Machine Learning Based Approaches to Predict Customer Churn for an Insurance Company

Yunxuan He
University of Virginia
School of Data Science
Charlottesville, VA, U.S.
yh7bb@virginia.edu

Ying Xiong
University of Virginia
School of Data Science
Charlottesville, VA, U.S.
yx4pt@virginia.edu

Yiting Tsai
University of Virginia
School of Data Science
Charlottesville, VA, U.S.
yt9mh@virginia.edu

***Abstract*—Customer churn prediction plays an important role in business success for insurance companies like Markel Corporation. Each year Markel loses premium because some of their customers choose not to renew their policies. Based on the fact that the cost of attracting new customers is much greater than that of retaining existing customers, it is important for Markel to take early action to engage their customers before a policy expires. The goal in this work is to apply various machine learning methods and obtain an optimal model to predict customer churn rate. The dataset includes customer demographics features, customer behavior features, and macro environmental features. Exploratory analysis is conducted on critical features including policy length and types of coverage to draw insight about the impact of these features on the target variable – customers renew or do not renew their policies. With a large dataset, one of the main challenges is conducting feature dimension reduction and extracting important features to be used with a set of potential ML models. It turns out that the ML model with the best performance on the Area Under the Curve (AUC) metric is Extremely Randomized Trees Classifier and Gradient Boosting Model. Some suggestions on additional features to be incorporated are provided in the final comments. These features will improve predictive performance for the ML model of customer churn for Markel Corporation.**

***Index Terms*—Customer Churn Prediction, Machine Learning, Neural Network, Gradient Boosting Machine, Random Forest, Support Vector Machine, Automated Machine Learning**

I. Introduction

Customer churn refers to the rate of customer attrition in a company. It is one of the most important metrics for a growing business to evaluate. Usually it is much easier and less costly to keep an existing customer than it is to gain a new customer. A small decrease in customer churn rate will increase the company's profit. Therefore, a predictive model to foresee those who are going to leave the company is important for common service industries such as insurance. This paper focuses on how different machine learning techniques work on a dataset from Markel Corporation, a company specializing in insurance, reinsurance and investment operations around the world. The goal is to build a model with predictive power and to provide insight on customer churn behavior for the company.

Due to significant variance between retention rates of groups with different coverages, the dataset is divided into two parts and sub-models are built to cater to those different groups. The machine learning models applied include logistic regression with penalty, random forest, extremely randomized trees classifier (Extra Trees Classifier), support vector machine (SVM), neural network (NN) and gradient boosting method (GBM). AUC scores of 0.68 and 0.61 are achieved for each of the sub-models built with Extra Tree Classifier and Gradient Boosting.

II. Related Work

There is a plethora of research that applies machine learning to customer churn problems.

Burez and Van den Poel [1] focus on class imbalance in a churn problem. They declare that AUC is an overall better metric compared to accuracy for class imbalance problems. They also verify advanced under-sampling techniques that lead to improvement of model predictive power for logistic regression and random forest models.

Veronikha, Adiwijaya and Baizal [2] use combination sampling (over-sample to increase minority class using Synthetic Minority Oversampling Technique, or SMOTE in short and under-sample the majority class) and weighted random forest (WRF) to perform the best prediction on an imbalanced dataset. SMOTE augments minority class by creating additional data close to existing examples. There is a trade-off using these sampling methods: Under-sampling reduces data dimension and boosts the training process, but loses possible information from the majority class. On the other hand, over-sampling increases learning time and may cause overfitting. A weighted random forest can also address imbalanced data problems by assigning penalized misclassificated data with higher class weights. Since Markel data is imbalanced, this cost-sensitive learning approach can fulfill the needs to strengthen the learning ability to the minority class.

Vafeiadis, Diamantaras, Sarigiannidis and Chatzisavvas [3] conduct comparison among popular classification methods for a customer churn prediction in the telecommunications industry. An advanced boosting version of each ML model has also been developed. A poly-SVM with AdaBoosting achieves the best performance for the telecom data.

Zhang, Li, Tan and Mo [4] develop a state-of-art approach for customer churn combining shallow and deep components (DSM) by using a weighted sum of their output log odds as

the prediction, which can take advantage of models with good memorization and models with good generalization.

Olson R.S., Moore J.H. [5] introduces an automated machine learning (AutoML) tool, Tree-based Pipeline Optimization Tool (TPOT), that trains machine learning models without human supervision. AutoML is a technique that designs and optimizes data pipelines automatically. TPOT explores thousands of possible pipelines and utilizes genetic programming(GP) to select the best hyperparameters and optimal machine learning method. This approach will reduce efforts and time to construct high accuracy data pipelines.

III. Data and Context Setup

A. Data Dictionary

Machine learning models are implemented on a real-life dataset received from Markel Corporation. There are 25,275 observations covering policies from 2011 to 2019, which contain 253 features about customer demographic, customer behavior, and macro environmental information. Among all features, 40 are company internal features and the rest are governmental features. Grouping by data type, there are 241 numerical, 8 categorical and 4 date-type features. The target variable is binary with 1 denoted as 'Renewed' and 0 as 'Not Renewed'. In the dataset, 72.6% of the entries are renewal instances and 27.4% are non-renewals.

B. Exploratory Data Analysis

Before fitting any machine learning models to the data set, thoroughly inspecting the data can strengthen the understanding of the dataset and indicate possible data issues. Exploratory Data Analysis(EDA) is applied for data summarizing, visualizing and getting a quantitative sense about the data.

- Missing values: Seven governmental features with more than 60% missing values were removed. An imputation based on regional information was conducted for other missing governmental data.
- Correlation: There are more than 100 features that are highly correlated with one or more other features. A heat map of 253 features is not clear and does not demonstrates the correlation clearly. Thus, features are grouped into several categories and correlation heat maps are checked respectively. For instance, high correlations exist in features related to climate where the insured is located and features related to the consumption situation of that area.
- Before fitting models, exploratory analysis was conducted to check how retention rates differ for different outcomes of some features.
  - As shown in Figure 1, customer retention rate for policies with "Product Line2 Code 4" (denoted as PLC4) have higher retention rate than "Product Line2 Code 1/2/3" (denoted as PLC123), where "Product Line2 Code" refers to different types of coverage (used for dataset splitting);

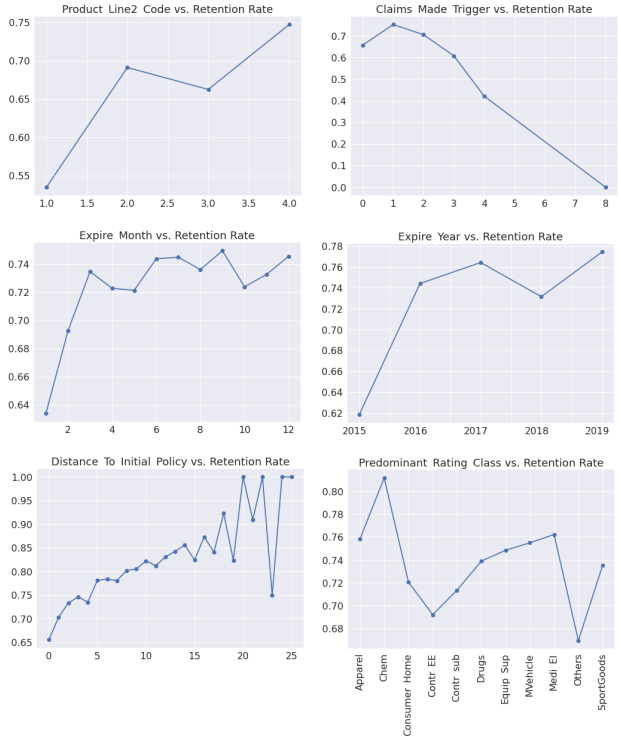

Fig. 1: How retention rates differ for different features

  - Customers who have made fewer claims and stayed longer with Markel have higher probability to renew their policies. Moreover, customers from the chemistry industry are more likely to renew their policies than those from other industries. Last but not least, retention rates have an overall increasing trend along time, meanwhile policies with January as the expiration month show the lowest retention rate among 12 months.

C. Data Pre-Processing

Missing governmental data is first imputed by corresponding zip code and state values, and missing company internal data is assigned by median. In order to avoid overfitting and reduce training time, feature selection strategies are introduced. Highly correlated features are deleted with a threshold of 90% and categorical features are converted into dummies. Top 10 values are chosen as representative if they contain more than 10 different classes. To select features that contribute most in prediction, introduction of white noise can help extract the top 21 features, sorted by XGBoosting feature importance 50 times out of 100 on bootstrap sets.

The last process is to limit outliers by Winsorizing, a method that changes outlier values to boundary values. Additionally, to ensure input is of the same scale and to reduce processing time, standardization prior to fitting SVM and NN is needed.

D. Model Evaluation: Area Under the Curve (AUC)

AUC is a measurement for classification performance with different thresholds. It indicates how well a model distin-

guishes randomly selected positive and negative values and measures the trade-off between true positive rate and false positive rate. As AUC score gives a broad view of model performance and can be explained intuitively to business users, AUC score will be used to evaluate model performance. We implement cross validation to get a robust AUC score.

E. Context Setup: Goal and Final Deliverables

The goal for this project is to build a predictive machine learning model to obtain the probability that a customer churns. This model will be applied 90 days before a policy expires in order to check customers who are likely to leave Markel and take early action to bring back those customers. Results from different machine learning models are listed and compared to find the one that gives the highest AUC score. Additional features that can improve the predictive power of the model are listed at the end of the paper. A machine learning pipeline is built as follows:

- Step1. Data Cleaning and Train Test Split
- Step2. Feature selection:
  - Add a white noise as a new feature
  - Run a GBM model on 100 bootstrap training sets
  - Select features that are more important than the white noise 50 out of 100 times
  - 21 important features have been selected
- Step3. Model building on whole training set and check model performance on testing set (AUC)
- Step4. Split dataset for sub-models:
  - 90% of predicted value for observations with PLC4 equals to 1
  - Split data into two groups: PLC4 and PLC123
- Step5. Final model building:
  - Build two sub-models on the above two sets.
  - Fine-tuned hyperparameters to get optimal models.

IV. Machine Learning Methods Applied

Different machine learning models are implemented on the splited dataset to achieve the final model.

A. Logistics Regression

It is common to apply logistic regression on a binary classification problem such as customer churn. In order to avoid overfitting, regularizations are incorporated into the algorithm to penalize high value parameters and ensure that the model can have a better performance on unseen data.

B. Random Forest

Tree-based methods work well for datasets with outliers and with high dimension. Random Forest is a collection of many decision trees that aggregate individual predictions to vote for final prediction. Each tree involves bootstrapping training samples and selects a subset of features randomly. This technique enables each decision tree to be more independent and hence improve the performance. As each individual tree only with a subset of samples, this results in faster computational time. Given Markel data, it turns out that random forest model will boost training speed as well as perform high prediction.

C. Extremely Randomized Tree (Extra Trees) Classifier

Extra trees classifier is a similar technique as random forest; the difference is that it builds trees without bootstrapping samples by default and it randomly splits the node instead of finding the best split. While the extra trees classifier tends to be biased to the majority class, class weights can be assigned to each class in the splitting process to penalize misclassifications and place emphasis on the training class with higher class weight.

D. Support Vector Machine (SVM)

SVM is a traditional technique that aims to solve binary classification problems. It transforms data by using kernel tricks, and finds an optimal boundary to separate data. SVM can always find a global optimal solution because the optimization function is convex, while some tree-based algorithm can only return local optimal solutions. A drawback of SVM is that it requires setting the correct hyperparameters and kernel to achieve the best result. It is also computationally intensive for large data sets. Given the size of Markel data, it is inefficient to apply svm to get the best performing model.

E. Automated Machine Learning (AutoML)

It is very time consuming to grid search optimal machine learning models and hyperparameters. AutoML is a technique to automatically design and optimize data pipelines from data pre-processing, model fitting, parameters tuning to model selection. AutoML is a highly efficient tool that explores thousands of process combinations and requires less prior knowledge from the user, which in some cases performs better than basic data analysis. TPOT is a Python automated machine learning tool that is built on top of scikit-learn, which utilizes genetic programming to optimize data pipelines. It tries thousands of possible pipelines and returns the best one. As AutoML learns better with more generations, it depends heavily on sufficient training time and computational power.

F. Neural Network (NN)

Deep learning models are now widely used for many artificial intelligence problems including image recognition, speech recognition and natural language processing (NLP). On one hand, those models can extract inexplicit features that are hard to detect by other shallow models like logistic regression. On the other hand, they can recreate feature combinations by iteratively adjusting feature weights during backpropagation (Zhang, Li, Tan and Mo [4]). A multilayer perceptron (MLP) neural network model, a supervised learning algorithm, is implemented on this classification problem.

G. Gradient Boosting Model (GBM)

As a tree-based model like Random Forest, Gradient Boosting Models (GBM) are another sophisticated and powerful method which ensembles a series of weak learners into strong learners. However, unlike Random Forest which grows trees in parallel on bootstrap sample sets, the main idea of GBM is to grow the weak learner in a gradual and sequential manner.

Each time when a new one is added to the sequence, it helps to optimize the loss function. The term "Gradient" demonstrates that the way of introducing new learners is to descend the gradient of the differentiable loss function, which makes the algorithm both accurate and fast.

Since GBM works in a way of growing a new tree sequentially and each time it focuses more on the points which are incorrectly predicted, it is a suitable way to deal with Markel imbalanced customer churn data. Besides, a sophisticated model like GBM can gain more interpretability with a package called SHAP.

H. Other Methods Applied

There are some other techniques that have been implemented to improve the model performance; however, they do not work well given Markel's data.

Principal Component Analysis (PCA) is a popular dimensionality reduction technique which has been commonly applied. It does not make much difference to the model performance given the data from Markel. One possible reason might be that it assumes there are linear relationships among features but there is no relationship among features in this dataset.

Survival analysis is sometimes used for customer churn analysis with a focus on investigating the attrition rate risk over time. Continuous time variables are needed in a survival model. Given the dataset from Markel, time variables are discrete and binary classification is required instead of risk change over time. Therefore, survival analysis does not help in this case.

V. Results

A. Model Performance Comparison

TABLE I: Comparison of AUC score of different methods for PLC123

Model	Parameter	AUC Score
Logistic Regression	class weight = {0: 3, 1: 1} penalty ='l2' C = 0.1	0.53
Random Forest	class weight={0: 1, 1: 2} max features = 15 max depth = 15	0.60
Extra Trees Classifier	class weight= {0: 1, 1: 4} max features = 5 max depth=18	0.60
Support Vector Machine	class weight = {0: 1, 1: 1} kernel= rbf C=0.5	0.57
Gradient Boosting	**learning rate=0.05** **max depth = 18** **max features=18**	**0.61**
Neural Networks	learning rate=0.001 batch size = 64 optimizer = Adam	0.56

TABLE II: Comparison of AUC score of different methods for PLC4

Model	Parameter	AUC Score (%)
Logistic Regression	class weight = {0: 6, 1: 1} penalty ='l2' C = 0.01	0.58
Random Forest	max depth = 15 max features = 16 max depth = 15	0.66
Extra Trees Classifier	**class weight= {0: 1, 1: 5}** **max features = 12** **max depth = 22**	**0.68**
Support Vector Machine	class weight = balanced kernel= rbf C = 0.5	0.63
Gradient Boosting	learning rate=0.01 max depth = 9 max features=18	0.65
Polynomial Transformation + Gradient Boosting	degree = 2 learning rate = 0.01 max depth=9 max features=15	0.66
Neural Networks	learning rate=0.001 batch size = 128 optimizer = Adam	0.64

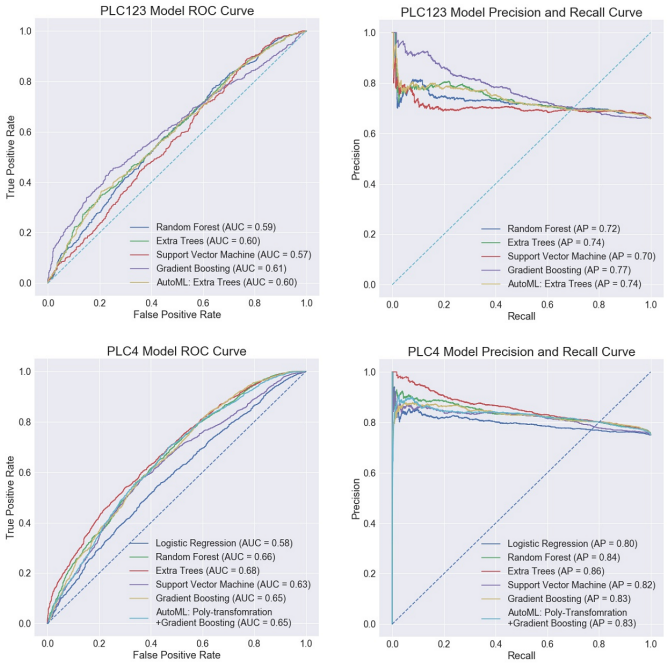

Fig. 2: Model performance comparison

The models are compared using AUC score and confusion matrices are then created to compare results for different thresholds in optimal models. As can be seen in Table 2, the best performing model is Extra trees classifier for PLC4 with 0.68 AUC score and GBM in PLC123 with 0.61 AUC score, as shown in Table 1.

After getting feedback from clients, it was ascertained that

more emphasis should be put on false-negative classification (Table III and Table IV), the percentage of predictions erroneously predicting not renew. The probability that a particular customer is retained is given by the model. The threshold by default is 0.5. A probability of lower than 0.5 is predicted to leave; those customers will be considered target customers and actions will be taken to engage them. In this case, a threshold of 0.5 is sufficient to meet the business context of Markel. By setting the threshold as 0.5, as many false-negative cases will be eliminated. Markel can target those customers who are most likely to churn to save spending and human efforts accordingly.

TABLE III: PLC123 Confusion Matrix

		Prediction	
		Renew	Not renew
True	Renew	1101	258
	Not Renew	530	172

TABLE IV: PLC4 Confusion Matrix

		Prediction	
		Renew	Not Renew
True	Renew	3661	88
	Not Renew	1086	171

B. Feature Importance Analysis

By combining a fitted classification model with the SHAP package, each feature is assigned a value which measures the marginal contribution of the feature to the target variable. Based on the SHAP values gained in the first GBM model over the whole training set, a feature importance summary plots can be generated.

Figure 3 shows a list of most important features in descending order. Figure 4 shows how these features affect the target variable - whether a customer renews or not. The horizontal position indicates whether the effect of that value is associated with a higher or lower predicted retention rate. The red or blue color shows whether the value of that feature is high or low for each observation.

"Distance To Initial Policy" is the most important one among the 21 selected features. According to Figure 4, it has a positive relationship with retention rate. The longer the customers have been with Markel, more likely they are going to renew the policies. On the other hand, "Policy Length", "Paid ALAE Amt" and "Paid Loss Amt" show an opposite trend. Policies with a longer policy length have a larger negative SHAP value which means that the customer is less likely to renew with Markel. And the other two features which represent the amount of losses allocated to the policy's claim, no matter whether they have been paid or not, would have a negative effect on policy retention.

Feature dependence analysis have been conducted respectively on the data sets with PLC123 and PLC4 in Figure 5.

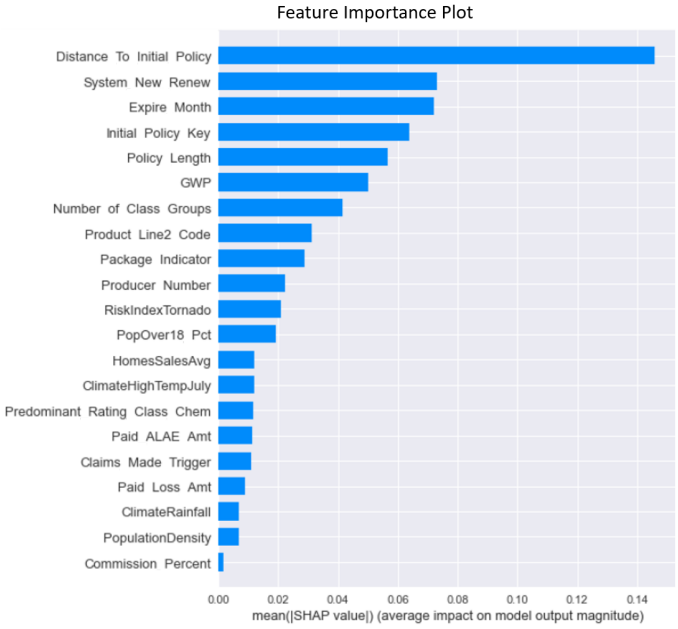

Fig. 3: Feature importance plot

Fig. 4: Model performance comparison

The first subplot on the left shows that for the first group, PLC123, whose expiration months are May, June and July, are less likely to renew their policies. While for the PLC4 group, customers who are going to expire on the first two months tend to choose not renew with Markel. The dependence plot for "System New Renew" shows that an renewal customer has a higher probability to renew their policy than a new customer in group PLC4. While some old customers in group PLC123 may have a different idea. The last point is about "Distance To Initial Policy": In group PLC4 there is a clear upward trend representing that the longer the policies stayed with Markel, the more likely they will get renewed. While in group PLC3

this kind of pattern can not be detected, which coincides with the effect the feature "System New Renew" brings.

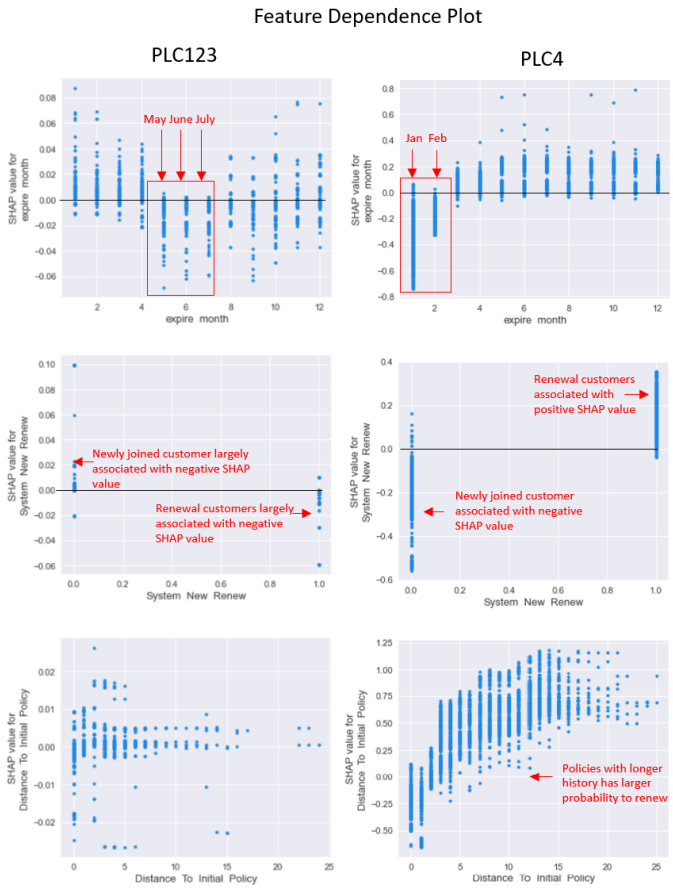

Fig. 5: Feature dependence comparison for each group

VI. Conclusion and Final Comments

This paper explores multiple approaches to predict churn classification and compared performance based on the AUC score. The study shows that Extra Tree Classifier and Gradient Boost are the optimal models for PLC4 and PLC123 respectively.

As all tuned models return similar results, optimal models are achieved given Markel's dataset. In order to increase prediction performance, other critical features need to be added. Policy Expiration Year, which refers to the year the policy expires, is a statistically important factor. By incorporating this feature in the final model, the AUC score would likely increase significantly. However, it is impractical for Markel to include policy year in their production and therefore this feature is excluded from the model. Additionally, some other features like market conditions and company service rating, which vary over year, are essential features to be included and will make the model more effective in predicting customer churn.

Customer churn is highly associated with subjective causes like individual business financial performance and customer-related factors, which are not incorporated in this dataset. Markel can consider include these other features for future use if possible.

There are some areas that warrant further study. First, January to March in PLC123 and May to July in PLC4 have the most contracts that are not renewed, as is shown in Figure 5. In particular, there may be reasons why clients choose not to renew during that period. Therefore, a better model can be developed by including these crucial variables.

As for the benefits of this research for Markel, actions can be taken based on the trend of important features. Our standardized data pipeline can save human labor and time as well as give better prediction. Through targeting potential churn customers, Markel can take actions 90 days before the contract expires to retain customers.

Acknowledgment

The research for this paper was sponsored by Markel Corporation and University of Virginia, School of Data Science. We would like to express our deep gratitude to Louis Mungin for the continuous support throughout the process. We also would like to thank our supervisor Professor Learmonth for his patient guidance and constructive recommendations on this project.

References

[1] Burez, J. and Van den Poel, D., 2008. Handling Class Imbalance In Customer Churn Prediction.

[2] V. Effendy, Adiwijaya and Z. K. A. Baizal, "Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest," 2014 2nd International Conference on Information and Communication Technology (ICoICT), Bandung, 2014, p. 325-330.

[3] Vafeiadis, T., Diamantaras, K., Sarigiannidis, G. and Chatzisavvas, K., 2015. A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, pp.1-9.

[4] Zhang, R., Li, W., Tan, W. and Mo, T., 2017. Deep And Shallow Model For Insurance Churn Prediction Service - IEEE Conference Publication.

[5] Olson R.S., Moore J.H. (2019) TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In: Hutter F., Kotthoff L., Vanschoren J. (eds) Automated Machine Learning. The Springer Series on Challenges in Machine Learning. Springer, Cham.

[6] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," Eighth International Conference on Digital Information Management (ICDIM 2013), Islamabad, 2013, pp. 131-136.

[7] K. Dahiya and S. Bhatia, "Customer churn analysis in telecom industry," 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, 2015, pp. 1-6.