Advances in Consumer Research
https://acr-journal.com/
Volume-3 | Issue-1 | Jan 2026

Original Researcher Article

"Customer churn prediction Using Machine Learning"

Dr. Prashant chordiya[1], Dr. Atvir Singh[2]
[1]Assistant Professor Dr.D.Y. Patil Centre for Mangement and Research , Pune
[2]Professor, Chaudhary Charan Singh University, Meerut, India

ABSTRACT

Rapid technology growth has affected corporate practices. With more items and services to select from, client churning has become a big challenge and threat to all firms. We offer a machine learning-based churn prediction model for a B2B subscription-based service provider. Our research aims to improve churn prediction. We employed machine learning to iteratively create and evaluate the resulting model using accuracy, precision, recall, and F1- score. The data comes from a financial administration subscription service. Since the given dataset is mostly non-churners, we analyzed SMOTE, SMOTEENN, and Random under Sampler to balance it. Our study shows that machine learning can anticipate client attrition. Ensemble learners perform better than single base learners, and a balanced training dataset should increase classifier performance..

.

1. INTRODUCTION:

Digitalization and globalization have led to new ways of doing business, and organizations around the world have had to adapt. Subscription based services are one result of the tremendous digitalization that has taken the globe by storm. With this comes both possibilities and challenges that require new solutions. Digitalization has revolutionized how business is performed and increased the supply of subscription-based services. Companies may find it harder to keep customers as a result. Digitalization can save labor expenses, boost efficiency, and provide a better picture of company operations. This is vital for keeping competitive and gaining an edge over other companies.

Since information technology is growing, the amount of data and information has expanded in recent years. This rapid rise has permitted the storage and processing of large volumes of data and increased the need to automatically identify and create knowledge. By extracting valuable information from stored data, organizations can grow. With this rise, data mining and machine learning are used more often because they can handle and analyze large amounts of data. The digitalization movement has also improved customer relationship management (CRM) data processing processes.

Knowledge management and customer relationship management have gained increased attention in the subscription-based business model. The concepts focus on allocating resources to customer-centric activities to increase competitive advantage. Customer knowledge is information gleaned from customers. Customer relationship management systems collect, store, and analyze data to give organizations an overview of their consumers. Such systems have grown over the years, and by employing technology and data analysis tools, businesses can detect patterns in client behavior that would be hard to discover manually. These patterns could represent a customer's buying habits or churning. In a subscription-based company model, success depends on minimizing client churn.

A consumer churns when they leave their service provider. The word has gained popularity with the rise of online services. Firms around the world see client churn as a huge loss since they've already invested in attracting them.

This is one reason why firm retention is important. Customers churn for numerous reasons; it's hard to define a general cause. The availability of information has given consumers negotiating power, and they can simply discover a provider who offers the same goods at a better price. To address this, corporations invest in 2 customer churn prediction, which implies they try to forecast which customers will quit so they can prevent it.

Depending on the reason a client may churn, preventive interventions could include a lower price or extra service. Previously said, studying customer behavior helps predict attrition, which is crucial for several reasons. For companies that rely on subscription-based income, it can affect whether they can sustain a stable income or need to adapt their offerings to keep clients. Compared to retaining consumers, acquiring new ones is more expensive, thus corporations save money by keeping existing ones.

Problem-solving:

Digitalization is creating subscription-based business models that allow companies a new way to do business. Digitalization makes data easier to acquire, store, and handle. In today's service economy, competition has increased, making it difficult to keep customers. Subscription-based companies must focus on CRM, especially churn management, due to data and replacements.

Blank & Hermansson argue a low churn rate is important to a subscription-based firm's profitability. Current customers buy more from a service provider than new ones. Customer acquisition costs demonstrate this. Verbeke et al. say recruiting new customers costs five to six times more than retaining them, underscoring the necessity of lowering turnover. Retention strategies have a higher net ROI than acquisition, which may boost repeat-customer revenue.

Predicting and avoiding turnover can increase a company's reputation and income, say Amin et al. A reduced churn rate enhances a company's profitability because of consumer revenues. CRM can lead to better, more valuable customer relationships and increase business loyalty, which can boost income.

Companies offer incentives to retain customers. When building retention initiatives, firms shouldn't focus on all 3 clients because such techniques aren't free and not all consumers churn. Target clients who are likely to leave with retention incentives; identifying churn is a prediction problem, and machine learning has been successful in many fields. B2C uses machine learning to accurately predict customer attrition. Cunningham said data is extracted using statistical procedures. Such tactics require expertise. Machine learning streamlines data extraction and improves statistical analysis. Most machine learning algorithms include statistical approaches, making separation hard.

Purpose

Digitalization and easy information access cause continual client turnover, especially for subscription-based service providers. Business-to-business features, including a higher transactional value per client, hurt revenue more when customers churn. Predicting turnover and helping firms manage it is key. Our research seeks to forecast B2B turnover. In this work, we examine B2B churn prediction models using machine learning. The purpose is to study how machine learning can predict churn for subscription-based companies. I'll develop a binary churn prediction model, compare its algorithms, and balance its training dataset.

Delimitations

Financial Services Company commissioned this investigation. The chosen company only deals with other businesse, hence this study will only involve B2B enterprises. Fewer studies have been done on B2B churn. Time and complexity required comparing and evaluating three supervised learning techniques. In a perfect world, algorithms and methods can predict customer attrition. Delimitation is using just particular algorithms. Limited raw data prevents calculating attrition rate. This study won't look at how predicting client attrition can cut costs and boost revenue.

Comparable

Many methods anticipated telecom churn. ML and DM are widely utilized. Most related research focuses on data mining and churn prediction. Gavril et al. suggested a data mining method to anticipate prepaid customer churn utilizing 3333 call information, 21 characteristics, and a Yes/No churn parameter. Message counts and voicemail are features. PCA reduced dimensionality. Using NN, Bayes and SVM predict churn. The author measured AUC. Bayes, Neural, and SVM AUCs were 99.10%, 99.55%, and 99.70%. This study's little dataset was complete. He et al. proposed a Neural Network-based model to predict customer turnover in a 5.23 million-customer Chinese telecom firm. Predictions were 91.1% accurate.

2. LITERATURE REVIEW

We used BTH summon and Google scholar to research. We used Diva portal, a searchable database. We mainly used books and peer-reviewed publications from scientific journals and conferences to filter results and ensure high reliability. As our study investigates customer churn prediction, we searched for customer churn prediction, customer churning, customer relationship management, churn management in subscription-based services, and churn prediction in B2B.

Machine learning-based churn prediction system Math Subject Classification, 14 Feb 2021, Praveen Lalwani, Manas Kumar, Mishra,Jasroop Singh,Chadha,Pratyush Seth.

In the prediction process, logistic regression, naive bayes, support vector machine, random forest, decision trees, etc. are implemented on train set as well as boosting and ensemble techniques to examine the effect on model accuracy. K-fold cross validation is performed over train set to tune hyper parameters and prevent model overstating. The test set outcomes were analyzed using confusion matrix and AUC curve. Adaboost and GBoost Classifier have the highest accuracy, 81.71% and 80.8%. Adaboost and GBoost Classifiers achieve the greatest AUC of 84%.

Analysis of machine learning techniques for churn prediction and factor identification in telecom IRFAN ULLAH1, BASIT RAZA1, Ahmad Kamran Malik, Muhammad Imra, Saif Ul Islam, and Sung Won Kim. IEEE, 4/30/19

This article discovered churn characteristics that help determine its causes. By recognizing important churn variables from customer data, CRM can boost productivity, offer relevant promotions to likely churn customers based on similar behavior patterns, and optimize firm marketing initiatives. Accuracy, precision, recall, f-measure, and ROC area are used to evaluate the churn prediction model. Our churn prediction technique improved RF churn classification and k-means clustering customer profile. Using the attribute-selected classifier technique, it also offers factors behind churning clients.

A B2B SaaS provider used machine learning to predict customer churn. Marie Sergue, Second-Cycle Physics Degree Project, 30 Credits Stockholm2020,

This thesis analyzes real-life data from a SaaS startup providing an innovative cloud-based business phone solution, Air call. This use case's dataset collects monthly customer data and has an uneven target distribution: most customers do not churn. Several ways are tried to reduce the imbalance's influence while staying true to the real world and temporal framework. Oversampling, under sampling, and time series cross-validation are used. To predict and explain churn, logistic regression and random forest models are used. Non-linear models fared better than logistic regression for our use case. Oversampling with under sampling improves precision/recall. Time series cross-validation improves model performance. The model is better at explaining churn than predicting it. It emphasized product-use-related factors that influence churn.

Defect Detection: Measuring and Understanding Customer Churn Models, Scott Neslin et al. JMR AMA ISSN 43 (Apr. 2019),

This article describes how methodological aspects affect customer churn models' accuracy. Academics and practitioners acquired data from a public website, estimated a model, and made predictions on two validation databases. The findings are significant. Methods count. Predictive accuracy differences could impact a churn management campaign's profitability by hundreds of thousands of dollars. Models last.

Wei and Chiu. "Using data mining to anticipate telecom churn," 23 (Aug. 2020),

This approach can identify contract-level churners for a certain projection period. The proposed technique uses a multi-classifier class-combiner to handle the skewed class distribution between churners and non-churners. The empirical evaluation results reveal that the suggested call-behavior-based churn-prediction technique is effective when more current call details are used. The proposed technique shows satisfactory or decent predictive power during one month between model construction and churn prediction. Compared to a previous demographics-based churn-prediction system, our suggested technique achieves satisfactory lift factors.

Hiziroglu and Omer Seymen. "Modeling client attrition using segmentation and data mining" (Jan. 2019),

This research provides a model with multi-dimensions of customer churning level by integrating segmentation and data mining. Compared to other prediction models, the suggested model delivers more accurate predictions on consumer behavior and a better knowledge of customer-company relationships. The report includes the model's managerial and practical consequences.

Ahsan Rehman1, Abbas Raza Ali, 2014 ASE Big Data/Social Informatic, Customer Churn Prediction, Segmentation, and Fraud Detection in Telecommunication Industry.

Phase one presents data-mining strategies for identifying churners and client segmentation based on KPIs (KPIs). In phase two, social network analysis is used to improve traditional learning algorithms at the individual subscriber level. Social network analysis improved Base model findings greatly.

Improved credit card churn prediction using rough clustering and supervised learning Mar. 2018, vol. 21, no. 1, pp. 65–77

C3P solely used supervised categorization techniques for customer retention. But it ended well. We can improve C3P accuracy with hybrid classification techniques. C3P lacks efficient approaches like rough set theory. In this work, we first do data processing, then present a modified rough K-means algorithm for clustering credit card holders, and last, hold-out approach splits the cluster data into testing and training clusters. Finally, classification is done using SVM, RF, DT, KNN, and Naive Bayes. Finally, we examine precision, recall (sensitivity), specification, accuracy, and misclassification error.

An empirical evaluation of approaches addressing the class imbalance problem in churn prediction, Inf. Sci. 408, pp. 84–99, Oct. 2017.

A freshly created projected maximum profit criterion is one of the primary cost-benefit performance indicators. Experiments reveal that evaluation metric affects technique performance. Intra-family comparisons within each solution group and global comparisons of representative procedures from different groups are used to explore reaction patterns to different measurements. Results show there's room to improve solutions' profit-based performance. Our analysis offers academics and professionals useful information and a foundation for developing new churn prediction methodologies.

Profit-maximizing logistic model for customer churn prediction using evolutionary algorithms, Swarm Evol. Computer, vol. 40, pp. 116–130, Jun. 2018.

Our proposed technique builds lucrative churn models for retention initiatives to maximize profits. ProfLogit has the highest out-of-sample EMPC and profit-based precision and recall in a benchmark study with nine real-life data sets. Due to the lasso resemblance, ProfLogit performs profit-based feature selection, selecting features that would otherwise be eliminated using an accuracy-based metric.

Mishra and U. S. Reddy, "A novel approach for churn prediction using deep learning," IEEE Int

In this paper, Deep learning by Convolutional Neural Network (CNN) is implemented for churn prediction and it showed good performance in terms of accuracy. The predictive model for churn prediction has an accuracy of 86.85%, error rate of 13.15%, precision 91.08, recall 93.18%, and F-score 92.06%.

On the operational efficiency of different feature types for telco churn prediction. 2018 Jun; 267(3):1141–1155.

We bridge the gap between prediction performance and operational efficiency by developing a new feature type categorization and a reusable approach to discover optimal feature type combinations using Pareto multi-criteria optimization. Our findings can guide industry practitioners.

Machine learning classifiers predict customer churn in telecommunications. Computer Science, 6 August 2019

The proposed methodology for churn prediction analyzes data, implements machine learning algorithms, evaluates classifiers, and chooses the best one for prediction. Data preprocessing involves cleaning, transforming, and selecting features. Logistic Regression, A.N.N., and

Random Forest are machine learning classifiers. The best classifier was found by evaluating accuracy, precision, recall, and error rate. Logistic regression surpasses AI and random forest, according to this study.

Machine learning on Big Data platform predicts telecom customer churn. April 2020, Muhammad Bin Abubakr, Engineering

The study aims to anticipate telecom customer attrition using massive machine learning data. Estimating churn using machine learning; this study predicts telecom consumer turnover using logistic regression and KNN with large data. Logistic regression is used to estimate churn based on customer characteristics or traits. For churn, K-Nearest Neighbor examines a customer's closeness to customers in every class to determine if they churn. This analysis predicts and analyzes churn using Kaggle data. The study indicated that consumer churn prediction accuracy was 0.80% and AUC was 0.71.

Customer churns prediction in telecommunication business utilizing data certainty, A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar. Jan. 2019 J. Bus. Res. 94:290–301

This research presents a novel CCP approach based on classifier certainty estimation utilizing distance factor. The dataset is separated into two categories: I data with high certainty, and (ii) data with poor certainty, for forecasting Churn and Non-churn behavior. Using different state-of-the-art evaluation measures (e.g., accuracy, f-measure, precision and recall) on publicly available TCI datasets show that I the distance factor is strongly co-related with the certainty of the classifier, and (ii) the classifier obtained high accuracy in the zone with greater distance factor's value (i.e., customer churn and non-churn with high certainty) than those placed in the zone with smaller

Idris and A. Khan, "Customer churn prediction for telecommunication: Using multiple features selection strategies and tree-based ensemble classifiers," Proc.

In this study, we compare tree-based ensemble classifiers with m RMR, Fisher's ratio, and F-score based feature selection strategies for churn prediction in telecom. Large telecommunication datasets are the biggest obstacle to churn prediction model classification performance. Tree-based ensemble classifiers are excellent for larger datasets; however we found rotation forest and rotboost to be more effective than random forest. These techniques boost through feature selection and increase diversity by using linear feature extraction methods like PCA.

3. OBJECTIVES:

The primary objective is to develop a machine learning model that accurately foretells clients' attrition.

Secondly to discover the reasons why clients churn

The purpose is to segment clients so that a marketing strategy may be developed.

As a result, the purpose of churn prediction is to spot clients who are going to abandon the service in advance, so that the service provider can utilize marketing to maintain them.

Problem Statement

The subscription-based business model is continuously growing due to digitalization and offers companies an innovative way of conducting their business .

At the same time, more and more services are being digitalized and data has become much easier to collect, store, and process making a good prediction of climate is always a major task now a day because of the climate change.

There is an abundance of different service providers to choose from, which has increased competition and made it more difficult to retain customers, in this modern-day service market. Due to the availability of data and substitutes, subscription-based businesses must adapt by focusing more on Customer Relationship Management, specifically customer churn management

4. RESEARCH METHODOLOGY

All the study of the Churn related has capitalized by different parameters and suggests various things to develop or improve on these crashes. Various methods apply for reducing the Churning rates on the Telecom Industries by which it help for reducing the fatality rate.

This project, machine learning is used to develop a model that is able to forecast the future before it actually happens. Machine learning is defined as an automated process that mines patterns from a dataset. The art of developing and using models that make predictions based on patterns extracted from data form the past is called predictive data analytics.

Algorithm:

A. Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML.

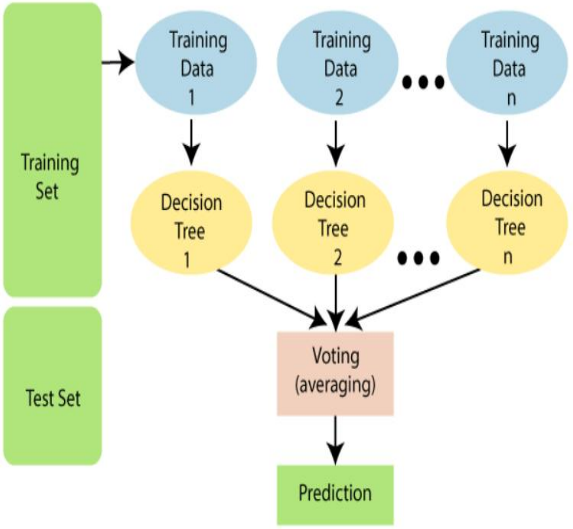


Figure1.1: Random Forest Algorithm

Assumptions for Random Forest:

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

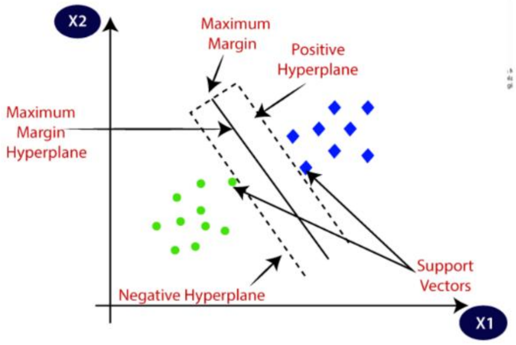
The predictions from each tree must have very low correlations

B. Support Vector Machine Algorithm

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



SVM can be of two types:

Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

C. KNN Algorithm:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors

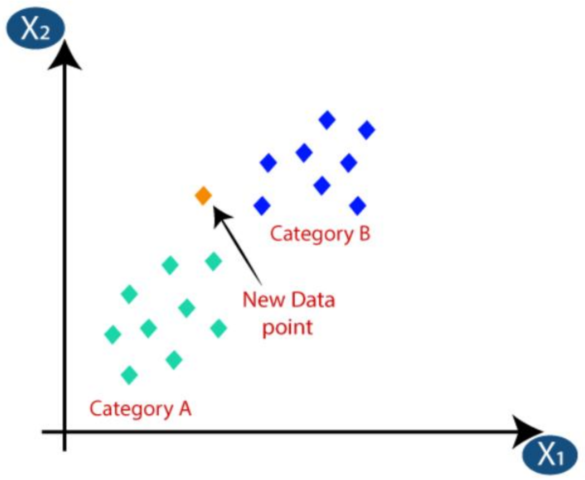
Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

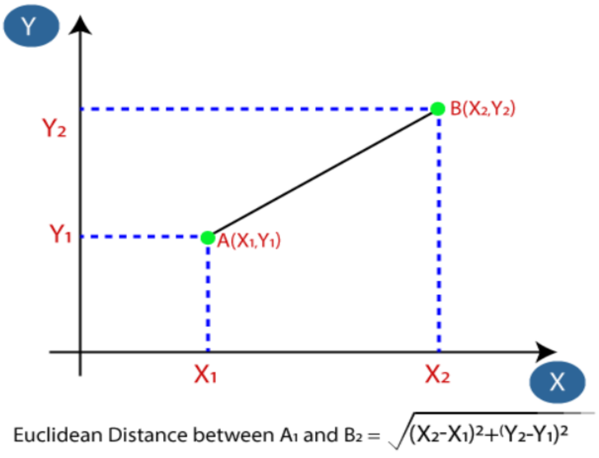
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.



Firstly, we will choose the number of neighbors, so we will choose the k=5.

Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



Training algorithm

For Churn prediction, there are several different optimization algorithms used models development.

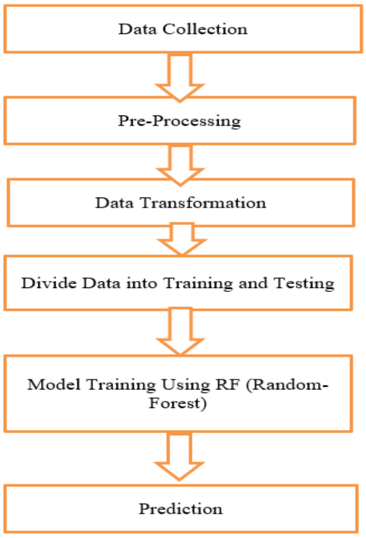

Data Collection
Pre-Processing
Data Transformation
Divide Data into Training and Testing
Model Training Using RF (Random-Forest)
Prediction

Fig 1.2: The Training Architecture

Construction of models

I) Data Preprocessing

1) Noise Removal

It is very important for making the data useful because noisy data can lead to poor results. In telecom dataset, there are a lot of missing values, incorrect values like ''Null'' and imbalance attributes in the dataset.

2) Feature selection

Feature Selection is a crucial step for selecting the relevant features from a dataset based on domain knowledge. A number of techniques exist in the literature for feature selection in the context of churn predictions.

Ii) Training the Network

The primary goal of training is to minimize an error using error techniques. In this Project We will training machine learning model using random forest algorithm. this is classifier algorithm .

Iii) Testing

Through this process, the RF fined the predicted and compares it with the input values using data that was not used in training or validation process. At this stage no adjustment occurs to weights.

Software and evaluation of model performance

Most of the researchers used SPSS, MATLAB and MS Excel software in developing models for the prediction of churn, which is also used for machine learning, signal and image processing, etc. Statistical software R, Minitab software is also used in models development for churn prediction by some researchers. More recently, Python is also used and has become more popular in machine learning and AI.

Customer Classification and Prediction

There are two types of customers in the telecom dataset. First, are the non-churn customers; they remain loyal to the company and are rarely affected by the competitor companies. The second type is churn customers. The proposed model targets churn customers and identify the reasons behind their migration. Furthermore, it devises retention strategies to overcome the problem of switching to other companies. In this study, a range of machine learning techniques is used for classifying customers' data using the labeled datasets. It is to assess which of the algorithm best classifies the customers into the churn and non-churn categories. First, the decision tree algorithm is used for classification. It is categorized as an eager learning algorithm where training data is generalized to classify new samples.

Performance Evaluation Matrix:

In this study, the proposed churn prediction model is evaluated using accuracy, precision, and recall, f-measure, and ROC area. Equation 1 calculates the accuracy metric. It identifies a number of instances that were correctly classified.

Accuracy = (TP + TN) (TP + TN + FP + FN) (1)

Here ''TN'' stands for True Negative,

''TP'' stands for True Positive,

"FN'' stands for False Negative and

''FP'' stands for False Positive.

TP Rate is also known as sensitivity. It tells us what portion of the data is correctly classified as positive.

For any classifier, the TP rate must be high. TP rate is calculated by using Equation 2.

TP Rate = True Positives Actual Positives (2)

FP Rate tells us which part of the data is incorrectly classified as positive. The result of the FP rate must be low for any classifier. It is calculated by using Equation3.

FP Rate = False Positives Actual Negatives (3)

Accuracy, also known as Positive Predictive Value (PPV), indicates which part of the prediction data is positive. It is calculated by using Equation 4.

Precision = True Positive (True Positive + False Positive) (4)

The recall is another measure for completeness i.e. the true hit of the algorithm. It is the probability that all the relevant instances are selected by the system. The low value of recall means many false negatives. It is calculated by using Equation 5.

Recall = (True Positive) (True Positive + False Negative) (5)

The F-measure value is a trade-off between correctly classifying all the data points and ensuring that each class contains points of only one class. It is calculated by using Equation 6.

F − measure = 2 ∗ Precision ∗ Recall Precision + Recall (6)

ROC area denotes the average performance against all possible cost ratios between FP and FN. If the ROC area value is equal to 1.0, this is a perfect prediction. Similarly, the values 0.5, 0.6, 0.7, 0.8 and 0.9 represent random prediction, bad, moderate, good and superior respectively. Values of ROC areas other than these indicate something is wrong.

DATA COLLECTION

Case Study

Our first customer data collection is from a telecom business, where we can review gender, age, tenure, balance, amount of items subscribed to, predicted wage, and if they terminated the subscription or not.

Churn is a telecom sector concern. According to research, the top 4 US wireless carriers have a 1.9%-2% monthly churn rate.

In this project, we'll analyze consumer data to boost customer retention using data insights and predictive modeling. Python and machine learning will be used for analysis.

Dataset and Data collection

We used a Kaggle public bank data set with 10127 customer records and 21 characteristics. Attrition Flag characteristic signals churn or not; 8500 are not churners and 1627 are, with percentages of 16.1% and 83.9%.

Feature Engineering/Data Scrubbing

This study preprocessed data five times. We dropped two irrelevant features first. The missing value analysis detected no missing values. We searched our dataset for duplicates and found none. Classifiers used for analysis convert object data to numerical data. Because the dataset was unequal, SMOTE was used to balance it. Cross-validation prevents over fitting. K-fold cross-validation

Exploratory data Analysis

Training and testing have three gender divisions. U, M, F (unknown). 5% of test and train cases are Unknown. Unknown cases have two explanations. 1) Non-disclosing customers. 2) Input or quality difficulties causing data loss. Unknown cases are difficult to divide between two scenarios. I assumed unknown cases were persons who didn't identify their gender. This suggests that this group may behave differently than those who identified.

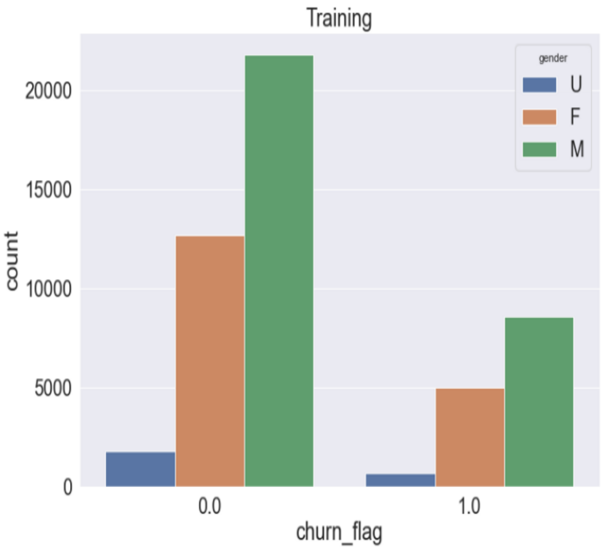


Figure: Gender distribution across churn flag

12 clients in the training data have missing or unknown country of residence. These customers have missing or unknown country of residence. This little percentage of the training set is likely due to input or collection errors. It's unlikely that a customer would join up without stating their country of residence. These observations won't be included in training data.

Some training data clients have join dates of 1901, indicating data quality issues. These observations weren't analyzed.

One 150-year-old consumer was tested. As the oldest person is 117, this is likely a data input error. Observations with ages more than 117 were replaced with the average customer age in each country. The ludicrous age observations have been corrected.

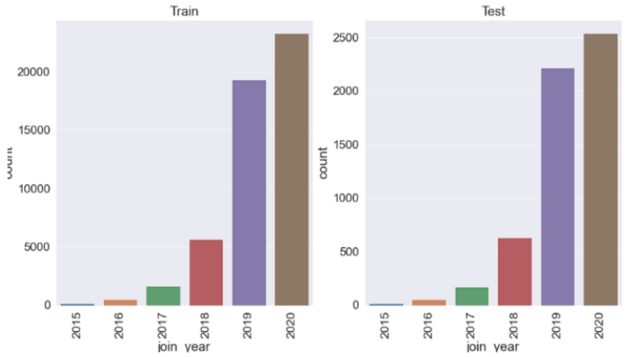
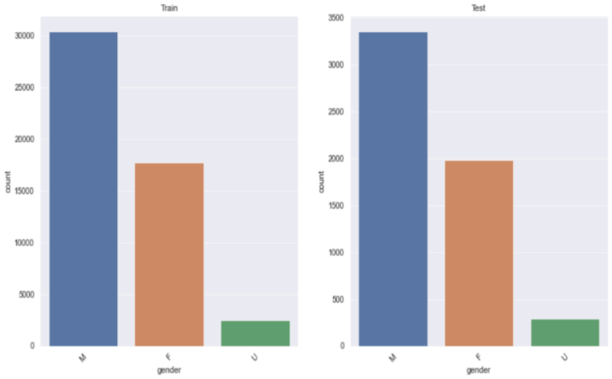


Figure: Join year Distribution

Train and Test data: churn flags separated training and testing. This seems sense given the purpose of inferring missing churn metrics. Verify that the exam population is representative of the training population. We can see this better by comparing the test set and training set distributions for all of our attributes. Visually, the test set seems to be a subset of the training set, which is what machine learning need



One Hot Encoding

Machine learning models interpret numbers not words. Our customer data has categorical variables that must be transformed such that the machine learning models are able to work with them. This is done via one hot encoding.

Modelling

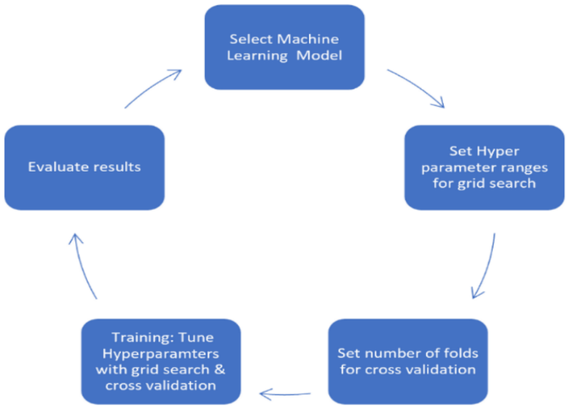
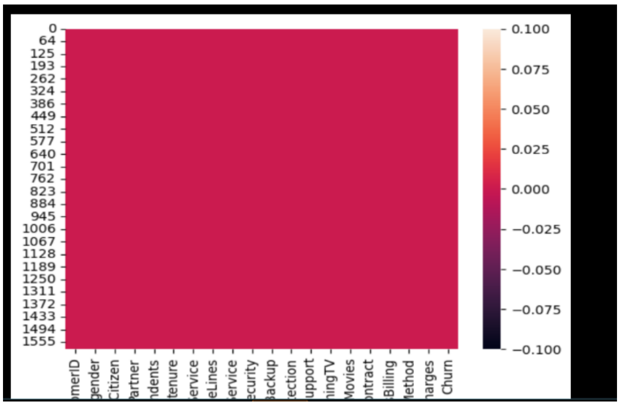

Select Machine Learning Model
Set Hyper parameter ranges for grid search
Set number of folds for cross validation
Training: Tune Hyperparamters with grid search & cross validation
Evaluate results

Figure: Machine Learning Pipeline

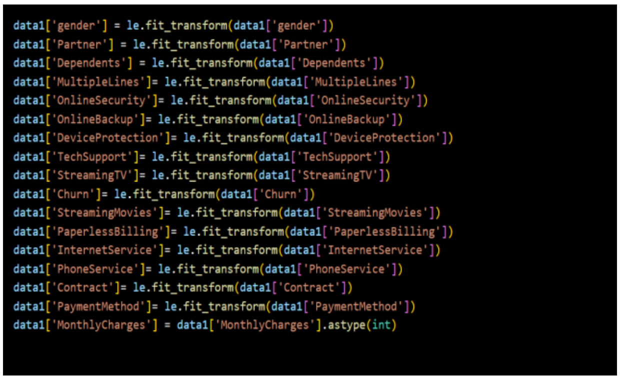
Four machine learning models have been trained following a standard modelling pipeline. A model is selected, and hyper parameters are carefully tuned with a combination of grid search and 5-fold cross validation.

Results:

Data Clean:

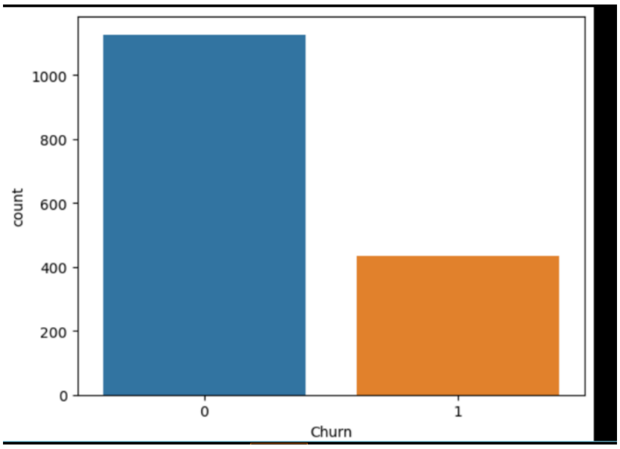


Categorical to Numerical Conversion:

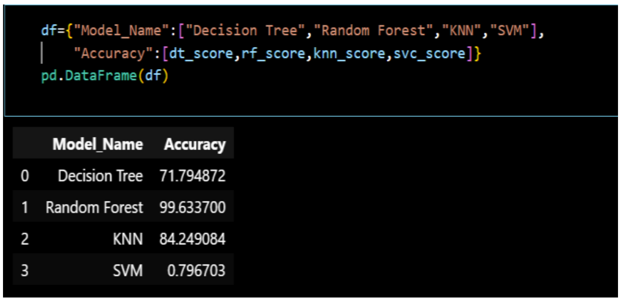


```
data1['gender'] = le.fit_transform(data1['gender'])
data1['Partner'] = le.fit_transform(data1['Partner'])
data1['Dependents'] = le.fit_transform(data1['Dependents'])
data1['MultipleLines']= le.fit_transform(data1['MultipleLines'])
data1['OnlineSecurity']= le.fit_transform(data1['OnlineSecurity'])
data1['OnlineBackup']= le.fit_transform(data1['OnlineBackup'])
data1['DeviceProtection']= le.fit_transform(data1['DeviceProtection'])
data1['TechSupport']= le.fit_transform(data1['TechSupport'])
data1['StreamingTV']= le.fit_transform(data1['StreamingTV'])
data1['Churn']= le.fit_transform(data1['Churn'])
data1['StreamingMovies']= le.fit_transform(data1['StreamingMovies'])
data1['PaperlessBilling']= le.fit_transform(data1['PaperlessBilling'])
data1['InternetService']= le.fit_transform(data1['InternetService'])
data1['PhoneService']= le.fit_transform(data1['PhoneService'])
data1['Contract']= le.fit_transform(data1['Contract'])
data1['PaymentMethod']= le.fit_transform(data1['PaymentMethod'])
data1['MonthlyCharges'] = data1['MonthlyCharges'].astype(int)
```

Count Plot:



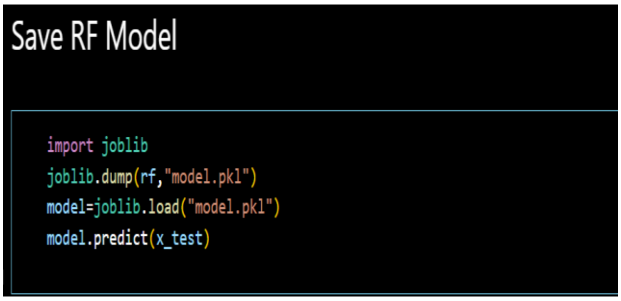
Model Accuracy:



```
df={"Model_Name":["Decision Tree","Random Forest","KNN","SVM"],
    "Accuracy":[dt_score,rf_score,knn_score,svc_score]}
pd.DataFrame(df)
```

	Model_Name	Accuracy
0	Decision Tree	71.794872
1	Random Forest	99.633700
2	KNN	84.249084
3	SVM	0.796703

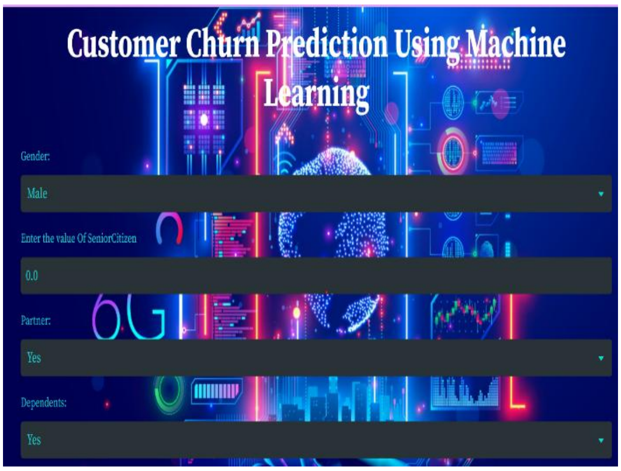
Model Ssaving:



Save RF Model

```
import joblib
joblib.dump(rf,"model.pkl")
model=joblib.load("model.pkl")
model.predict(x_test)
```

GUI:



Customer Churn Prediction Using Machine Learning

Gender:

Male

Enter the value Of SeniorCitizen

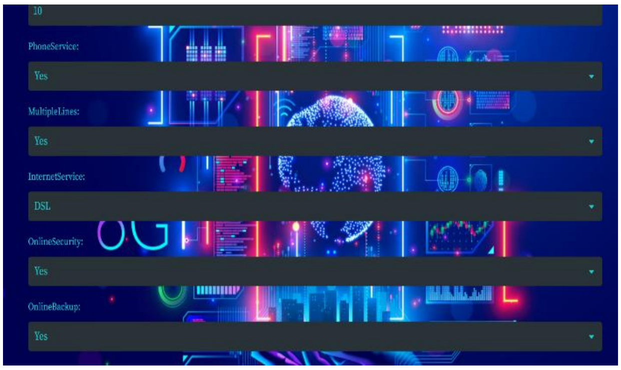
0.0

Partner:

Yes

Dependents:

Yes



10

PhoneService:

Yes

MultipleLines:

Yes

InternetService:

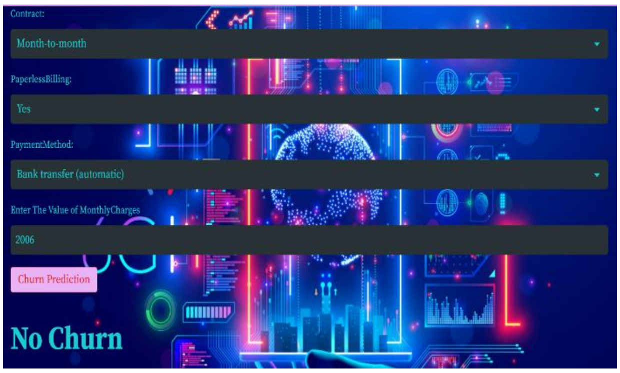
DSL

OnlineSecurity:

Yes

OnlineBackup:

Yes



Contract:

Month-to-month

PaperlessBilling:

Yes

PaymentMethod:

Bank transfer (automatic)

Enter The Value of MonthlyCharges

2006

No Churn

Scope of Project

It's a predictive model that estimates at the level of individual customers the propensity (or susceptibility) they have to leave. For each customer at any given time, it tells us how high the risk is of losing them in the future.

It is important to note that this is the probability of belonging to the group of clients who leave. Thus, it is the propensity to leave and not the probability of leaving. However, it is possible to estimate the probability through a churn model.

5. CONCLUSION:

In conclusion, customer churn prediction using machine learning is a valuable tool for businesses to proactively identify customers who are likely to churn and take appropriate actions to retain them. By leveraging historical customer data and relevant features, machine learning models can be trained to accurately predict churn probabilities or binary churn outcomes. Through the process of data collection, preprocessing, feature engineering, model selection, training, evaluation, and optimization, businesses can develop a predictive model that aligns with their specific requirements. The chosen machine learning algorithms, such as logistic regression, random forests, gradient boosting, or support vector machines, are trained on historical data to learn patterns and relationships that can indicate potential churn..

REFERENCES

1. S. Babu, D. N. Ananthanarayanan, and V. Ramesh, ''A survey on factors impacting churn in telecommunication using datamininig techniques,'' Int. J. Eng. Res. Technol., vol. 3, no. 3, pp. 1745–1748, Mar. 2014.
2. C. Geppert, ''Customer churn management: Retaining high-margin customers with customer relationship management techniques,'' KPMG & Associates Yarhands Dissou Arthur/Kwaku Ahenkrah/David Asamoah, 2002.
3. W. Verbeke, D. Martens, C. Mues, and B. Baesens, ''Building comprehensible customer churn prediction models with advanced rule induction techniques,'' Expert Syst. Appl., vol. 38, no. 3, pp. 2354–2364, Mar. 2011.
4. Y. Huang, B. Huang, and M.-T. Kechadi, ''A rule-based method for customer churn prediction in telecommunication services,'' in Proc. Pacific– Asia Conf. Knowl. Discovery Data Mining. Berlin, Germany: Springer, 2011, pp. 411–422.
5. A. Idris and A. Khan, ''Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers,'' in Proc. 15th Int. Multitopic Conf., Dec. 2012, pp. 23–27.
6. M. Kaur, K. Singh, and N. Sharma, ''Data mining as a tool to predict the churn behaviour among Indian bank customers,'' Int. J. Recent Innov. Trends Comput. Commun., vol. 1, no. 9, pp. 720–725, Sep. 2013.
7. V. L. Miguéis, D. van den Poel, A. S. Camanho, and J. F. e Cunha, ''Modeling partial customer churn: On the value of first product-category purchase sequences,'' Expert Syst. Appl., vol. 12, no. 12, pp. 11250–11256, Sep. 2012.
8. D. Manzano-Machob, ''The architecture of a churn prediction system based on stream mining,'' in Proc. Artif. Intell. Res. Develop., 16th Int. Conf. Catalan Assoc. Artif. Intell., vol. 256, Oct. 2013, p. 157.
9. P. T. Kotler, Marketing Management: Analysis, Planning, Implementation and Control. London, U.K.: Prentice-Hall, 1994.
10. F. F. Reichheld and W. E. Sasser, Jr., ''Zero defections: Quality comes to services,'' Harvard Bus. Rev., vol. 68, no. 5, pp. 105–111, 1990.
11. J. Hadden, A. Tiwari, R. Roy, and D. Ruta, ''Computer assisted customer churn management: State-of-the-art and future trends,'' Comput. Oper. Res., vol. 34, no. 10, pp. 2902–2917, Oct. 2007.
12. H.-S. Kim and C.-H. Yoon, ''Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market,'' Telecommun. Policy, vol. 28, nos. 9–10, pp. 751–765, Nov. 2004.
13. Y. Huang and T. Kechadi, ''An effective hybrid learning system for telecommunication churn prediction,'' Expert Syst. Appl., vol. 40, no. 14, pp. 5635–5647, Oct. 2013.
14. A. Sharma and P. K. Kumar. (Sep. 2013). ''A neural network based approach for predicting customer churn in cellular network services.'' [Online]. Available: https://arxiv.org/abs/1309.3945
15. Ö. G. Ali and U. Aritürk, ''Dynamic churn prediction framework with more effective use of rare event data: The case of private banking,'' Expert Syst. Appl., vol. 41, no. 17, pp. 7889–7903, Dec. 2014.
16. A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, ''Customer churn prediction in telecommunication industry using data certainty,'' J. Bus. Res., vol. 94, pp. 290–301, Jan. 2019.
17. S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, ''Telecommunication subscribers' churn prediction model using machine learning,'' in Proc. 8th Int. Conf. Digit. Inf. Manage., Sep. 2013, pp. 131–136.
18. V. Lazarov and M. Capota, ''Churn prediction,'' Bus.

Anal. Course, TUM Comput. Sci, Technische Univ. München, Tech. Rep., 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.7201&rep=rep1&type=pdf

19. R. Vadakattu, B. Panda, S. Narayan, and H. Godhia, ‘‘Enterprise subscription churn prediction,’’ in Proc. IEEE Int. Conf. Big Data, Nov. 2015, pp. 1317–1321.
20. V. Umayaparvathi and K. Iyakutti, ‘‘Applications of data mining techniques in telecom churn prediction,’’ Int. J. Comput. Appl., vol. 42, no. 20, pp. 5–9, Mar. 2012.
21. A. T. Jahromi, M. Moeini, I. Akbari, and A. Akbarzadeh, ‘‘A dual-step multi-algorithm approach for churn prediction in pre-paid telecommunications service providers,’’ J. Innov. Sustainab., vol. 1, no. 2, pp. 2179–3565, 2010.
22. V. Yeshwanth, V. V. Raj, and M. Saravanan, ‘‘Evolutionary churn prediction in mobile networks using hybrid learning,’’ in Proc. 25th Int. FLAIRS Conf., Mar. 2011, pp. 471–476.
23. G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, ‘‘Credit card churn forecasting by logistic regression and decision tree,’’ Expert Syst. Appl., vol. 38, no. 12, pp. 15273–15285, Nov./Dec. 2011.
24. S. V. Nath and R. S. Behara, ‘‘Customer churn analysis in the wireless industry: A data mining approach,’’ in Proc. Annu. Meeting Decis. Sci. Inst., vol. 561, Nov. 2003, pp. 505–510.
25. Y. Zhang, J. Qi, H. Shu, and J. Cao, ‘‘A hybrid KNN-LR classifier and its application in customer churn prediction,’’ in Proc. IEEE Int. Conf. Syst., Man Cybern., Oct. 2007, pp. 3265–3269.
26. H. Yu et al., ‘‘Feature engineering and classifier ensemble for KDD cup 2010,’’ Dept. Comput. Sci. Inf. Eng., National Taiwan Univ., Taipei, Taiwan, Tech. Rep., 2010, vol. 1, pp. 1–16. [27]
27. L. Zhao, Q. Gao, X. Dong, A. Dong, and X. Dong, ‘‘K-local maximum margin feature extraction algorithm for churn prediction in telecom,’’ Cluster Comput., vol. 20, no. 2, pp. 1401–1409, Jun. 2017.
28. G. Holmes, A. Donkin, and I. H. Witten, ‘‘WEKA: A machine learning workbench,’’ in Proc. Austral. New Zealnd Intell. Inf. Syst. Conf., Dec. 1994, pp. 357–361.
29. F. S. Gharehchopogh and S. R. Khaze. (2013). ‘‘Data mining application for cyber space users tendency in blog writing: A case study.’’ [Online]. Available: https://arxiv.org/abs/1307.7432
30. J. Vijaya and E. Sivasankar, ‘‘An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing,’’ Cluster Comput., pp. 1–12, Sep. 2017. doi: 10.1007/s10586-017-1172-1.
31. A. Amin et al., ‘‘Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods,’’ Int. J. Inf. Manage., vol. 46, pp. 304–319, Jun. 2019.
32. A. Amin, B. Shah, A. M. Khattak, T. Baker, H. ur Rahman Durani, and S. Anwar, ‘‘Just-in-time customer churn prediction: With and without data transformation,’’ in Proc. IEEE Congr. Evol. Comput., Jul. 2018, pp. 1–6.
33. A. Amin et al., ‘‘Customer churn prediction in the telecommunication sector using a rough set approach,’’ Neurocomputing, vol. 237, pp. 242–254, May 2017. 60148 VOLUME 7, 2019 I. Ullah et al.: Churn Prediction Model Using R
34. A. Amin et al., ‘‘Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study,’’ IEEE Access, vol. 4, pp. 7940–7957, 2016.
35. M. Ahmed, H. Afzal, A. Majeed, and B. Khan, ‘‘A survey of evolution in predictive models and impacting factors in customer churn,’’ Adv. Data Sci. Adapt. Anal., vol. 9, no. 3, Jul. 2017, Art. no. 1750007.
36. R. Rajamohamed and J. Manokaran, ‘‘Improved credit card churn prediction based on rough clustering and supervised learning techniques,’’ Cluster Comput., vol. 21, no. 1, pp. 65–77, Mar. 2018.
37. B. Zhu, B. Baesens, and S. K. van den Broucke, ‘‘An empirical comparison of techniques for the class imbalance problem in churn prediction,’’ Inf. Sci., vol. 408, pp. 84–99, Oct. 2017.
38. E. Stripling, S. van den Broucke, K. Antonio, B. Baesens, and M. Snoeck, ‘‘Profit maximizing logistic model for customer churn prediction using genetic algorithms,’’ Swarm Evol. Comput., vol. 40, pp. 116–130, Jun. 2018.
39. A. Mishra and U. S. Reddy, ‘‘A novel approach for churn prediction using deep learning,’’ in Proc. IEEE Int. Conf. Comput. Intell. Comput. Res., Dec. 2017, pp. 1–4.
40. S. Mitrović, B. Baesens, W. Lemahieu, and J. D. Weerdt, ‘‘On the operational efficiency of different feature types for telco churn prediction,’’ Eur. J. Oper. Res., vol. 267, no. 3, pp. 1141–1155, Jun. 2018.
41. A. D. Caigny, K. Coussement, and K. W. D. Bock, ‘‘A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees,’’ Eur. J. Oper. Res., vol. 269, no. 2, pp. 760–772, Sep. 2018.
42. M. Hassouna, A. Tarhini, T. Elyas, and M. S. AbouTrab. (Jan. 2016). ‘‘Customer churn in mobile markets a comparison of techniques.’’ [Online].
43. Available: https://arxiv.org/abs/1607.07792

.

.