3 No.1, January 2024
Print ISSN: 2834-8923 Online ISSN: 2832-8175
DOI: https://www.doi.org/10.56830/IJAMS012024

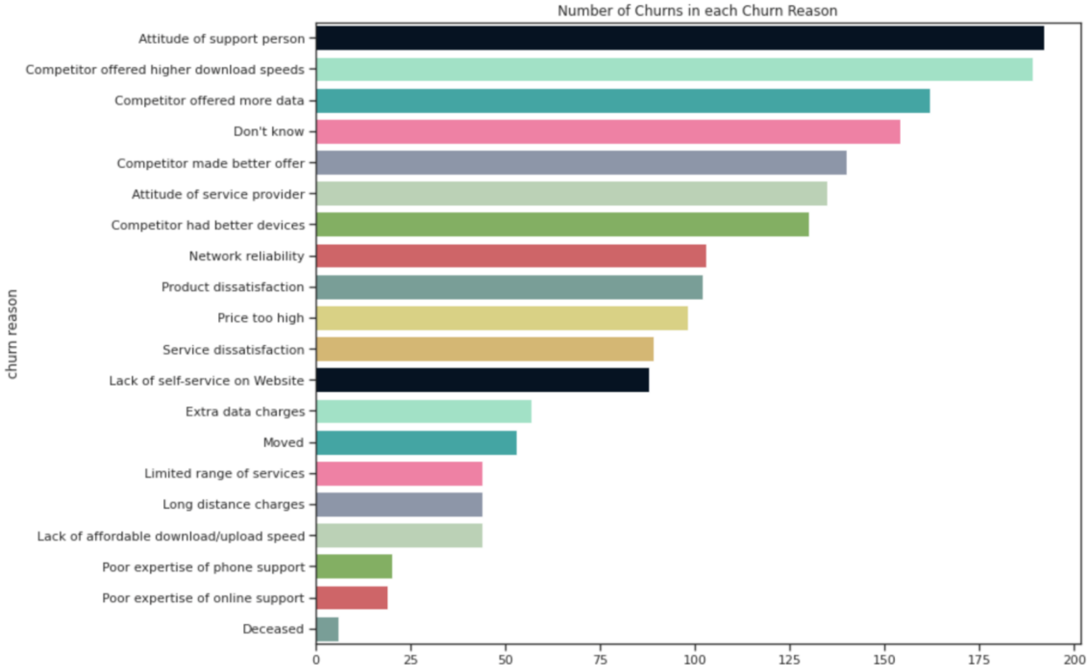


Figure 5. Bar chart for counting nubmer of churns in each churn reason

It’s clear from three pervious chart that competitor offers, and attitude of support person is the most important columns or customer churn reason. Also, internet service and technical support and the contract type.

5. Conclusion

The feature-selection approach could be used to identify and get the necessary characteristics. The performance results demonstrated the significance of selecting a feature-selection procedure to develop a higher-quality churn prediction model. The study recommended employing feature selection to obtain relevant features, which results in improved customer-churn prediction framework performance. Using SBFS-Sequential Backward-Floating Selection, feature selection technique (with feature number), and SBS-Sequential Backward Selection, the framework provided optimized higher performance in churn prediction.

No research addressed the method of cognitive churn analysis. This architecture allows for the monitoring of a broader range of consumer behavioral attributes (unmeasurable parameter values), including customer feedback. Similarly, the problem of improving prediction accuracy and performance capabilities is dependent on churn analysis. The customer behavioral attributes must be valued based on the weight criteria of those churn effects. As a result, future studies should include a score evaluation of customer behavior aspects. The statistics ought to be utilized by the research to determine which category of attributes has the greatest impact on churn rate.

The study includes a thorough churn prediction techniques which are useful in a variety of businesses. The analytics clearly demonstrate the high significance of the churn-predictive method, particularly in the communication industry, for maintaining customer retention and producing high essentials for all service sectors.

References:

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data, 6(1)*, 1-24.

Dwiyanti, E. A., & Ardiyanti, A. (2017). Handling imbalanced data in churn prediction using rusboost and feature selection (case study: Pt. telekomunikasi indonesia regional 7). In Recent Advances on Soft Computing and Data Mining. *August 18-20, 2016, Proceedings Second* (pp. pp. 376-385). Bandung, Indonesia,: Springer International Publishin.

Gui, C. (2017). Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. Artif. *Intell. Res., 6(2)*, 93.

Nguyen, N. N., & Duong, A. T. (2021). Comparison of two main approaches for handling imbalanced data in churn prediction problem. *Journal of advances in information technology, 12*, (1).

Nurhidayat, M. M., & Anggraini, D. (2023). Analysis and Classification of Customer Churn Using Machine Learning Models. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 7(6)*, 1253-1259.

Petkovski, A. J., Stojkoska, B. L., Trivodaliev, K. V., & Kalajdziski, S. A. (2016). Analysis of churn prediction: a case study on telecommunication services in Macedonia. *In 2016 24th Telecommunications Forum (TELFOR)*, (pp. 1-4). IEEE.

Tiwari, A., Sam, R., & Shaikh, S. (2017). Analysis and prediction of churn customers for telecommunication industry. *In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* , (pp. (pp. 218-222). IEEE).