Strategic Analytics for Telecom Churn Prediction using Machine Learning Approaches

MD. IRTIZA HOSSAIN, Brac University, Bangladesh
JUNAID AHMED SIFAT, Brac University, Bangladesh
FARHAN ISHRAQ FAGUN, Brac University, Bangladesh

In the telecommunications market where customer retention is extremely important, churn prediction becomes the key factor to maintain or gain the industry stability and manage the continuing growth. This research is dedicated to the proposing of complex machine-learning algorithms which would enable the prediction of customer's churning propensity, relying on the collection of a wide range of customer's behavioural data over a three-month period. All-important data in the data set comprises of call frequencies (inbound and outgoing), data consumed through 3G and 2G networks, recharge amount, duration of service with the network, the monthly average revenue per user and so on. We endeavour to use machine learning models like, Logistic Regression, Random Forest, and Support Vector Machine (SVM) to address the prediction challenge well. The performance of each model is assessed thoroughly in order to identify which approach is a better option for the problem of potential churners. The aim of this research not only relates to the accurate predictions of groups of algorithms but also to improving decision-making for customer retention programs.

Additional Key Words and Phrases: Customer Churn, Machine Learning, Telecommunications

ACM Reference Format:
Md. Irtiza Hossain, Junaid Ahmed Sifat, and Farhan Ishraq Fagun. 2024. Strategic Analytics for Telecom Churn Prediction using Machine Learning Approaches. 1, 1 (May 2024), 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

One of the key factors that determine whether a company will succeed and compete in the telecommunications industry is the level of customer retention, it is important to note that there are thriving industries where a high degree of customer retention will ensure a strong market position. Now that the world has experienced great utilizations of technological innovations, the data that this segment of customers produces has increased manifold, providing a new avenue for managers to comprehend the behaviour of consumers. Consequently, the power of a data-driven system is limited by competitors' ability to collect and analyze the given information preceding customer churn prediction. The topic of the report under discussion is about the predictive analysis of customer churn within the telecommunication industry but through employing advanced machine learning techniques prediction is done of the customer behavior. We will run different models that include Logistic Regression, Random Forest, and Support Vector Machine (SVM) for predictive modelling. Our models are selected based on their track records of successfully dealing with big data and their great capability to rank classification for different enterprises. The essence of our work is based on sophisticated data analysis of ultra-diverse encompassing attributes of customer activities, characteristics including calling patterns,

Authors' Contact Information: Md. Irtiza Hossain, Brac University, Dhaka, Bangladesh, md.irtiza.hossain@g.bracu.ac.bd; Junaid Ahmed Sifat, Brac University, Dhaka, Bangladesh; Farhan Ishraq Fagun, Brac University, Dhaka, Bangladesh.

Manuscript submitted to ACM

data usage, account retention, recharge habits and revenue data. These factors form crucial part of strategy which deals with intricate relationships between user engagement and satisfaction. By employing this forward-looking analysis, our goal is to ensure that the companies are able to incorporate these suggestions into their customer retention approaches in a manner that is more precise. Consumer abandonment can be prevented by using predictive analytics to identify major dropout factors and implementing targeted interventions to minimize the risks and so increase client loyalty, and hence, reduce the rate of churn in order to maximize profitability. This report is broken to help the reader understand how churn prediction works. The essay starts with an introductory note which is section 1 and then is followed by a review of the literary-related works which is section 2. Section 3 contains research methodology where one will describe data preprocessing and model selection. How to build the experimental setup and a model is discussed in section 4. Section 5 includes the findings and the investigation in it. The last section of the report addresses the sectional issue which includes the Statement of findings and strategic recommendations. This is the layout of the churn prediction's contribution to the firm's competitive advantage and how machine learning can be helpful in strategic decision processes.

2 LITERATURE REVIEW

Machine learning which is able to predict customer churn in the telecom sector being so important, is covered by [2] misclassification calculations and the assumption that ensemble methods are the best models for these. The paper highlights the CART cost-sensitive model as the chosen approach. Scholars who have researched this model have shown that the misclassification costs of this approach are low and it outperforms other classifiers in accuracy. What makes this model distinctive is its demonstration through a real-world dataset, emphasizing its effectiveness in resolving problems stemming from churn prediction that lurk on budget constraints. Similarly, another study focusing on the application of ensemble learning techniques for churn prediction will also be reviewed. The research establishes ensemble learning as the technique that properly tackles the prediction of churn, albeit it acknowledges the involvement of several other strategies to meet peculiar integration needs and predictive performance. The difficulties associated with the choice of method for ensemble classification are considered in detail which urges to handle the problem tactfully before the application of this model in real settings. An analysis section is presented through which certain metrics for each model are provided, and based on this, the gradient boosting model turns out to be the most accurate model as it has an AUC value of 84.57%. The high AUC value indicates the power of the gradient boosting model to pick up churners from the non-churners which means the model is successful in preventing customer churn in our Google Play store. This indicates the key role of selecting the correct model by performance metrics and they give a numeric data about which machine learning algorithm is more effective for churn prediction.

The work of [1] does a detailed study of effective machine learning model implementations for the prediction of customer churn for the case of the telecommunications industry. The exploration of the following models is Random Forest, Support Vector Machine (SVM), Extreme Game Boosting (XGBoost), Ridge Classifier, K-nearest neighbours (KNN) and Deep Neural Networks (DNN). The Random Forest model in its turn could be regarded as the best one, which performance is equal to 90.96% of the test data accuracy. The model is based on the fact that the large amount of information and inconsistencies of certain variables can be well processed by the program. Following the grid search process in linearizing the parameters of the model, accuracy was increased by 91.26%. The model works based on the approach of constructing several decision trees at the training period of time and then generating the class that is the most typical (classification) of individual trees. SVM demonstrated an advantage to smaller datasets and even to

non-linear decision boundaries through the kernel trick, but ensembles were more efficient than single clauses, such as random woods and XGBoost. XGBoost also performed well; apart from this, they were exceptional at effectively dealing with the imbalanced classes, which are common in churn prediction datasets. Ridge and KNN were also tested. Ridge has an L2 regularization parameter that requires data to weigh the opposite sides of the classifier equally. KNN is a non-parametric approach that only looks at proximity between classes. Appraising Deep Neural Networks used for that cause showed promising results which in turn predicted a turning point in the hybridization process towards more sophisticated models capable of revealing non-linear associations between the given large amount of data. The limitation of the above model relates more to the balance between the complexity of the model and overfitting aspects, which are particularly difficult when using deep learning models. It was essential for the grid search and cross-validation techniques to be used in order to improve the models which presented some toughness in computational efficiency and model tuning, however. Through this, the paper not only describes the efficiency of ensemble methods, particularly the Random Forest, but also points out the superiority of Random Forest over multiple algorithms when handling large and complex datasets usually observe in the telecommunication sector.

The work [3] employ three machine learning classifiers to predict customer churn in the telecommunications sector: Logistic Regression, ANN i.e. Artificial Neural Network and Random Forest. The models were assessed for performance based on metrics such as accuracy rate, precision, recall, and error rate. In the end, the Logistic Regression model gave the best results. Accordingly, it attained amazing numbers, including 100% accuracy, 1.00 for both churn and non-churn and, again, 100% recall. The fact that such a model demonstrated high accuracy and correctly identified both churners and non-churners indicates its magnificent purpose for future use. The Random Forest model also yielded excellent results with an accuracy rate of 0.9844, strength measurements of 1.00 for both churn and non-churn, 0.93 recall for churn and 0.99 recall for non-churn 0.99. Random Forest´s classification accuracy rate was slightly less than that of Logistic Regression, but showed high capability, particularly in the case of dealing with several data attributes concurrently and avoidance of over-fitting. Artificial Neural Network (ANN) was a poorer performer, with a precision of 0.8555, whereas the others delivered better accuracy. The model had a precision of 0.70 for churn and 0.80 for non-churn, recall of 0 for churn and 0.92 for non-churn. The returned values for confused customers are lower for precision and recall, meaning that ANN had less chance against Logistic Regression and Random Forest of correctly predicting the rates of churners. The authors mentioned that a limitation of Logistic Regression is its time and space complexity, thus its performance is outstanding but it could be a drawback due to resource requirements for data applications that require a lot of space or real-time handling. Summally, the best outcome was yielded by Logistic Regression, having perfect classification and the highest accuracy, but it can be resource-demanding in some cases. Random Forest is an alternative that adopts a strong approach to classification with a minimal decrease in accuracy but yet a good ability to break down complex data sets. ANN, although the one with the least accuracy among the three algorithms in the current case, can be improved with further adjustment and may require either more complex architecture or architecture depending on the scenarios.

3 RESEARCH METHODOLOGY

Data preparation and processing stages were the introductory methods to our work. Data stored from different origins was brought to the same table, and there were cases of omitting or removing errors in the table. The next step was feature engineering being done through the use of Principal Component Analysis (PCA) to unveil uncorrelated linear combinations of features and to find the right feature for churn prediction. In the process of data modelling, the data set

was separated to produce train and test subsets to avoid the case where the model was being evaluated itself. The models of Logistic Regression, Random Forest and SVM with an RBF kernel were initialized with class imbalances considering and then trained using data which had the smallest subset from PCA. In this evaluation, the model was assessed on parameters of accuracy, ROC AUC score, and recall and then performance indices were calculated to compare the churn to the prediction given by each model. Through the comparative analysis, the model that performed better based on such metrics as precision, recognition and other accuracy characteristics was identified. In order to guarantee reliability, the model was tested with the aid of a separate test set that aimed to show the real results and prevent the problem of overfitting. Ultimately, it was the most powerful version developed that was made available by our model in which PCA components were exploited to pre-process the input variables. This strategy therefore resulted in a careful evaluation of historical data and modeling the derived patterns to predict the high occurrence of churn data over the years. It enabled the application of the model in real-life scenarios.

4 EXPERIMENTAL SETUP AND IMPLEMENTATION

4.1 Dataset Description

Analyzing customer churn in the telecommunications sector utilizes a comprehensive dataset of 69,999 records with 171 attributes. This dataset encapsulates a wide array of features representing various aspects of customer usage and engagement patterns across different telecom circles. Attributes such as local and STD call details, call types (e.g., on-net, off-net, roaming), usage volumes (both voice and data), and various financial metrics like ARPU (Average Revenue Per User) and different recharge amounts provide a holistic view of customer behaviour. The dataset also includes several categorical and continuous variables that track usage patterns over a three-month period (June, July, and August), labelled as KPIs for respective months. Additional variables indicate the type of network (2G, 3G), special services like ISD, roaming, and internet data packs, reflecting both short-term (sachet) and long-term (monthly) usage plans. This detailed dataset will be used to train machine learning models to predict customer churn, employing algorithms capable of handling large-scale data to uncover patterns and predict future churn behaviour effectively. The ultimate goal is to identify key factors that influence customer decisions to leave the network, thus enabling targeted interventions to improve customer retention.

4.2 Data Processing

In data preprocessing of our churn prediction study, we performed an audit on the dataset by noting that some of the columns had a null rate of up to 74% while some were completely absent. The part for maintaining data integrity was the most tiring task during our data cleaning process, where the columns were dropped when the occurrences of null values were over 50%, referring to the 30 columns as being dropped. Besides, more in-depth research showed there existed some more columns with approximately 3000 omitted entries, which were also not included in the statistical study. As shown in the heatmap visualization in 1 before and after column removal, the dataset consists of getting rid of the data sparsity, and this shows improvement of the coherence of the dataset for subsequent analysis. To complete the task, we then threw away the columns with the unique values, because they don't add predictive power-related churn analysis. In the end, the categorical features were applied one-hot by means of, the technique in which they are converted into a binary matrix representation that allows to use of those features in machine learning algorithms. Therefore, the systematic preprocessing systematically processed the data, intervening only in the case of the coarse

features, so subsequently, the model would use only reliable features. We ended up with only 130 columns after all of the data pre-processing, down from 171.

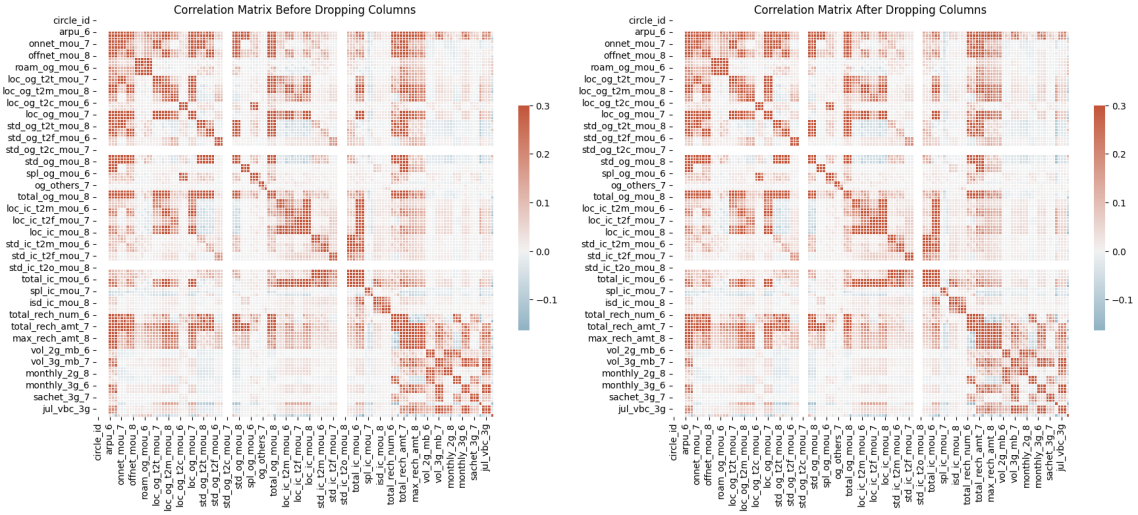

Fig. 1. Correlation Matric Before and After Dropping Columns

4.3 EDA Analysis and Outlier Treatment

Concerning the EDA, there was a low churn rate in the data which is representative of data imbalance can be seen in 2, hence the need to address it to avoid a bias in the model. This pattern corroborated that the adopted approach allowed the organization to scale up to a higher number of observations, eventually increasing customer churn from 1st to 3rd year of the relationship this can be seen **??**. The increasing importance of this trend justifies wider searches for information on the peculiarities of consumer behaviour during these periods. On the other hand, the analysis showed that outliers existed in a number of vital metrics that used call usage in and outgoing, recharge amounts, average revenue per user (ARPU) and data volume for 2G and 3G networks, especially in high quantiles. To eliminate the anomalous observations and standardize the data in order to achieve a more accurate model, we introduced a capping method by developing cutoff limits with 1st percentile and 99th percentile as the terms. This approach entailed that for every column a floor and ceiling value were set below and above which the lowest 1st percentile data point was replaced with the first percentile column and no top 99% point data was replaced with the 99% point column. The treatment was systematically applied across all the columns with outlier values which significantly diminished their impact on the prediction models as well as reduced the misrepresentation of the data main disturbed distribution.

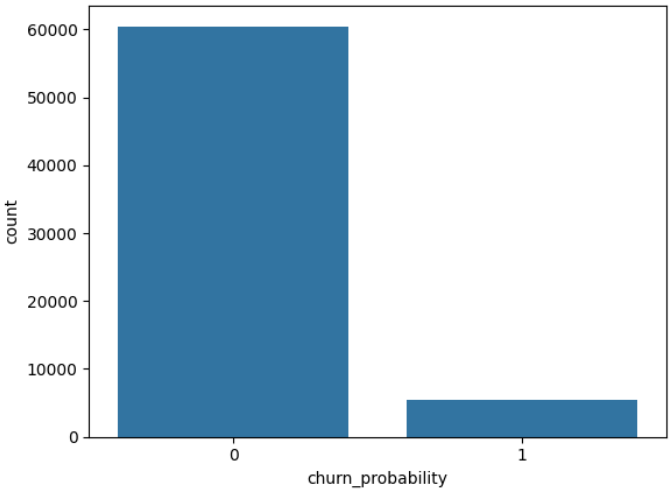

Fig. 2. Churn Probability

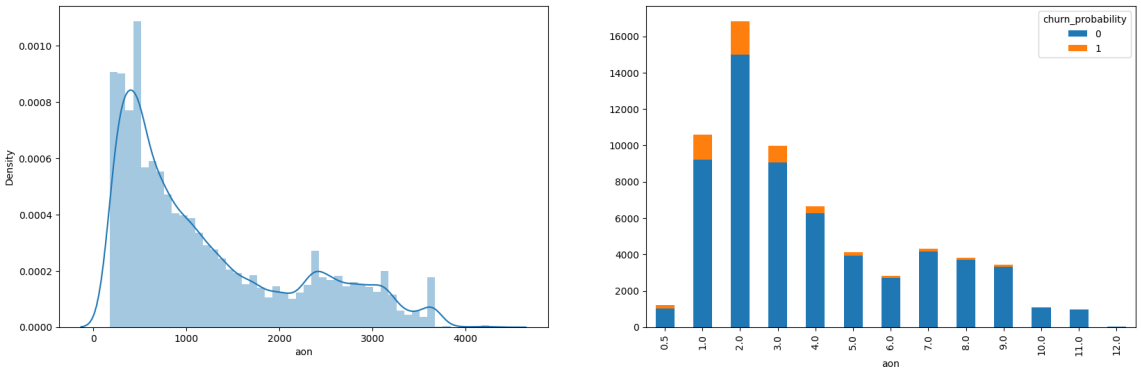

Fig. 3. Year-Wise Churn Probability

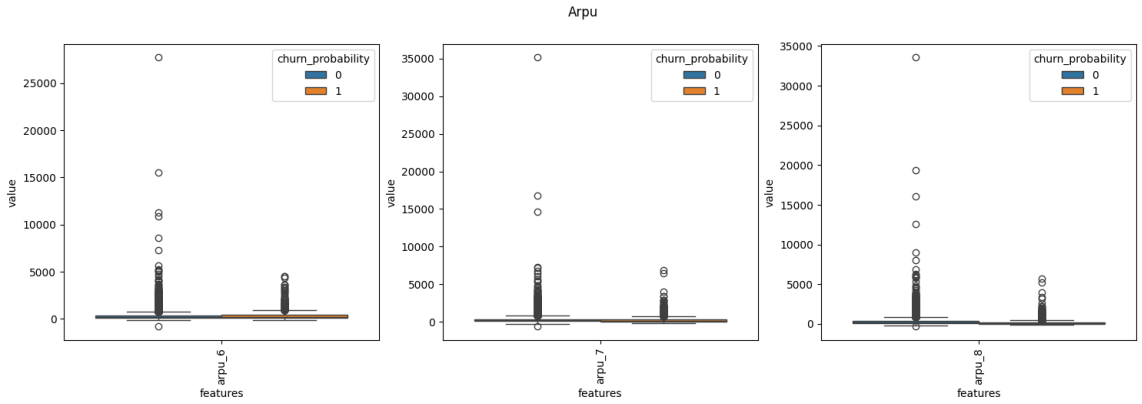

Fig. 4. Distribution of Average Revenue Per User Values by Churn Probability

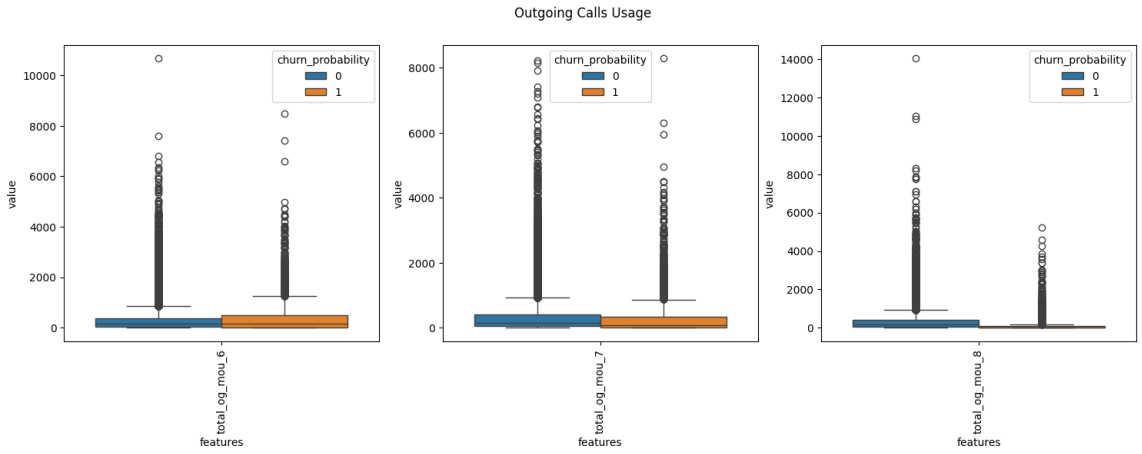

Fig. 5. Distribution of Outgoing Call Values by Churn Probability

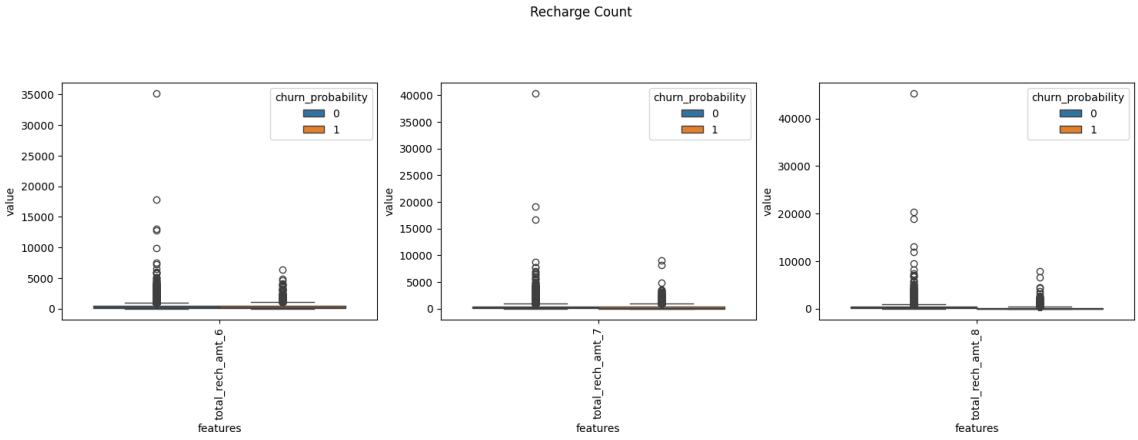

Fig. 6. Distribution of Recharge Count Values by Churn Probability

4.4 Data Imputation and Dimensionality Reduction with PCA for Model Optimization

Before proceeding with the training set split, we started to handle missing values through mean imputation, forward filling, and linear interpolation choosing the strategies which were suitable for the characteristics of each feature. We next narrowed our dataset through feature selection, which led to the selection of only the most significant variables. We adopted PCA, the dimensionality reduction technique which is very powerful, to determine the number of the principal components in the smaller set of uncorrelated variables. Through PCA, there comes down the complexity in the feature space, and multicollinearity is removed, and computational efficiency is improved. This gives importance to the relevant variables containing the most variances in the data. This projection is accomplished by projecting the original data onto a new subspace that is spanned by an idempotent decomposition of the data covariance matrix. This simplifies our modelling, thus increasing the predictive power. After PCA, we were left with only 30 columns where we started with 171 columns.

5 RESULT AND ANALYSIS

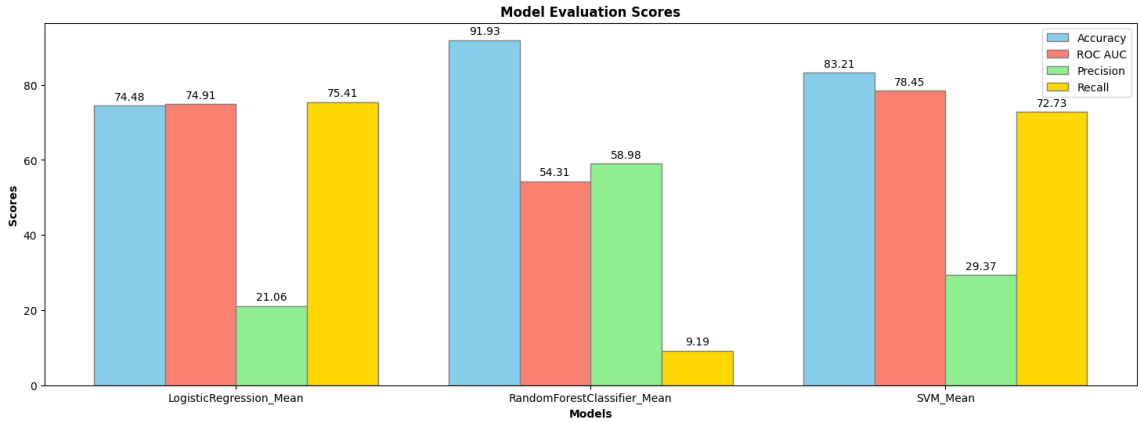

Fig. 7. Comparison of Implemented Models Matrices

From the result of model performance, the SVM model eclipses others by demonstrating an accuracy rate of around 83.21%. This number is notably higher than what is demonstrated by the LR model at roughly 74.48% and the RR at approximately 54.31%. It comes out that the SVM ranked first with 78.45% concurrently, the following will be the logistic regression with 74.91%, while the last will be the random forest with 58.98%. It's Logistic Regression that has the highest score (75.41%) and SVM comes the second (72.73%). However, the Random Forest is the most lagging one with an awful score of only 21.06%. These statistics altogether suggest, if a business assigns more importance to correct identification of churn cases, then SVM can be the choicest classifier but, if the enterprise would assign more weightage to recall rates then Logistic Regression can be the best classifier and Random Forest could need to work as precision and recall balancer. In the churn prediction logistic regression has middle opening scores proving this model's stability as a model. In spite of its scorecard not shining in any particular metric, it did not underperform and, as a result, it constituted a good benchmark for comparison. It follows that, by virtue of its simplicity compared to more complicated models, it could potentially provide a better generalisation capability to unseen data and have a reduced tendency to over-fit The Random Forest classifier proved to be the strongest one in ROC AUC score, thus being a model of choice for the severance prediction model due to its importance of finding the customers who are likely to churn. However, the lower precision and recall suggest the model has issues with classification purposes, particularly where correctly distinguishing true positives from outliers is vital. As for Random Forest, the component of the model consists of multiple trees working together by a voting mechanism through which different decisions lead to the final verdict, which helps prevent overfitting while ensuring robustness as this is very useful to imbalanced datasets. The SVM model that clearly outperforms others by its top accuracy and precision could be stated. It performs excellently in conditions where the possibility of addressing high False positives is significant. This will help the business to position its customer financing strategy accurately. Also, the support vector machine (SVM) operates with a radial basis function (RBF) kernel which enhances the ability of the SVM to handle the non-linear features within the data. This makes SVM captivating and suitable for complex datasets where linear separability is not an option. The reason SVM is good is because SVM is proved to be the best possible option for churn risk determination due to maximum accuracy and precision. Such measures become all the more critical when accurate forecasts are the top priority and expensive maintenance programs

deserve careful selection to be deployed effectively on the target audience. In fact, its capability to efficiently cope with nonlinear patterns of data significantly increases likelihood of being succeed in this scenario. Therefore, in cases where the success of making the right prediction is high and the cost of producing false positives is great, SVM model proves to have the highest predictive power and thus is the most preferred model. This assessment has a limitation mainly because of how inherent assumptions and trade-offs are built into the models that are used. On the one hand, Logistic Regression is powerful enough and provides a good starting point that might not be information-rich or aggravated to fully capture intricate details in the data, leading to overfitting. Since it's a ROC AUC score, Random Forest can be computationally intensive and naturally might not achieve good results for lower precision and recall. Therefore, it is highly possible that some cases would be misclassified. Moreover, this kind of model's Random Forest algorithm could also become complicated, and thus harder is to read and interpret the rule it produces. SVM's high accuracy and precision point towards the model's robustness, but it has a downfall which lies in the specifications of kernel function and hyperparameters and one should not ignore this if one is going to use it, on the other hand, like Random Forest, the model is also difficult to interpret. Besides the fact that SVM may not deliver the same performance with a very large dataset due to high computational demands, it is ill-advised. Furthermore, the performance evaluation relies on the current data set including the feature engineering method; data set changes as well as new feature introductions could reverse model performance drastically

6 CONCLUSION

The interpretative and practical applications of the learnings from the machine learning algorithms in this research have remarkably improved the understanding of customer churn between telecom organizations. Two models, logistic regression as well as random forest, and the SVM have been separately utilized as tools to explore the issue which is customer behaviour and based on the specific insights one can strategically make decisions which can be directly linked to customer retention. The highest accuracy among tested models belongs to the SVM. Therefore, within the scenarios where exactness is more than critical for the decrease of churn and increase of profitability, SVM becomes a valuable model. The research will be, therefore, not without the angular constraints. Another shortcoming is the dependence on the present dataset and the specific feature engineering ways. If the changes in data or the emergence of new attributes were to modify the models' performance level, it could be drastic. For that matter, the SVM algorithm has relatively high accuracy, but its complexity and the burden of computations may hinder scalability, particularly on very large datasets. The use of implicit kernel assumptions and tuneable hyperparameters also makes modelling generalization and comprehensibility a little bit daunting. In addition to that, even though logistic regression and random forest proved to be less precise than SVM, they have underlying perks such as analyzing and understanding the data. Random Forest, with its capabilities of combating overfitting and dealing with imbalanced data sets, does suffer from decreased precision and recall, a phenomenon that may result in instances of misclassification in churn cases. To find a way over these restrictions and enrich this predictional ability, future studies can be developed by the integration of more dynamic and diverse datasets, deploying different feature selection techniques, and probably cost-sensitive learning, to be derived at better model outcomes. Moreover, making the models more explainable and computationally efficient to sync up with real-time situations wherein quick and correct decision expectations must be kept would be vital. The conducted research enhances a comprehensible idea of the influence of advanced analytics on the process of transition from tactical strategies to strategic ones within a competitive field. Through the constant enhancement of such systems and reducing their drawbacks, an enterprise finds a way to generously improve the accuracy of their

models and the effectiveness of their operations, which brings only positive consequences including business thriving and customer satisfaction.

REFERENCES

[1] Pushkar Bhuse, Aayushi Gandhi, Parth Meswani, Riya Muni, and Neha Katre. 2020. Machine Learning Based Telecom-Customer Churn Prediction. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. 1297–1301. https://doi.org/10.1109/ICISS49785.2020.9315951

[2] Abhishek Gaur and Ratnesh Dubey. 2018. Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques. In *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*. 1–5. https://doi.org/10.1109/ICACAT.2018.8933783

[3] Nurul Izzati Mohammad, Saiful Adli Ismail, Mohd Nazri Kama, Othman Mohd Yusop, and Azri Azmi. 2020. Customer Churn Prediction In Telecommunication Industry Using Machine Learning Classifiers. In *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing* (Vancouver, BC, Canada) *(ICVISP 2019)*. Association for Computing Machinery, New York, NY, USA, Article 34, 7 pages. https://doi.org/10.1145/3387168.3387219