A Comparison of Various Class Imbalance and Dimensionality Reduction Techniques on Customer Churn Prediction

Submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

Submitted by

Mada Datta Sai Sri Phani Bhushan (AP19110010110)

Nandini Thimmireddy Gari (AP19110010128)

Thondapu Sneha Teja Sree Reddy (AP19110010124)

Rahul Tata (AP19110010051)

Research Supervisor

Saleti Sumalatha

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SRM UNIVERSITY AP, ANDHRA PRADESH

GUNTUR– 522502, INDIA

APRIL 2022

Table of Contents

A Comparison Of Various Class Imbalance and Dimensionality Reduction Techniques on Customer Churn Prediction

Sri Phani Bhushan Mada
CSE Dept
SRM University-AP
Andhra Pradesh
India.
phanibhushandss_mada@srmap.edu.in

Nandini Thimmireddygari
CSE Dept
SRM University-AP
Andhra Pradesh
India
nandini_t@srmap.edu.in

Rahul Tata
CSE Dept
SRM University-AP
Andhra Pradesh
India.
rahul_tata@srmap.edu.in

Sneha Teja Sree Reddy Thondapu
CSE Dept
SRM University-AP
Andhra Pradesh
India.
snehatejasreereddy_thondapu@srmap.edu.in

Sumalatha Saleti
CSE Dept
SRM University-AP
Andhra Pradesh
India.
sumalatha.s@srmap.edu.in

***Abstract*—With the advancement of technology, companies are able to foresee the customers who are going to leave their organization much before. Our main aim is to compare the performance of machine learning models with a combination of different class imbalance and dimensionality reduction techniques. Our proposed methodology comprises of six phases. In the first two phases, we performed data selection and data preprocessing. In the third phase and fourth phases, we performed class imbalance and dimensionality reduction methods. Next, we implemented machine learning models like Logistic regression(LR), Decision tree, Naive Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and a boosting algorithm called ADA-Boost. Further, we evaluated the performance of the models using AUC curves, Confusion matrix, and K-fold cross-validation techniques. It is observed that for SMOTE class imbalance technique, ADA Boost performed better under both feature selection and feature extraction and for the remaining class imbalance techniques such as TOMEK and Random Oversampling, KNN performed best followed by ADA Boost and SVM.**

I. Introduction

Gaining a competitive advantage is essential for businesses in today's world. The companies need to focus on profit maximization to sustain this competitive market. Customer retention is a less economic burden to the companies when compared to other plans like gaining new customers and making the customer buy new products. Thus, the organization must take necessary steps to satisfy the customer and encourage them to continue utilizing their services. If a customer is unsatisfied with the assistance or quality of a service provided by the companies, then the customer is more likely to shift to other competitors which in turn is a loss for the company. Customer churn is the word that is being used extensively to address this issue. It indicates that the customer has canceled his or her subscription with the company and has joined another for some reason.

The principal causes for customer churn are the lack of proper service from the companies, better price from the competitors and it can be due to client-side problems. Churn occurs in two ways. One is Voluntary and other way is involuntary. Involuntary churn happens when the customer optionally quits the company and shifts to another company. This type of churn only happens when the customer is dissatisfied with reasons like non-payments, policies of the company and bad quality of the product or services. In Voluntary churn, customers may leave the company in two ways. Firstly, due to personal reasons of the customers such as a change in financial circumstances or current location may lead to incidental voluntary churn. Deliberate churn is the second way that happens due to issues related to technology such as customers wanting advancement in technology, quality of service and other psychological factors.

There are two ways to handle this problem: one is proactive and the other is reactive. In the reactive approach, the company waits until the customer cancels a subscription but in the proactive approach the customer churn is predicted before the customer leaves the company. If the companies can identify the customers and categorize them into churners and non-churners then it would be easy for them to target that churner group and offer them incentives and improve relationships. Because there was not much technology in the past, companies used to offer incentives to customers after they left the company in order to retain their services i.e. reactive approach. However, in today's modern world, companies can predict customer churn way in advance because of technological advancements i.e. proactive approach.

This prediction process will reduce the churn and accelerate companies' profit. This predictive process depends on data mining techniques because of their efficiency and accuracy. The data mining approach uses machine learning algorithms and statistics to extract knowledge from the data. These techniques can be applied to further develop predictive models, identify the behaviors of customers and allow companies to make better decisions based on the knowledge extracted from the data. This whole process can be identified as a classification task in the machine learning approach.

We can classify the products with different levels of demand for a product, divide a product for a certain amount of time, the spending of a customer, etc. By solving the above sub problems, we can order products at particular times at required quantities so that we will be able to keep the customer without increasing the cost to keep a customer with the company, as a result, increasing profit.

II. Literature Review

Wouter Verbeke et al [2011] [1], states that predictive accuracy, comprehensibility and justifiability are three main aspects of customer churn prediction model. Advanced rule induction rule mining techniques such as AntMiner+ and ALBA are used on dataset and measured to traditional rule induction techniques such as C4.5 and RIPPER. The Results show that ALBA improves learning of classification techniques whereas AntMiner+ is accurate, comprehensible and justifiable in the study.

Shin-Yuan Hung et al [2006] [2], provided empirical evaluation of various data mining techniques and found out that decision tree and neural network techniques show more accurate churn prediction models by using customer structured data.

Kristof Coussement et al [2009] [3], focused on two main aspects of improving churn prediction that are relying on a wide variety of customer types and selecting most effective classification techniques. In the paper, Logistic Regression, Support Vector Machines and Random Forests techniques are used and shows that random forests technique is more suitable for churn prediction.

Nilam Nur Amir Sjarif et al [2019] [4], uses K-Nearest algorithm and Pearson correlation function to propose an effective churn prediction model. On dividing the obtained dataset into training and testing in a ratio of 70:30, results shown that K-nearest algorithm showed more accuracy than others.

Faris et al [2018] [5], author proposed an intelligent hybrid model with Particle swarm Optimization and Feedforward neural network for churn prediction. The model is easy to understand and also greatly improves the coverage rate of churn customers in contrast with other classifier techniques.

Arno De Caigny et al [2018] [6], author proposed a new hybrid algorithm called logit leaf model which is benchmarked against decision trees, logistic regression, random forests and logistic model trees. This hybrid model showed more comprehensibility than its parent models logistic regression and decision trees.

Abdelrahim Kasem Ahmad et al [2019] [7], developed a model using four algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM" and Extreme Gradient Boosting "XGBOOST" using Area under curve and Social Network Analysis features to compute the performance of the model. The best results are obtained by using the Xgboost algorithm. AUC value obtained is 93.3% whereas Social Network Analysis(SNA) enhanced the model performance from 84 to 93.3% against AUC standard.

Ali Rodan [2014] [8], used the most parametric model: Support vector machine for customer churn prediction. The dataset is collected from a local telecommunication industry. The SVM model is compared with several machine learning models like Multilayer Perceptron Neural Network with back propagation learning algorithm, k-Nearest Neighbour, Naive Bayes, C4.5 Decision Trees algorithm and the best accuracy is obtained by SVM with 98.7%.

Pınar Kisioglu [2011] [9], developed a model using Bayesian belief Networks and used CHAID (Chi-squared Automatic Interaction Detector) algorithm to discretize continuous variables. At last, authors suggested three scenarios to investigate the characteristics of churners.

E.W.T. Ngai et al[2008] [10], proposed that the application of data mining techniques helped in forming an effective CRM(Customer relationship management) and mentioned the steps to carefully divide the research papers according to the various CRM framework and data mining applications with the help of association rule and decision tree.

Jonathan Burez et al[2007] [11] mentioned about a European pay TV company that offers premium content that is suddenly suffering from a high customer churn rate. The dataset is collected and different binary classification techniques such as Markov chain, logistic regression and random forest are applied from which an appropriate model is chosen.

John Haden et al[2005] [12], proposed detailed steps on how to reduce the customer churn rate and retain existing customers that include identification of best data, data semantics, feature selection, validating results and development of predictive model. Various methods are mentioned to accurately and effectively follow the above mentioned steps.

[13] [14] [15] [16] [17] [18] [19] [20]

III. Preliminaries

In this section of the paper, we have attempted to describe the notations, abbreviations, data cleaning, preprocessing techniques that are used to make predictions more robust and machine learning models used for categorization.

A. CLASS IMBALANCE

A dataset apparently comprises of class imbalance when one class has higher number of observations than other classes. It is becoming a common issue since the volume of data is increasing and mainly affects classification algorithms.

*1) **Synthetic Minority Oversampling Technique (SMOTE)**:* This works by selecting a point from minority class and computing the nearest points for these points and add a synthetic point between them. In this method we will select the minority class data point and find all the nearest points from that minority class data point and draw lines between them. We add minority class points on that lines until count of both data points become same. Since we are adding data points it is called Oversampling technique.

*2) **Tomek Links**:* Tomek Links is a Tomek-developed variant of the Condensed Nearest Neighbours under sampling approach (1976). Unlike the Condensed Nearest Neighbours technique, which chooses samples with nearest neighbours from the majority class that needs to be deleted at random, the Tomek Links method applies a rule to select pairs of observations that is the pair of belonging to two different classes and find the shortest distance. Among those two data points, we remove majority class samples. Since the total points are decreasing it comes under Undersampling.

*3) **Random Over sampling**:* Oversampling is defined as boosting the number of copies in the minority class. When the dataset has less number of entries, oversampling might be a smart option. Replacement is used to select examples from the training dataset at random. This means that minority class examples can be selected and added to the new "more balanced" training dataset multiple times; they are selected from the original training dataset, added to the new training dataset, and then returned or "replaced" in the original dataset, allowing them to be selected again. This method can help machine learning algorithms that are influenced by skewed distributions and where several duplicate samples for a particular class might damage the model's fit.

B. DIMENSIONALITY REDUCTION

It is defined as the process of transforming a dataset having vast dimensions into a dataset with fewer dimensions. It can be done in two ways.

1. Feature selection: It is the process of selecting essential features from the given dataset.

2. Feature extraction: It creates new set of features from the existing features.

There are 2 popular feature extraction techniques that are available. The first one is PCA(Principal Component Analysis) and LDA(Linear Discriminant Analysis). Feature selection techniques include filler, wrapper and embedded techniques.

*1) **Feature extraction:**:* Principal Component Analysis (PCA) is a well-known dimensionality reduction technique that helps in identifying correlations and patterns in a data set and converts them into a dataset of fewer features without losing any important information. The new set of variables that are transformed is called principal components. These principal components are linear combinations of original variables and are orthogonal. PCA removes inconsistencies in the data, redundant data, and highly-correlated features from the original dataset. There is a need for principal component analysis due to dimensionality reduction. High dimensional data that is used in image processing, image translation, and natural language processing is extremely complex to process due to the instability and the variability in the features that increase the computational time and make data processing and exploratory data analysis(EDA) difficult. The steps involved in this algorithm involve gathering the dataset, computing the mean vector, and covariance matrix, calculating the eigenvalues, eigenvectors and normalizing eigenvectors, choosing components, and forming a feature vector finally we derive a new dataset and use it for further analysis.

*2) **Feature selection:**:* Some of the predictive models have vast number of variables that requires a large amount of system memory and time. The performance may not be effective if the input variables are not relevant to the target variable. So we use feature selection to reduce the number of input variables which makes the model fast and effective. The main objective of feature selection is to find the best set of features that allow us to build the best user model. In this research, we have used embedded method for feature selection. In embedded methods, interactions of features are included and also maintain reasonable computational cost. In each recursive step of the tree growth process, tree algorithms chooses a feature and divides the sample set into smaller subsets. The more child nodes in the same class in a subset, the more informative the features become. The most common embedded techniques are the tree algorithms like the random forest, Decision tree, XGboost, etc. In this paper, random forest is used for feature selection. Random Forests are a type of Bagging Algorithm that aggregates a set of decision trees. Random forests' tree-based tactics are naturally ranked by how well they increase node purity or in other words, how well they reduce impurity over all trees. The nodes with the largest drop in impurity are found at the beginning of the trees, while the notes with the least decrease in impurity are found at the conclusion. We can produce a subset of the most essential features by pruning trees below a specific node.

C. MACHINE LEARNING MODELS

*1) **Decision Tree**:* This technique is commonly used for classification and prediction in both Machine Learning and Data Mining applications because to its excellent characteristics. One of the advantages of employing this strategy is that the Decision Tree contains concepts that are easy to understand and comprehend due to its simplicity and comprehensibility in revealing enormous or microscopic data structures and anticipating them. A Decision Tree Classifier is a flowchart classifier that is similar to a Tree Structure in that it uses a decision tree to classify data.

1. A test on an attribute is represented by a non-leaf node. 2. One of the test's results is represented by the tree branch. 3. The value of the destination attribute identifies a terminating node. 4. In a tree, the root node is at the very top.

Because of the following significant properties, we chose the decision tree algorithm:

1. The decision tree can readily handle high-dimensional data. 2. Trees of small size can be easily understood. 3. The procedures for correctly classifying decision tree induction are simple.

To choose the dividing attribute, we need to find Information gain. Consider D to be the full training dataset.

$$\sum_{i=1}^{m} -p_i log_2 p(i) \tag{1}$$

Gain = Info(D) - Info(attribute)

*2) **Naive Bayes**:* The Naive Bayes method is a supervised learning strategy for dealing with classification problems that is based on the Bayes theorem. It's generally used for jobs that require a large training dataset, such as text classification. The Naive Bayes Classifier is a simple yet efficient classification technique that can help you develop rapid machine learning models that can make accurate predictions. It's a probabilistic classifier, which means it produces predictions based on an object's likelihood. Spam filtration, sentiment analysis, and article classification are all common uses for the Naive Bayes Algorithm. The Bayes'theorem, often known as Bayes'rule or Bayes'law, is a mathematical formula for calculating the likelihood of a hypothesis based on past data. It is conditional probability that determines this. It depends on conditional probability. Bayes Theorem is as follows:

P(X | Y) = P(Y|X) * P(X) / P(Y).

Where P(X|Y) is Posterior probability: Probability of hypothesis X on the observed event Y. P(Y|X) is Likelihood probability: The probability of the evidence given that the probability of a hypothesis is true.

*3) **Logistic regression**:* Regression analysis is the study of relationships between variables. We study the relationship between the input variable is referred to as the explanatory variable and the output variable which is referred to as the dependable variable. It consists of finding the best fitting line through all the points. The best-fitting line is called the regression line. Basic formulas like mean, variance and correlation, and a sum of mean squared error are used. In case of linear regression the outcome is continuous where as in case of logistic regression the the outcome is discrete. In case of logistic regression a function is used to to predict the values. Logistic regression has become particularly popular in online advertising, enabling marketers to predict the likelihood of specific website users who will click on particular advertisements as a yes or no percentage. Logistic regression can also be used in healthcare to identify risk factors for diseases and plan preventive measures, weather forecasting apps, and voting apps.

*4) **Support vector machine**:* A common Supervised Learning technique for handling classification and regression problems is the Support Vector Machine (SVM). However, it is usually utilised to tackle classification problems in Machine Learning. The SVM method's purpose is to find the optimum line or decision boundary for classifying n-dimensional space into classes so that following data points can be conveniently placed in the appropriate category. A hyperplane is a boundary that is the optimum option.

*5) **K−nearest neighbours**:* K- Nearest neighbor classification is a simple algorithm that stores all available cases and classifies new classes based on a similarity measure called distance function. It is used for nonlinear data and has been used in statistical estimation and pattern recognition. It is one of the simplest supervised learning-based machine learning algorithms. KNN algorithm is mostly used for classification. It is non-parametric and is also specified as a lazy learner algorithm. There are steps that are performed in the KNN algorithm: 1. Get the data 2. Select the number of K neighbors. 3. Compute euclidean distance and compare it from new data points to all the other points in the dataset. 4. Take K nearest distances from the calculated Euclidean distance. 5.

Among these K neighbors, count the number of data points in each category. 6. Assign new data points to the category for which the number of neighbors is maximum. 7. Our model is ready The preferred value of K is odd value for efficient results. Values like 1 and 2 for K are not considered as it involves noisy data.

*6) **ADA−Boost**:* ADA is known as Adaptive Boosting which is a Boosting strategy used as an Ensemble Method in Machine learning. Each instance has re-assigned weights, where larger weights applied to instances that were incorrectly recognised. In order to decrease bias and variance, Boost is the best technique to be used. With the exception of the first, each succeeding student is developed from previously developed learners.

For the construction of First decision model/tree, the indelicately categorised record in the first model is given precedence. Only those records are supplied to the next model as input. The procedure continues until the quantity of base learners we wish to produce is specified.

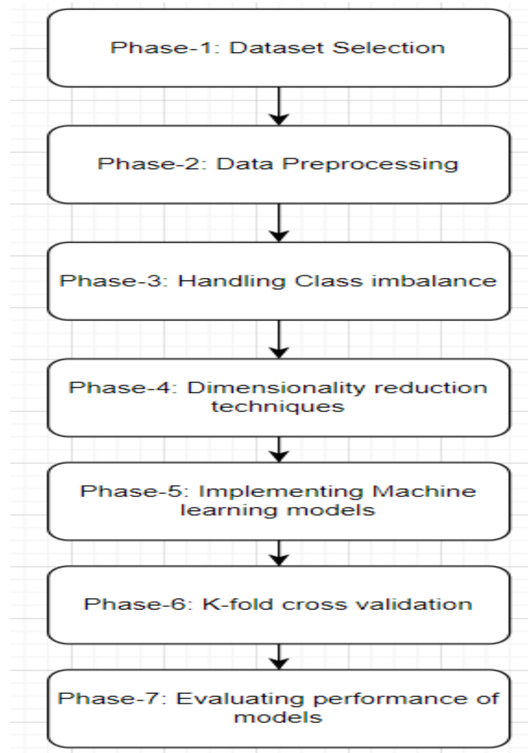
IV. PROPOSED WORK

A. SYSTEM ARCHITECTURE

This part contains the description and algorithm of proposed work.

System Architecture The system architecture contains 7 Steps: 1. Identifying Suitable Data 2. Cleaning & Transforming Data 3. Balancing the Data 4. Feature Selection & Reduction 5. Models used 6. Cross Validation 7. Evaluation of results

Data preprocessing is an important step in data mining which helps to clean the data and remove any outliers to make it more consistent. In this step we also remove the null values in the dataset and we also check and resolve the data for imbalanced class distribution. In our Steps – 1 to Step – 4.



*1) **Identifying the Suitable Dataset**:* In this step we will select dataset which is appropriate to our problem and we also remove the unwanted columns such as serial number, names, unique identifiers. Which we do not need for analysis as the languages itself gives a unique column.

Feature Names	Data Type
state	object
account_length	int64
area_code	object
international_plan	object
voice_mail_plan	object
number_vmail_messages	int64
total_day_minutes	float64
total_day_calls	int64
total_day_charge	float64
total_eve_minutes	float64
total_eve_calls	int64
total_eve_charge	float64
total_night_minutes	float64
total_night_calls	int64
total_night_charge	float64
total_intl_minutes	float64
total_intl_calls	int64
total_intl_charge	float64
number_customer_service_calls	int64
churn	object

*2) **Data Preprocessing**:* The real-world data is subjected to noise and null values. So we need to clean the dataset and apply a machine learning model for better results. First, we checked for any duplicates in the dataset and then went for Data Cleaning which includes removing noise and null values. Also, in this step we converted the columns of the dataset with strings into numeric values in case if the columns two we replaced one value with 1 and another value with 0. If we have more than two values, we used categorical encoder to encode the data. Converting the string data into numerical data is very important to implement all the machine learning models. After encoding we scaled the dataset using Standard Scaler technique for better accuracy of the models.

*3) **Balancing the Imbalanced Classes**:* In this step used different class imbalance techniques such as TOMEK, SMOTE and Random Over Sampling. Class imbalance problem is the one of the major concern in classification problems. Handling class imbalance prevents overfitting of the machine learning models and helps in better prediction of the models.

*4) **Feature Selection and Reduction**:* The main reason for this step is to reduce the size of data. In this step for the feature selection, we used Embedded feature selection which involves random forest classifier for identify the importance of each feature. We performed hyper parameter tuning to find the best parameters for random forest classifier. And we included the features which has feature importance greater than 3 in the final dataset and for feature extraction we used Principal Component Analysis(PCA).

*5) **Models Used**:* In this step predictive models are applied. In order to see how different class balancing techniques and dimensionality reduction techniques effect a model so we will be applying different frequently used models Decision Tree, Naïve Bayes, Logistic Regression, K-nearest neighbours, Support Vector Machine and boosting algorithm named ADA Boost Classifier.

*6) **Cross Validation**:* It's a resampling method for evaluating machine learning models on a small data set. The process only includes one parameter, k, which is the number of splitted groups in a given data sample. The k-fold cross validation method shuffles the data set at random and then divides it into k groups. One group is chosen at random as a test set, while the remaining groups are used as train sets. Following that, the model is fitted and the score is confirmed using previously unknown data. The results of k-fold cross validation are provided in the table below. It turns out that the k-fold cross validation was used to fine-tune the models and keep them from overfitting on the train set.

*7) **Evaluation of Results**:* The key to analysing the performance of the presented model is model evaluation. While evaluating the models we use confusion matrix and the AUC curve. These outputs are compared to find the best model for the data.

Model Name	Accuracy
Decision tree	79.18968055
Naive Bayes	75.71215
Logistic Regression	77.902553
K-Nearest Neighbours	77.9025
Support Vector Machine	70.07131
ADA Boost	88.089336

TABLE I: Accuracies of the models after applying PCA and SMOTE technique

Model Name	Accuracy
Decision tree	79.1759
Naive Bayes	74.84
Logistic Regression	78.09
K-Nearest Neighbours	88.06
Support Vector Machine	86.63
ADA Boost	88.45

TABLE II: Accuracies of the models after applying Embedded feature selection and SMOTE technique

Model Name	Accuracy
Decision tree	83.26
Naive Bayes	81.366
Logistic Regression	78.35
K-Nearest Neighbours	91.31
Support Vector Machine	70.52
ADA Boost	83.78

TABLE III: Accuracies of the models after applying PCA and TOMEK technique

Model Name	Accuracy
Decision tree	82.324
Naive Bayes	80.942
Logistic Regression	77.3822
K-Nearest Neighbours	93.22
Support Vector Machine	87.88
ADA Boost	83.78

TABLE IV: Accuracies of the models after applying Embedded feature selection and TOMEK technique

Model Name	Accuracy
Decision tree	83.269
Naive Bayes	81.366
Logistic Regression	78.35
K-Nearest Neighbours	91.31
Support Vector Machine	83.88
ADA Boost	83.8854

TABLE V: Accuracies of the models after applying PCA and random oversampling technique

Model Name	Accuracy
Decision tree	82.324
Naive Bayes	92.71
Logistic Regression	76.88
K-Nearest Neighbours	92.716
Support Vector Machine	87.91
ADA Boost	83.57

TABLE VI: Accuracies of the models after applying Embedded feature selection and random oversampling technique

V. PERFORMANCE ANALYSIS

A. *Confusion matrix*

On test data, the performance of the models for predicting customer turnover was assessed using a variety of criteria, including precision, recall, accuracy and AUC curves. These are used to assess the models' ability to forecast. The confusion matrix may be used to determine measurements like precision, recall, and accuracy. The performance of a classification model is measured using a M x M matrix, where M is the number of target classes. The matrix compares the actual target values to the predictions of the machine learning model. This provides us with a comprehensive picture of how well our classification model is working and the types of errors it makes. True positive(T_p) refers to a number of churn customers that are correctly predicted by the model. True negative(T_n) refers to a number of negative samples correctly predicted. False-positive(F_p) refers to customers who are wrongly predicted as churners. False negatives(F_n) refer to the number of customers who are wrongly predicted as churners.

Recall = $T_p/T_p + F_n$

Precision = $T_p/T_p + F_p$

Accuracy = $T_p + T_n/T_p + F_p + T_n + F_n$

B. *AUC Curve analysis*

The AUC curve was used to measure the model's performance on the positive and negative classes of the test set. The better the model performs on both positive and negative classes, the higher the AUC score. Fig-2 shows the derived AUC values of predictive models that were used to forecast the churn prediction. From the figure-2 AUC Curves we can see that ADA-Boost performed well for Random over sampling(Feature Extraction) and SMOTE(Feature selection). And KNN performed well in the other combinations.

C. *Obtained outcome analysis*

Different class imbalance approaches, including as SMOTE, TOMEK, and Random over sampling, as well as dimensionality reduction techniques, such as PCA and Embedded feature selection using random forest, were used to assess the final preprocessed data. We used a variety of machine learning models to predict churn, including decision trees, logistic regression, support vector machines, ADA-Boost classifiers, KNNs, and Naive Bayes. It is observed that ADA-Boost technique performed well under Random under sampling(Feature Extraction) and SMOTE(Feature Selection). And KNN performed well in other combination of techniques.

It is observed that for SMOTE class imbalance technique ADA Boost performed better under both feature selection and feature extraction. And for the remaining class imbalance techniques TOMEK and Random oversampling KNN performed best followed by ADA-Boost and SVM. The findings and accuracy scores achieved are listed below in figure-3.

VI. CONCLUSION

The research demonstrates how data science and machine learning algorithms may be used to anticipate customer churn. The dataset consists of 4200 samples acquired from telecommunication firms. Several data mining techniques were utilised to retrieve hidden knowledge. To analyse the data set and forecast the churn rate, we used decision trees, KNN, Logistic Regression, Naive Bayes, ADA-Boost classifier, and SVM. Companies will be able to track down lost consumers and spend in obtaining new ones with this information. Understanding what keeps customers engaged is really helpful knowledge since it may help with the creation of retention strategies and the adoption of operational measures to keep customers from leaving.

REFERENCES

[1] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354–2364, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417410008067

[2] S.-Y. Hung, D. C. Yen, and H.-Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications*, vol. 31, no. 3, pp. 515–524, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417405002654

[3] K. Coussement, D. F. Benoit, and D. Van den Poel, "Improved marketing decision making in a customer churn prediction context using generalized additive models," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2132–2143, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417409007325

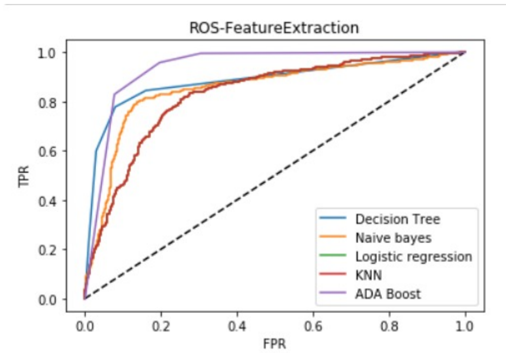
[4] D. H. W. S. Y. R. I. Nilam Nur Amir Sjarif, Muhammad Rusydi Mohd Yusof and M. Z. Osman, "A customer churn prediction using pearson correlation function and k nearest neighbor algorithm for telecommunication industry," *Int. J. Advance Soft Compu. Appl*, vol. 11, no. 2, 2019. [Online]. Available: http://188.247.81.52/PapersUploaded/2019.2.4.pdf

[5] H. Faris, "A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors," *Information*, vol. 9, no. 11, 2018. [Online]. Available: https://www.mdpi.com/2078-2489/9/11/288

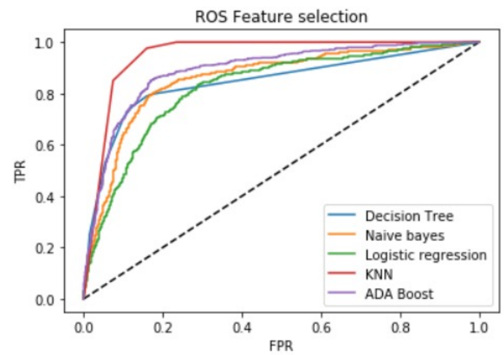
[6] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *European Journal of Operational Research*, vol. 269, no. 2, pp. 760–772, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221718301243

[7] K. A. Abdelrahim Kasem Ahmad, Assef Jafar, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data volume*, vol. 6, no. 28, 2019. [Online]. Available: https://link.springer.com/article/10.1186/s40537-019-0191-6

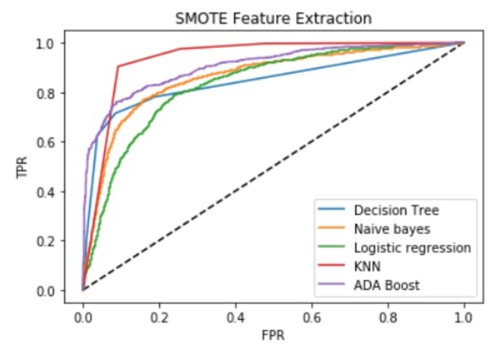
[8] A. Rodan, H. Faris, J. Alsakran, and O. Al-Kadi, "A support vector machine approach for churn prediction in telecom industry," *International journal on information*, vol. 17, no. 8, pp. 3961–3970, 2014.



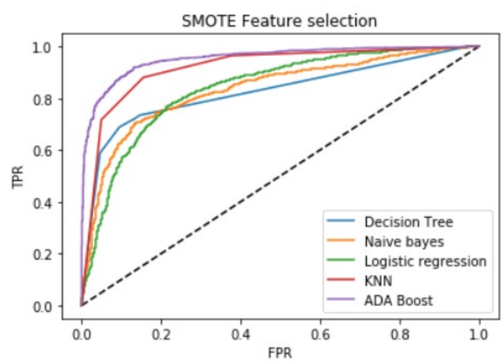
a. ROS Feature Extraction



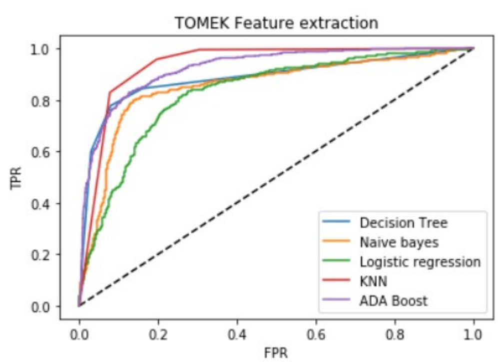
b. ROS Feature Selection



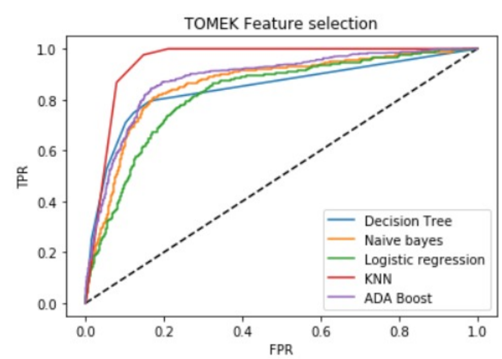
c. SMOTE Feature Extraction



c. SMOTE Feature Selection

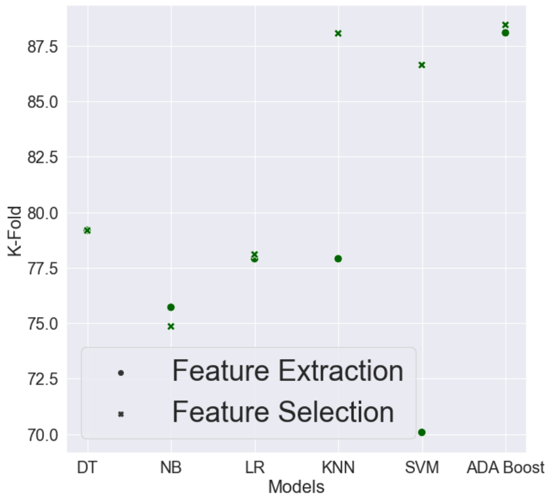


d. TOMEK Feature Selection

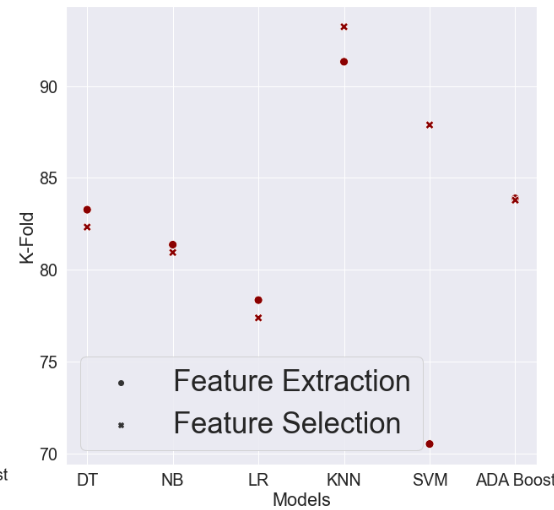


e. TOMEK Feature Extraction

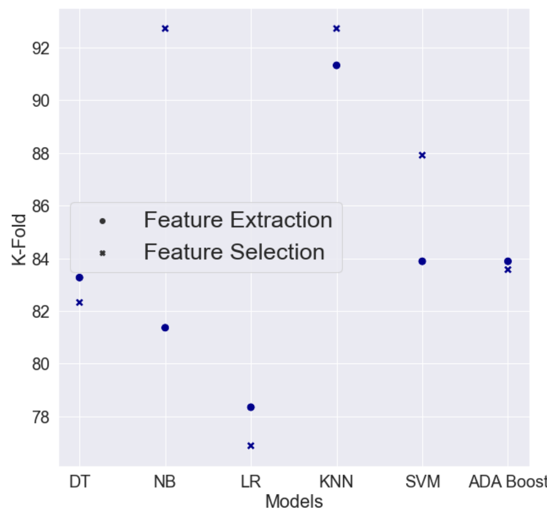
Fig. 2: AUC Curves for Random Over Sampling - Feature Extraction.



a. SMOTE



b. TOMEK



c. ROS

Fig. 3: Accuracy plots.

[9] P. Kisioglu and Y. I. Topcu, "Applying bayesian belief network approach to customer churn analysis: A case study on the telecom industry of turkey," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7151–7157, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417410014089

[10] E. Ngai, L. Xiu, and D. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 2592–2602, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417408001243

[11] J. Burez and D. Van den Poel, "Crm at a pay-tv company: Using analytical models to reduce customer attrition by targeted marketing for subscription services," *Expert Systems with Applications*, vol. 32, no. 2, pp. 277–288, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417405003374

[12] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Computer assisted customer churn management: State-of-the-art and future trends," *Computers & Operations Research*, vol. 34, no. 10, pp. 2902–2917, 2007.

[13] T. C. S. . B. M. T. Renato Alexandre de Lima Lemos, "Propension to customer churn in a financial institution: a machine learning approach," *Neural Computing and Applications*, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s00521-022-07067-x

[14] S. De, P. P, and J. Paulose, "Effective ml techniques to predict customer churn," pp. 895–902, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9544785

[15] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109 960–109 975, 2021.

[16] N. N. Vo, S. Liu, X. Li, and G. Xu, "Leveraging unstructured call log data for customer churn prediction," *Knowledge-Based Systems*, vol. 212, p. 106586, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705120307152

[17] D. Mukhopadhyay, A. Malusare, A. Nandanwar, and S. Sakshi, "An approach to mitigate the risk of customer churn using machine learning algorithms," pp. 133–142, 2021.

[18] A. Shukla, "Application of machine learning and statistics in banking customer churn prediction," pp. 37–41, July 2021.

[19] V. C. Nwaogu and K. Dimililer, "Customer churn prediction for business intelligence using machine learning," pp. 1–7, June 2021.

[20] Y. K. Saheed and M. A. Hambali, "Customer churn prediction in telecom sector with machine learning and information gain filter feature selection algorithms," in *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, Oct 2021, pp. 208–213.