A Study of Ensemble Machine Learning to Improve Telecommunication Customer Churn Prediction

Dissertation submitted in part fulfilment of the requirements
for the degree of
MSc Data Analytics
At Dublin Business School

Name: Nikhil Vilas Raut

Roll Number: 10521287

ACKNOWLEDGEMENTS

I cannot express enough thanks to my professor and research supervisor, Ms. Terri Hoare for her continued support, patient guidance from the start, valuable and constructive suggestions provided throughout the planning and development of my research. I offer my sincere appreciation for the learning opportunities, and her willingness to give her time generously has been very much appreciated.

I would also like to extend my thanks to H2O.ai for giving me the work on the Driverless AI cluster and the RapidMiner for providing machine learning and predictive analytical platform.

Finally, to my family for supporting and encouraging me throughout my entire project.

Declaration I, Nikhil Vilas Raut, declare that this research is my original work and that it has never been presented to any institution or university for the award of Degree or Diploma. In addition, I have referenced correctly all literature and sources used in this work and this this work is fully compliant with the Dublin Business School's academic honesty policy.

Signed: Nikhil Vilas Raut

Date: 25th Aug 2020

Abstract

The telecom industry has an ever-increasing number of service providers. With the competition between service providers to attract more customers by making attractive offers, there is an incentive for customers to keep switching service providers. Service providers can stay ahead of the game by making proactive attractive offers and retain customers. Machine learning is applied to predict those customers most likely to churn. This research uses state of the art H2O stacked ensemble methods to improve the performance accuracy of the base models.

The tools that are used for the research are RapidMiner, which is an analytical automation tool, and python coding implementing the H2O framework specifically designed for artificial intelligence. Traditional models are implemented in RapidMiner, namely the Generalized Linear Model, Gradient boosted trees, Deep Learning, and logistic regression. The models implemented in Python using the H2O framework are Stacked ensemble, deep learning, and distributed random forest trees. Base results are compared against ensemble methods and evaluated on various performance metrics.

Contents

Chapter 1: Introduction

1.1 Introduction

Churn Prediction Process is an integrated data analysis and classification process that enables a business to analyze the churn rate among the customers gained over a definite period. With the "enterprises in the telecom sector" looking forward to the retention of "loyal customers" for their business growth, the enterprises are also keen on acquiring new customers" at the same time. In order to improve the possibilities of gaining more customers and retaining them to earn a greater profit, the organizations establish strategies to predict the churn rate and identify their drawbacks to improve the situation by using different data analysis mechanisms and classifiers. This provides a greater opportunity for administering "customer relationship management." The "ensemble learning methods" will successfully combine the data mining methods and also "improve the accuracy and robustness of the process."

The study is based on the process of understanding the "ensemble learning process corresponding to customer Churn Prediction." Thus, the introduction chapter shall discuss the critical aspects of churn prediction and "ensemble learning methods." The background of the study provides an emphasis on the pre-acquired knowledge that is to be implemented in conducting the overall research. Furthermore, the "aims, objectives, and research questions" are outlined in this chapter to guide the research towards the specific goals of understanding the utility of the "churn prediction in the telecommunication sector." Additionally, the problems associated with the research area are strategically presented in the introduction to demonstrate the issues that should be managed for the appropriate application of the process. The chapter also defines the purpose of the study and the significance it holds with respect to the fulfilment of the core objectives of the research.

1.2 Background of Study

It is quite challenging for an organization to retain the customers and also ensure growth by acquiring new customers. However, this challenge is mostly perceived by the telecommunication industry, where the customers are prone to frequent switching among the different providers in the telecom sector, owing to their better performance of one provider over another. This fuels the need to establish a definite mechanism or a model that can identify the limitations of the organization due to which the organization fails to retain its customers and improve the sales revenue and growth profits as presented in (Pretam Jayaswal et al.,

2016). Therefore, it is essential to identify the number of customers who switched to a different telecommunication organization and also predict the number of customers that can move to a different organization depending on the existing features and facilities provided by the company to its customers. This phenomenon of customers moving to a different organization is designated as customer churn, and the rate at which this occurs is termed as the "customer churn rate". ***[Referred to Appendix 1]***

The organizations need to identify and predict the churn rate in order to develop their services and to administer appropriate facilities in the customer service. There are several methods of churn prediction that can be effectively implemented with machine learning techniques and other data analysis methods. However, "ensemble learning method" is explored to understand its potential in effectively handling large data sets and predicting the churn rates using different algorithms and techniques to provide accurate and precise results from the evaluation of customer data (Lessmann, Stefan & Coussement, 2018) (K. Mishra and R. Rani, 2017). The "Churn Prediction process" allows the telecom providers to retain loyal customers through the development of strategies that ensure enhanced performance in the market and provide customer satisfaction in various ways.

The chart prediction methods can be implemented with the use of different classifiers such as "Random Forest classifier, Logistic Regression, Bagging classifier, Decision Tree, Support Vector Machine (SVM), Naïve Bayes and Booting classifier". (Mumin Yildiz et al., 2020) discusses several methods that can be implemented. The online article provides a detailed discussion about the hybrid models using "Artificial Neural Network" and "Self-organising maps", "Bayesian Belief Network Models", "Ant-Miner+", "Rotation Forest" and "AdaBoost". These algorithms correspond to various "ensemble classifiers" and "rule extraction methods." With a distinctive evaluation of the discussed methods, the authors have analyzed a set of data. "Ensemble Learning method" is a highly effective "machine learning mechanism" that incorporates different models into one system model to produce the best results of data mining and analysis. The concept that is implemented behind the "ensemble learning methods" is that several weak models can provide extensively powerful inaccurate results when these models are combined strategically. Since the choice of a particular model in analyzing a set of data with the help of a machine learning mechanism depends on several factors such as the size of the data set, dimensionality, and distribution. The data set may have "high bias and high variance." The Ensemble method aims to develop a complex model that would help in "reducing the bias and variance of weak learners" to achieve accuracy and performance. Hence this model also requires the appropriate choice of the individual base models so that they are

"homogeneous weak learners trained in different ways" combined to form a "heterogeneous ensemble model". This also indicates the need for the implementation of several meta algorithms such as bagging, boosting, and stacking.

This explains the overall concept that is to be integrated into this research to develop a suitable Ensemble method that assists the "prediction of Churn rates among the customers of the telecommunication sector."

1.3 Research Aim

Precisely forecast the future churn to assist administrative decisions for the company's telecom customer.

1.4 Research Objective

Comparing standard machine learning algorithms with Ensemble machine learning techniques in order to find the most effective approach for forecasting and what variables affect them.

1.5 Research Question

The research seeks to address the following research questions:

1. How efficiently can the “ensemble learning methods predict the customer churn rate over traditional machine learning techniques”?
2. How does the “feature selection techniques” provide improved “accuracy of models”?

Hypothesis: Will the Stacked Ensemble Model outperform the traditional Machine learning algorithms.

Scope: The aim of this research is to help develop a Churn Prediction tool, which includes a broad range of statistical and machine learning algorithms to predict consumer churn accurately.

Limitation: Telecom companies won't share their private data since it is confidential, and data leak could be a serious offense. An accurate and huge amount of data will be helpful in training the Machining Learning algorithm more precisely.

1.6 Research Motivation

The research motivation was taken from the paper published by the authors Pretam Jayaswal, Bakshi Rohit Prasad, Divya Toma, and Sonali Agarwal. The paper was published on topic of “An Ensemble Approach for Efficient Churn Prediction in Telecom Industry”. The researchers

aimed to use customer usage and similar information from a telecom service provider to analyze churn in the telecom industry. The machine learning model they used in their research are decision tree and its ensemble variants random forest trees and gradient boosted decision trees. These ensemble-based algorithms Gradient boosted tree outperformed the other classifier models. The Gradient boosted tree performed even better than Random forest in terms of accuracy and sensitivity. The gradient boosted trees had an accuracy score of 96.78% (Page 227), random forest trees had an accuracy score of 95.32% (Page 228) and the lowest performance was by Decision trees having the accuracy score of 89.99%.

1.7 Research Rationale

The telecommunication sector faces a lot of trouble due to the frequent switching of customers between different providers. This results in the loss of customers and thereby affects the profits and overall revenue of the organizations. This phenomenon of service migration takes place when a company fails to ensure the satisfaction of the customers. Since the telecommunication market is inhabited with numerous telecom providers who practice active competition in acquiring customers and improving their performance through a cheap pricing and several other marketing strategies, service migration poses a major threat to the companies. Therefore, customer retention is a significant problem within the telecommunication industry that requires effective solutions and business strategies in place to retain loyal customers and also to acquire new customers, as mentioned in (Debjyoti Das Adhikary et al., 2019). This problem affects not only the financial growth but also downturns the "reputation of the enterprise". The lack of an efficient predictive model that can provide significant results corresponding to the Churn rate of customers poses a considerable amount of problems to the organizations. This also prevents enterprises from determining and identifying their limitations that should be handled effectively. The company also suffers from severe mismanagement and lack of administrative qualities regarding the "customer relations management," thereby losing the market to more competitive enterprises that retain loyal clients.

There are several challenges to the data mining process and the machine learning models which are outlined in the journal as the "large volume of data, variety of the data, unbalanced dataset, missing values and an extensive number of data features" (U. Ahmed, A. Khan et al., 2019). Overcoming these barriers in the dataset are also essential for appropriate "data analysis and machine learning."

Also, the identification of the appropriate ensemble method based on the classification of individual base models plays a vital role in determining the net customer churn rates of the company with extensive accuracy and precision. There are several methods of combining the

base models and identifying the particular algorithms that can be used to develop a suitable "ensemble learning model" which can provide powerful results from the "training and testing of data sets" (A Hammoudeh et al., 2019). Furthermore, the improper combination of weak learners can provide inappropriate results from the data mining and analysis of the Ensemble method.

This covers the different problem areas corresponding to which the research is implemented and seeks to address these issues by developing an innovative "ensemble learning model" and deploying it for predicting the customer churn of telecommunication providers.

1.8 Research Significance

The evaluation of the study will help determine the process in which the "churn prediction" is being conducted by implementing the "ensemble learning method." Moreover, the study shall emphasize on the aspects of ensemble models to enable the combination of different models with the use of suitable algorithms and methods and ensure an accurate data analysis.

Most essentially, the research provides extensive ideas about the models and techniques that have been incorporated by the authors and researchers regarding the appropriate modeling of the learning mechanism that ensures the precise and accurate data prediction (R. Andrews, 2019). The research shall also consider effective solutions that cover the critical areas of feature selection to provide the correct solution to the "data mining and the machine learning process."

The purpose of the study is also to provide practical, evidence-based analysis through the implementation of a large customer data set and apply the mechanism of "ensemble learning" to find the accurate and precise predictive data from the data set regarding the customer churn rate (W. Bi, M. Cai, et al., 2016).

The study also helps to understand the specific algorithms that are associated with the implementation of the learning methods such as the ANN, CNN, RF, and SVM. With the elaboration of the strategic approaches of conducting a churn prediction with respect to "ensemble learning methods". This indicates the overall purpose of the experimental research study and the overall significance the study has in the context of analyzing the customer data for performing the "churn prediction based on ensemble learning methods."

1.9 Research Roadmap

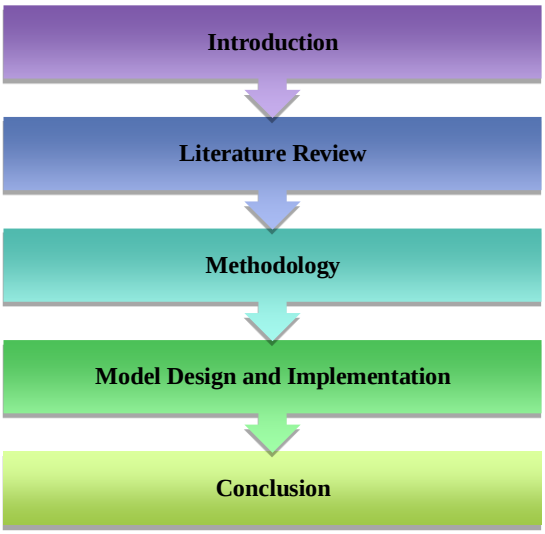


Figure 1: Research Roadmap

(Source: Self-created)

1.10 Conclusion

The introduction to the research topic regarding the "Ensemble Learning methods" that can be implemented by the organizations in the telecommunication sector to develop their current strategies of customer retention and acquiring through the effective prediction of churn rates have been elaborated in this chapter. The aims, objectives, and questions adequately guide through the process of the research and provide a direction to the determination of the specific research goals that are to be achieved at research completion. Overall, this chapter is highly sufficient in building the appropriate knowledge required to conduct the research. The problem areas linked with the implementation of the churn prediction techniques and the associated problems in the "ensemble learning methods" have been significantly considered in this chapter. Furthermore, the significance of the research explains the definite purpose of the study. It allows for the development of suitable strategies and methods to examine the intricacies of applying "Ensemble learning methods for churn prediction." The final section of the research Roadmap shows the procedural approach towards conducting the research.

Chapter 2: Literature Review

2.1 Introduction

The method involved in the "Ensemble learning" is considerably complex with respect to the application of the various algorithms and other critical factors associated with this method of data mining and analysis. Apart from the complexities of the "ensemble learning methods," there are varying aspects of the churn prediction methods which need to be studied elaborately in order to perform the churn prediction on a particular dataset based on the application of "ensemble learning methods." The reviewing of literature provides a critical overview of the concepts developed by other researchers and authors as well as the approaches that have been implemented by the researchers in applying the "ensemble learning methods for churn prediction." The literature evidence that is collected from the peer-reviewed journals and other scholarly articles can provide a detailed overview of the problems associated with the study of "ensemble learning methods."

The literature review also explores the theories and models of the corresponding research area. It defines the factors responsible for the appropriate implementation of "ensemble learning methods in respect to churn prediction." Therefore, the chapter shall highlight the independent variables and the dependent variables of the study to develop a conceptual framework of the research area. Finally, this chapter also produces the limitations in the literature sources and aims to resolve them in the current research.

2.2 Empirical Study

The study of the empirical literature is highly significant in researching as it provides an insight into the varied approaches and concepts that have been developed by the researchers on the "machine learning methods" involving the "ensemble models in predicting the customer churn rate in the telecom industry." Furthermore, the review of literature provides a comparative study of the previous works of the researchers regarding the “churn prediction” and weighs the advantages and disadvantages of the specific methods used by them in achieving the results to further incorporate them in the current research.

According to (Amin *et al.,* 2017), Churn Prediction allows the organizations to understand the "behavior of the customers" towards the service provider and further gives them the opportunity to differentiate among the "non-churn customers and churn customers" through a forecasting model. With a focus on the customer retention objective, the authors have found

that industries are severely inadequate in having a "rule-based Customer Churn Prediction (CCP) approach" within the telecommunication sector. The authors have proposed a "Rough Set Theory (RST)" for the "rule-based decision making" regarding the differentiation of "churn and non-churn customers." This journal has implemented for specific "rule generation algorithms corresponding to RST based CCP approach," namely, "exhaustive algorithm (EA), covering algorithm (CA), genetic algorithm (GA), and LEM2 Algorithm (LA)". With consequent evaluations and empirical results from the study, it has been identified that the "Genetic Algorithm is the most suitable technique" to perform learning on public data using decision rules. ***[Referred to Appendix 2]***

Another journal presented by the authors (Amin, Al-Obeidat *et al.*, 2019) has discussed the generation of appropriate "Churn prediction methods" in the telecom sector based on uncertain situations. This journal provides an extensive understanding of the approach that should be undertaken within the telecommunication industry to establish a secure business for the retention of customers. The paper provides a detailed analysis of the factors which are responsible for the customer churn using the "proposed model of churn prediction" that investigates the two different "types of samples, namely the upper distance samples and the lower distance samples". The samples were tested based on certain critical factors such as "accuracy, f measures, precision and recall" relevant to the uncertainty in the increment of sample size. The results of the proposed approach indicate that the "lower distance data set" have demonstrated better performance in comparison to the "upper distance data set". The Bayesian binomial test has been conducted to analyze the data distribution. ***[Referred to Appendix 3]***

As per the theories and approaches presented by (Mishra and U. Reddy, 2017), the "Ensemble-based classifiers" can extensively be used for the "churn prediction in the telecom industry." The "Churn Prediction process" has been demonstrated by the authors as a systematic process with various levels of risks involved in the model. Defining the business objectives of the telecom service provider forms for step in the prediction model, which is followed by the identification of the essential parameters for the prediction. In view of this process, the next step involves the development of a suitable prediction model that is tested with respect to the "low, medium, and high-risk levels". Finally, the model word predicts the "churn rate among the customers" with the major differentiation among the "churner and non-churner group". The authors have provided a detailed explanation of the "Ensemble-based classifiers and well-known base classifiers" that have been employed in the formulation of the "Churn Prediction model". The classifiers that have been chosen by the authors are the decision tree approach,

Bayesian Classifier, "support vector machine", "bagging, boosting and random forest methods". As per the results obtained based on the performance analysis of the classifiers, it is evident from the study that "the random forest classifier" provides the highest accuracy and sensitivity as well as the lowest error rate. Also, the study has indicated that the "random forest model" provides the best "precision, recall, and f score." Therefore, the authors suggest that the "random forest method" can provide the best solutions in response to the "Churn Prediction problems" corresponding to the telecom sector. ***[Referred to Appendix 4]***

These approaches and investigations support the knowledge required for the implementation of the "Ensemble-based machine learning in the prediction of the churn rate of customers" in this particular study.

2.3 Models and Techniques

The review of literature from different secondary sources has provided certain advanced models and methods based on the idea of innovating new classified models which can enhance the in-depth research considering the prediction of churn rates among customers in the telecommunication industry. The research conducted by eminent researchers and authors in recent years has proposed several techniques and methods which can be developed and implemented in further study.

"Prediction of customer churn rate" forms the "most dominant learning algorithms" for measuring the various business metrics such as the net revenue earned, net amount of profit and loss, which are related to the percentage of customer churn and "acquiring of new customers". The research presented by (A Hammoudeh, M. Fraihat, *et al.* 2019) furnished a "machine learning model" corresponding to churn prediction. In the study presented by the authors, an ensemble model was used to implement "a number of different models within a single system" to improve the "performance of the SEM model". The presented ensemble model in the study is highly effective in overcoming the limitations offered by single system models such as "high features or bias". The proposed "selective ensemble model" incorporates the "correct and incorrect result ratio" that is produced during the model training and effectively predicts the suitable model corresponding to the results ratio. The authors have demonstrated this model with the help of a data set consisting of 3333 customers and 21 attributes. The evaluation of this model in the research presented by the authors indicated that the model could effectively analyze 70% of the entire data set during the training and the rest of the data set during the testing. The model implemented the data preparation in the systematic flow from

"feature selection followed by feature transformation and imbalanced data." The primary evaluation criteria that were followed by the model are "accuracy, precision, and recall." Different algorithms have been implemented in the model, such as SVM, RF, ANN, and CNN. The use of the ANN technique has also been supported by (Khan *et al., 2019*).

Apart from the classification of models, the review of literature has also presented several techniques based on the models of implementing Churn prediction. The study presented by the authors (Ullah, B. Raza, *et al., 2019*) Indicates a model that incorporates machine learning techniques in combination with identification mechanisms to evaluate the sole factors responsible for the high customer churn rate. This technique was implemented on two different sets of data, where the customers were well labeled on different segments. The first data set consists of two segments of clients, namely, false clients and true clients, while the second data set consists of Churn clients and non-churn clients. Based on the evaluation of the model, it was observed by the author that the k-means clustering and RF algorithm produced the most "effective performance results for turn prediction."

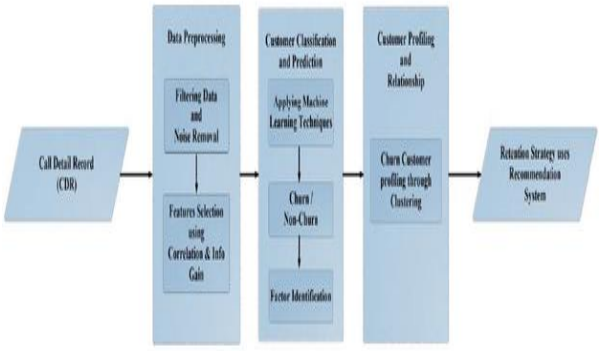


Figure 2: "Churn Prediction Model"

(Source: Ullal *et al.* 2019)

Another research work presented by (M. Saghir, Z. Bibi, *et al.*, 2019) broadly focuses on the importance of neural network algorithms for the churn prediction corresponding to the ensemble models and individual models. The models that have been presented by the authors have considered several methodologies, including "data acquisition, pre-processing, classifier training and testing using neural networks, deep learning and auto MLP, and finally the ensemble classification via bagging, majority voting, and AdaBoost." The "individual and ensemble models" are evaluated based on various classifiers such as "absolute error, accuracy, classification error, Kappa, precision, F measure, recall, and relative error." The results obtained from the implementation of the model with respect to the maximum performance have been exhibited by "Bagging MLP" in terms of precision, accuracy, Kappa, F- measure, and

recall. In contrast to that, "Bagging Neural Network" exhibits a high range of error classification along with precision and accuracy that renders it to be the best algorithm for the objective of Churn prediction.

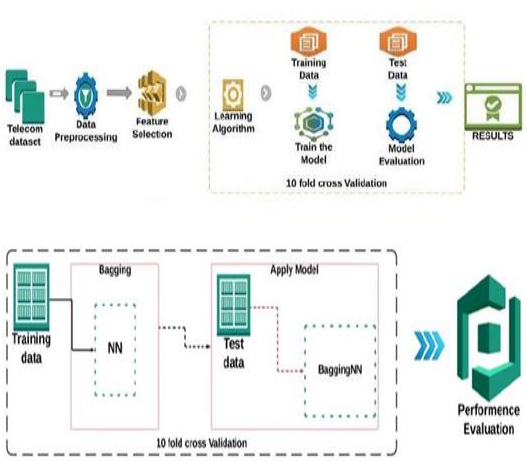


Figure 3: "Churn Prediction Model"

(Source: Saghir *et al.* 2019)

These models can be considered to further expand the areas of developing a suitable model for "Churn prediction using the methods of ensemble learning", thereby enhancing the accuracy and efficiency in enabling data prediction and analysis of a large data set of customers.

2.4 Literature Gap

Depending on the analysis of the empirical research that has been conducted, several gaps in the approaches and methods of "ensemble machine learning practices" employed by the authors have been identified for the improvement of the model in the subsequent study.

The authors (Krawczyk *et al.*, 2017) have indicated several limitations in their research corresponding to the implemented "Ensemble method of prediction". The limitations that have been highlighted in the journals are that the "classifier structure" is not updated with the appropriate models of the prediction that accounts for the unbalanced and delayed information. Another limitation of the research is the lack of "data stream classifiers that are "suitable for complex scenarios," such as "Changing class imbalance, verification latency, and multiple data streams." The study has also considered a survey which is significantly different from the actual scenario data sets, which are large and complex. The study also lacks the implementation of a "self-tuning ensemble mechanism," which would be able to produce better results for the prediction.

Another limitation perceived in the empirical study of the journal produced by the authors (Ke et al., 2017) based on the theories corresponding to the "Gradient-based One-Side Sampling

and Exclusive Feature Bundling" for the development of the GBDT algorithm indicates that the "Exclusive Feature Bundling" that is used in the research is only based on the sparsity of the dataset. At the same time, a better performance analysis could be done if the research considered both the sparse and non-sparse datasets for the development of the suitable "machine learning algorithm" in "Churn Prediction".

2.5 Conceptual Framework

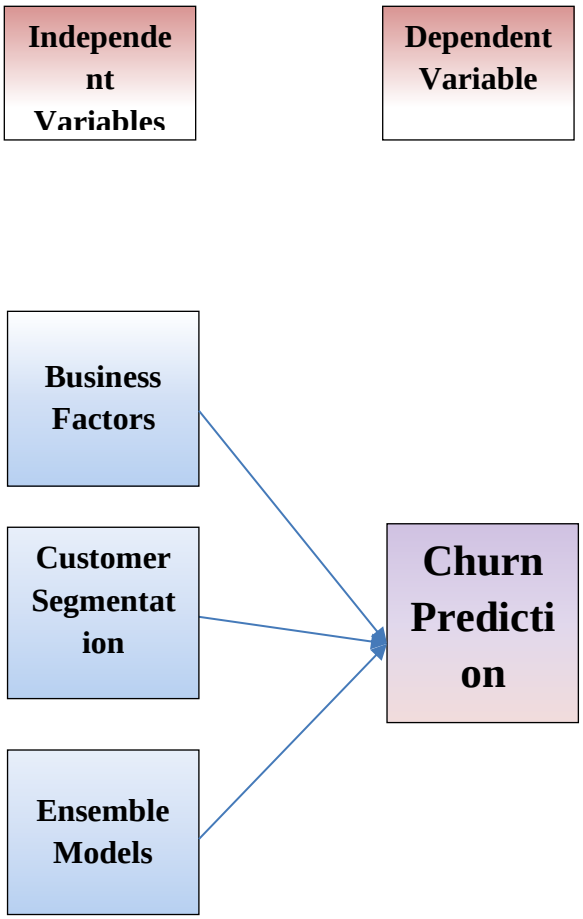


Figure 4: Conceptual Framework

(Source: Self-created)

Dependent Variable

"Churn Prediction": Churn prediction corresponds to the evaluation of the overall rate of customers who are most likely to switch to a different telecom service provider. This also indicates the process of estimating the number of churn customers in the organization as well as helps to provide differentiation among the "churn and non-churn customers" of the business. The churn prediction also involves investigating the reasons for the customer churn. Also, it helps to predict the risk labels for the prediction corresponding to a current dataset and a suitable "Machine learning model." There are several methods and algorithms which can help

in estimating and predicting the customer churn, as mentioned in the journal presented by the authors (Xia *et al., 2016*).

Independent Variables

Business Factors: (J. Xiao, L. Huang, et al., 2018) Indicate that the financial and cost of services are one of the prime factors that are responsible for customer churn. Apart from that, the business factors that are responsible for the customer churn prediction are the extensive competition in the telecom market that constantly requires the organizations to improve their base and employ an appropriate CRM through the churn risk estimation and development of policies that would help to ensure the customer retention. Also, in the journal by (Dahiya and Bhatia, 2015), the "cost of acquiring the new customers" are more with respect to retaining old customers. Therefore, the business factors are highly responsible for the incorporation of a strategic machine learning model that would enable the companies to estimate the churn rate and retain the customers.

Customer Segmentation: The need to understand customer behavior is broadly dependent on the analysis of the customer background and their requirements. Classification of the dataset of the customers and identifying the customer segments into the "churn and non-churn groups" are significantly responsible for the choice of the appropriate "ensemble learning method" corresponding to "the churn prediction." The classification of the dataset of the customers is an essential factor for the accurate "churn prediction." Authors (Yu *et al., 2018*) have proposed a "DBN based framework" that facilitates the classification of the customer data and their characteristics in order to provide a customer segmentation with respect to the objective of performing a "churn prediction" on an imbalanced dataset.

Ensemble Models: This corresponds to the developed approaches and models based on which the prediction is conducted. This is a formal machine learning method that supports the big data analysis of the customer data set. Authors (Umayaparvathi and Iyakutti, 2019) have provided an effective analysis of different models and their utilities in the tabular form from the extensive literature review in the study. All the ensemble methods have specific performance metrics that are used for the churn prediction.

2.6 Conclusion

The detailed review of the literature sources has been presented in this chapter to provide an insight into the techniques, methods, approaches, and "ensemble models" that have been used by different researchers in context to the extensive analysis performed by them on various

customer data sets. The literature review presents the elaborate discussion of the algorithms and models corresponding to the "ensemble methods of machine learning." The chapter has also explained the challenges faced by the telecom industries and their strategic adaptation of "churn prediction models" in order to prevent the outflow of customers from the company. Apart from that, the chapter has also highlighted the functionalities and intricacies involved with the use of different algorithms that are used for the appropriate feature selection and other classification processes. Furthermore, the secondary data also provided extensive knowledge about the methods involved with integrating several suitable methods to achieve accuracy and preciseness to the churn prediction process. The discussion of the models and techniques proposed by the researchers has a crucial role to play in the subsequent stages of developing the relevant "ensemble method" for the data analysis.

Moreover, the gaps in the literature studies have indicated the specific areas of limitations which can be considered for improvements in the current research studies. Finally, this chapter provides a detailed overview of the factors that are responsible for the appropriate data analysis and implementation of the "ensemble learning method for predicting customer churn."

Chapter 3: Methodology

In this research, we aim to use two tools RapidMiner which is data science software, which automates the process of data preparation, machine learning, and predictive modeling models. The second tool that we are using is Python, using a jupyter notebook as IDE and H2O machine learning models. The aim is to find the best models using these having high accuracy and predicting the number of churns effectively than the other models. Figure 5 below shows the process flow diagram of the research. CRISP-DM methodology has been followed to maintain the uniform flow of research work.

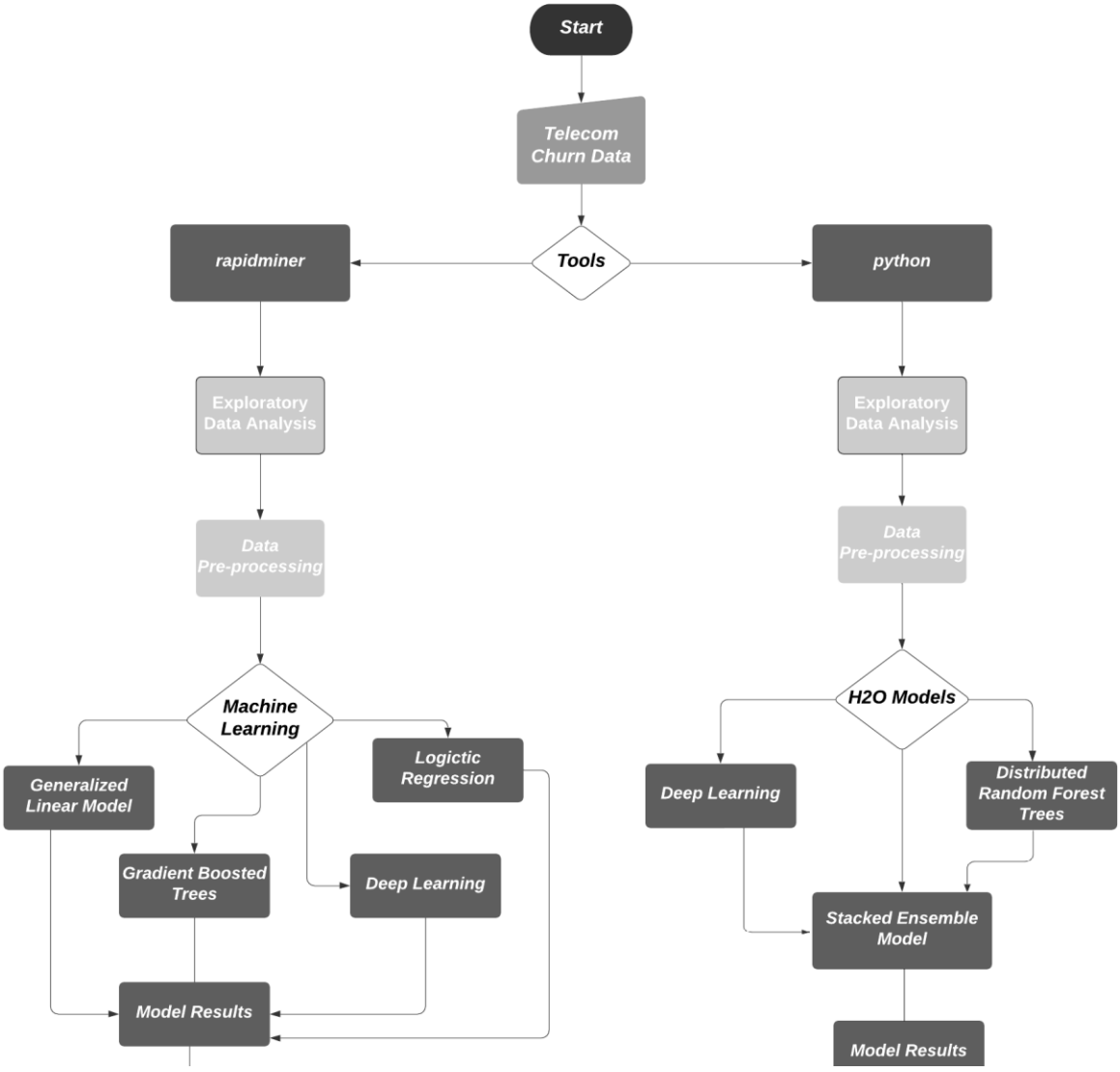
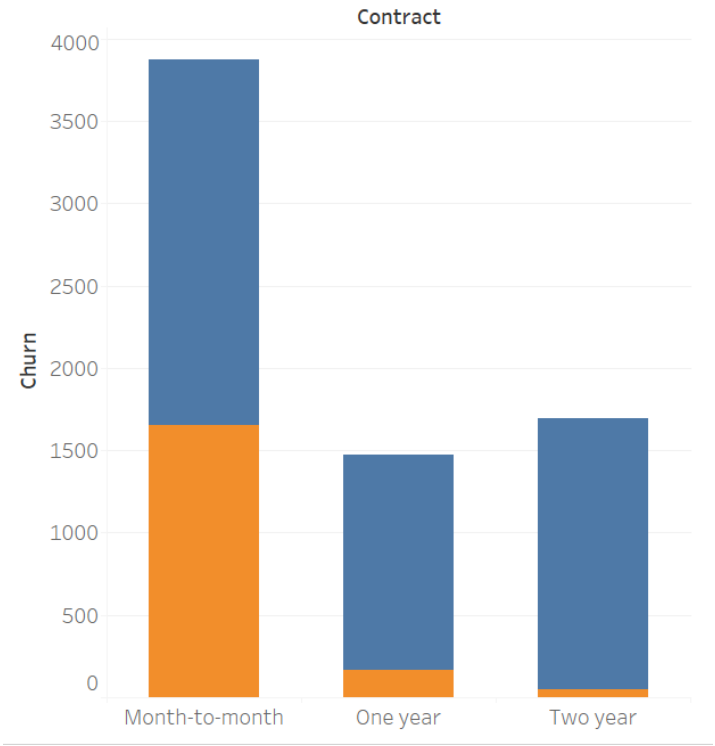


Figure 5: Model Framework

(Source: Self-created)

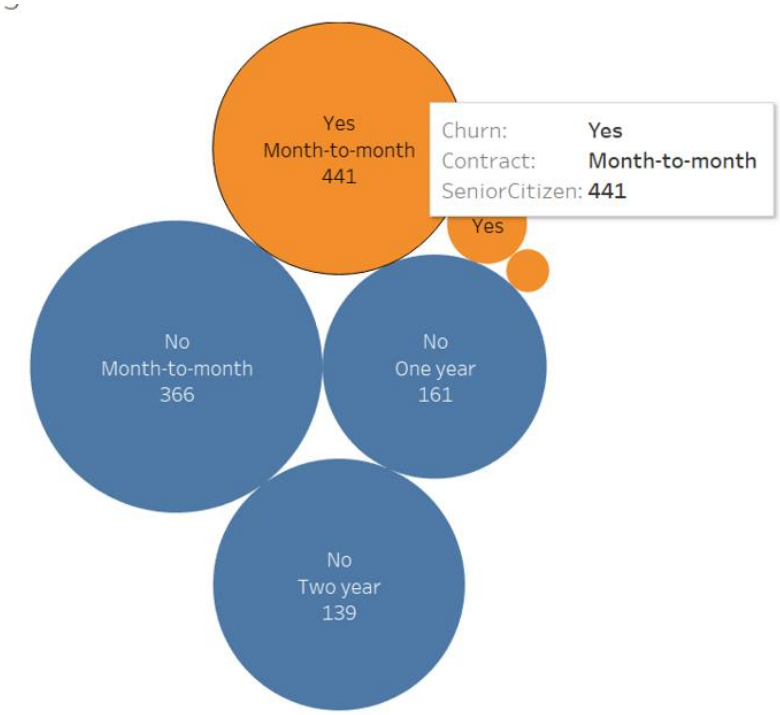
3.1 Business Understanding

Here, Tableau is used as a Business Intelligence tool to understand the Business notion with the Problem statement that is Churning of telecom Customers, as a start of CRISP-DM methodology. Insights provided below will give a visual understanding, where red color represents churned customers that is the Area of interest.



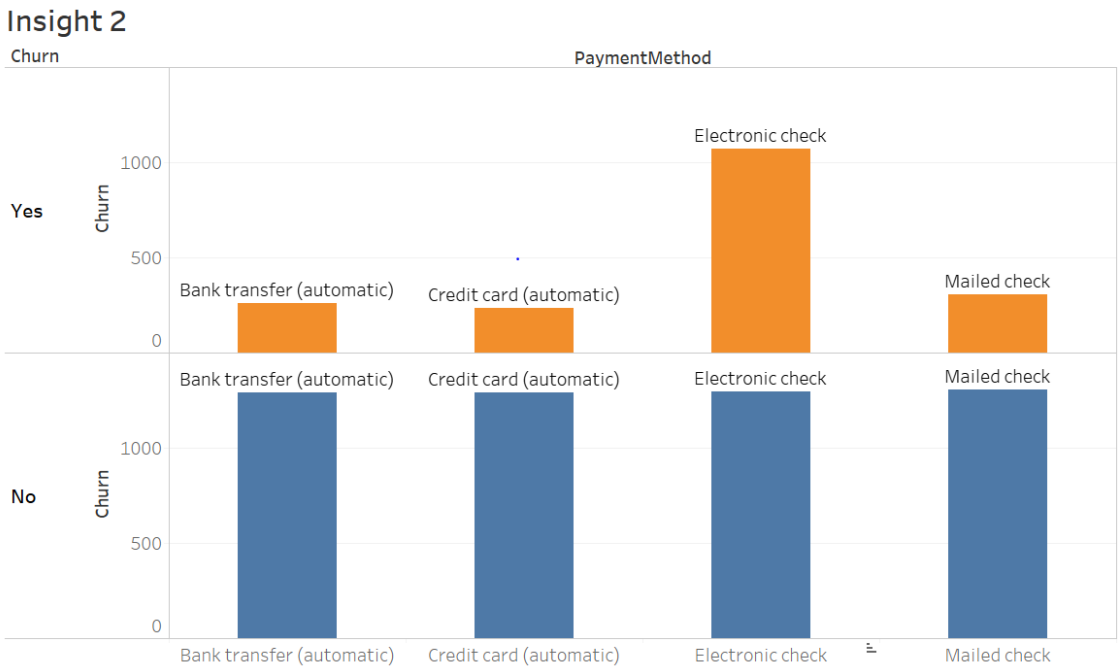
Insight 1

Here it is observed that customers having monthly plans are carrying more percentage of churn. So, it is clearly observable that the competition in prepaid customers who are paying monthly bills is high. So the telecom companies need to be one step ahead while offering monthly plans for customers to keep up in the market.



Insight 1.2

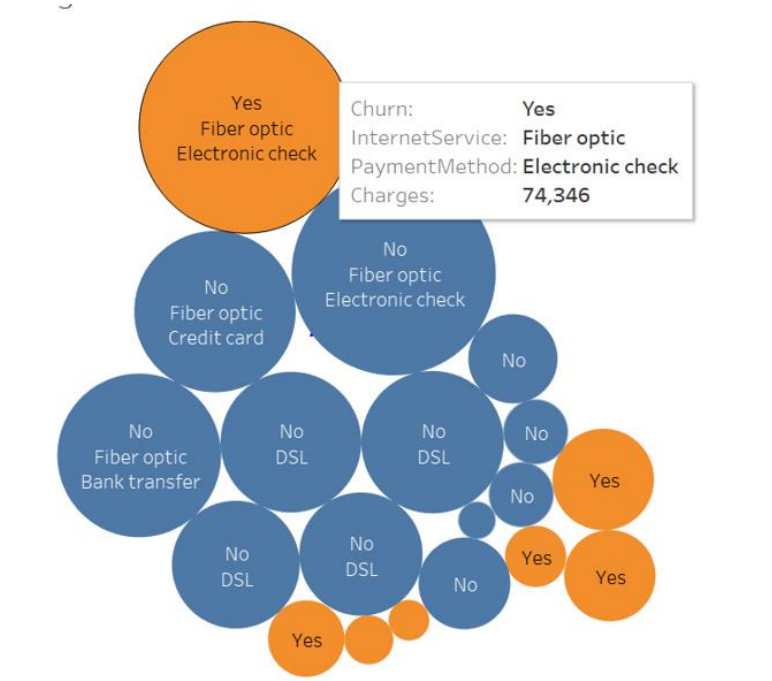
When mined dip into Insight 1, It was noticed that the senior citizen with a monthly plan should be given special attention. Here the Insight 1.2 gives us slight an overview of a senior citizen having a monthly plan basis that needed to be addressed timely.



Insight 2

In our dataset, four payment methods are existing, namely Bank transfer, Credit card, Electronic Check, Mailed check.

Now Insight 2 gives a clear understanding that electronic check customers are more likely to churn comparatively.



Insight 2.1

Now, digging dip into Insight 2 churning based on electronic check payment, it was observed that the customers with fiber optics internet service incur maximum charges and are more likely to churn.

3.2 Data Understanding

The process starts by importing the data into the tools. The data that is being used for this research is available on https://www.kaggle.com/blastchar/telco-customer-churn. The data is a sample of IBM Sample Datasets used to predict customer churns. The purpose of churn prediction is to the organization can formulate in advance the schemes and offers for such customers. The data follows the same steps for Exploratory Data Analysis and Pre-processing in both the tools. In Exploratory data analysis, we familiarize ourselves with the data. The statistics in the data, the data structure, and shape, checking if the data has missing values, checking if the classes in the data are balanced, and so on.

3.3 Data Preparation

In data pre-processing, we perform data cleaning, data balancing, and features selection part. In data cleaning, the missing values, if present in the data, is treated. The 'total charges' column in the data was the only one having missing data in 11 rows; these rows are removed from the system. To balance the data, we resampled it by SMOTE Upsampling in RapidMiner and Random oversampling in Python for class balancing.

3.4 Modelling

The machine learning models that are used in the RapidMiner tool are Generalized linear model, Gradient Boosted Trees, Deep Learning, and Logistic Regression. A generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. Gradient boosted Trees is the ensemble model where decision trees are boosted with random step sizes; this generates a prediction model in the form of an ensemble of weak prediction models. Deep learning is an AI function that mimics the workings of the human brain in processing data for use in detecting objects, recognizing speech, translating languages, and making decisions. Deep learning AI is able to learn without human supervision, drawing from data that is both unstructured and unlabeled. Logistic regression is a statistical model that, in its basic form, uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

For the python coding, the H2O platform is used. H2O.ai is the open source leader in AI and machine learning with a mission to democratize AI for everyone. H2O's Stacked Ensemble method is a supervised ensemble machine learning algorithm that finds the optimal combination of a collection of prediction algorithms using a process called stacking. Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised, or unsupervised. H2O's Deep Learning is based on a multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation. The network can contain a large number of hidden layers consisting of neurons with tanh, rectifier, and maxout activation functions. Advanced features such as adaptive learning rate, rate annealing, momentum training, dropout, L1 or L2 regularization, checkpointing, and grid search enable high predictive accuracy. Each computes node trains a copy of the global model parameters on its local data with multi-threading (asynchronously) and contributes periodically to the global model via model averaging across the network. Distributed Random Forest (DRF) is a powerful classification and regression tool. When given a set of data, DRF generates a forest of classification or regression trees, rather than a single classification or regression tree. Each of these trees is a weak learner built on a subset of rows and columns. More trees will reduce the

variance. Both classification and regression take the average prediction over all of their trees to make a final prediction, whether predicting for a class or numeric value.

The model's performances are then evaluated. The metrics used for the performance evaluation are accuracy score, classification error, AUC, precision, recall, F-measure, sensitivity, specificity, and confusion matrix. Accuracy is the ratio of correctly predicted observation to the total observations. The classification error rate is simply the fraction of training observations in a region that do not belong to the most common class. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. Precision is the ratio of correctly predicted positive observations of the total predicted positive observations. The recall is the ratio of correctly predicted positive observations to the all observations in the actual class. F-Measure or F1 Score is the weighted average of Precision and Recall. Model sensitivity is the ability of a test to correctly identify those that belong in the class(true positive rate), whereas model specificity is the ability of the test to correctly identify those that do not belong in the class (true negative rate). A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

3.5 Evaluation

Data is divided into 80% training and 20% testing set to track the testing error. Testing error plays a vital role as it is based on a completely disjoint set.
Once models were developed, sharp and improved findings were tested based on various merit matrix; Key factors included accuracy, AUC, f1 score, confusion matrix and etc.
For any Machine Learning process, accuracy is very important since it's purposeful to know how well the model will perform.

3.6 Deployment

The deployment stage is crucial to project success. The deployment phase is the final stage in crisp DM; when the model is ready for use, there needs to be a business operations strategy to work. The knowledge obtained would need to be organized and delivered in a way that can be seen by the customers.
Deployment may be on a variety of platforms, such as the web portal, smartphone phone, tool-based, or report production, according to end-user needs. ***[Referred to Appendix 5]***

Chapter 4: Finding Analysis

This section is divided into two parts the firsts section covers the implementation and results of RapidMiner and the second section covers the implementation and results of Python

4.1. RapidMiner

RapidMiner is the tool that automates the entire process of data analysis and predictive modeling. The data is first uploaded into the environment. Once this is done, the attribute selection is made. By attribute selection, we mean that the target variable is set which is 'churn' column in our data, 'customerid' was dropped from the data as it was found to be insignificant, 'gender' and 'PhoneService' was also dropped as it high correlation factor that would lead to the biases in the data. The role of the features and target variables is done.

For data pre-processing, the class balance was checked, and the difference between the total number of samples for both the classes 'yes' and 'no' was huge. For balancing the classes, SMOTE Up-sampling is used. Synthetic Minority Oversampling Technique (SMOTE) is a very popular used to improve oversampling. In this method, the data is upscaled to replicate the class that is a minority.

The data is then split into training and test set. The split that was performed is 80-20 split, which means that 80% of the entire data is being saved into the training set, and the remaining 20% is loaded into the test set. The training set is then used to train the models, and the test set is used to validate the model performance. Figure 6 below shows the RapidMiner flow of the data.

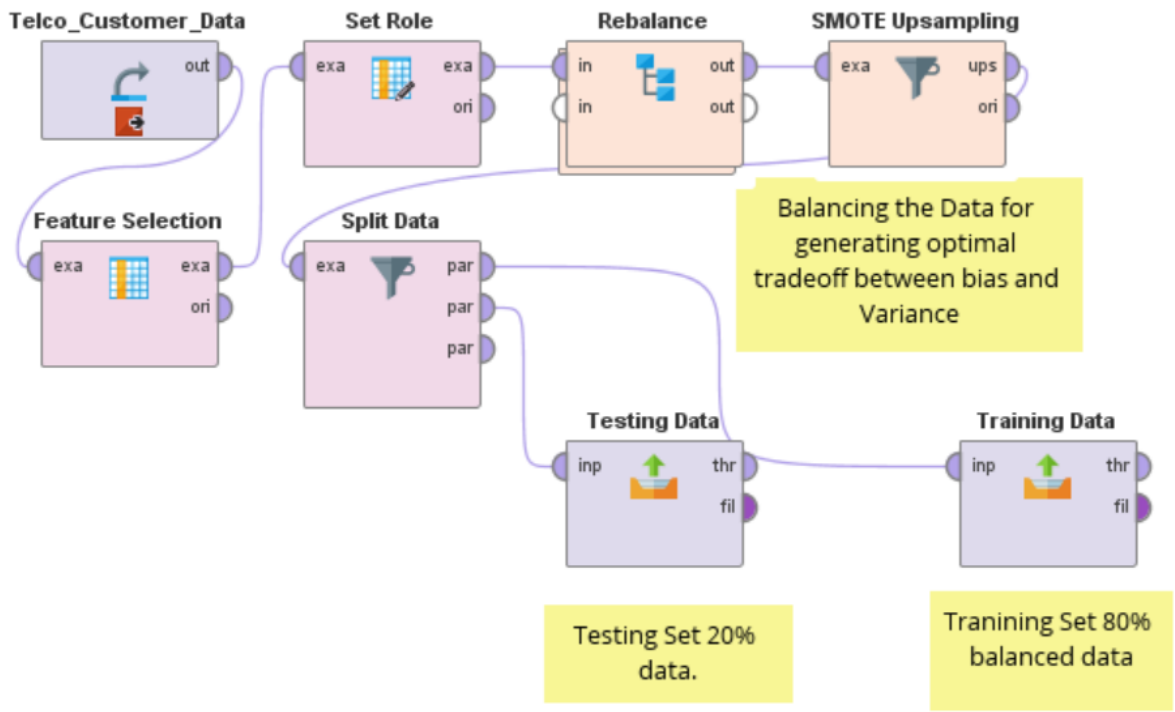


Figure 6: RapidMiner flow of the data.

(Source: Generated by RapidMiner)

Figure 7 below shows the Resampling process output in RapidMiner, here Filter positive filters the positive results which are churn = Yes for the data and similarly operates Filter negative which filters negative which is churn = No

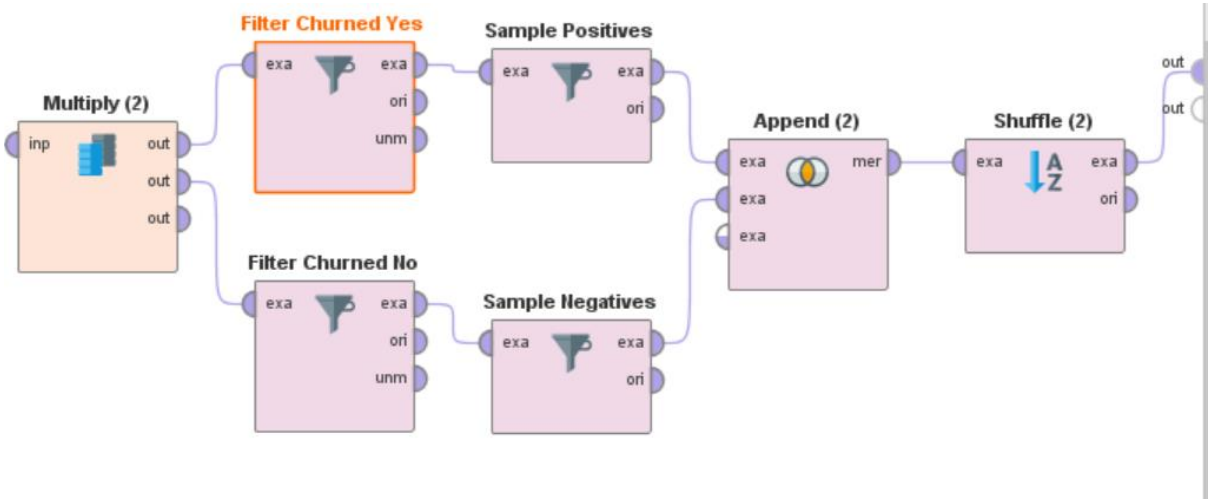


Figure 7: Resampling process

(Source: Generated by RapidMiner)

The generalized linear model is implemented on the data to predict the class of the data into 'yes' or 'no' for the Churn target field. The generalized linear model is the generalized version of the linear regression model. The model was trained on the training set and then evaluated for its performance using the test set. The performance evaluation report of the model is shown below in figure 8. The accuracy of the model was recorded as 74.2 %, classification error that the model made was found to be 25.8%, AUC is recorded as 83.1%, the precision of the model was 69.5%, the recall for the model was found to be 87%, the f-measure also known as F1 Score was recorded at 77.2%, the model had 87% sensitivity and 61.4% specificity.

Generalized Linear Model - Performance

Criterion	Value	Standard Deviation
Accuracy	74.2%	± 1.3%
Classification Error	25.8%	± 1.3%
AUC	83.1%	± 1.1%
Precision	69.5%	± 2.4%
Recall	87.0%	± 2.8%
F Measure	77.2%	± 0.8%
Sensitivity	87.0%	± 2.8%
Specificity	61.4%	± 4.1%

Figure 8: Generalized Linear Model Performance

(Source: Generated by RapidMiner)

Figure 9 below shows the predictions that the model has made.

Generalized Linear Model - Predictions

Row No.	Churn	prediction(C...	confidence(...	confidence(...	cost	Partner	Dependents	MultipleLines	InternetServ.
1	Yes	Yes	0.484	0.516	0.032	Yes	Yes	No	No
2	Yes	Yes	0.201	0.799	0.598	Yes	Yes	Yes	Fiber optic
3	Yes	Yes	0.257	0.743	0.486	No	No	No phone ser...	DSL
4	Yes	Yes	0.299	0.701	0.401	No	No	No	DSL
5	Yes	Yes	0.166	0.834	0.668	No	No	Yes	Fiber optic
6	No	Yes	0.259	0.741	0.482	No	No	No	DSL
7	No	No	0.694	0.306	0.388	Yes	Yes	No	Fiber optic
8	No	No	0.748	0.252	0.495	Yes	Yes	Yes	Fiber optic

Figure 9: Generalized Linear Model Predictions

(Source: Generated by RapidMiner)

Figure 10 below shows the confusion matrix of the model. The true 'yes' predictions in that the model are 373, and false 'yes' that the model made are 56. Similarly, the true 'no' predictions in that the model are 261, and false 'no' that the model made is 164.

Confusion Matrix

	true No	true Yes	class precision
pred. No	261	56	82.33%
pred. Yes	164	373	69.46%
class recall	61.41%	86.95%	

Figure 10: Generalized Linear Model Confusion matrix

(Source: Generated by RapidMiner)

Gradient Boosted tree is implemented on the data to predict the class of the data into 'yes' or 'no' for the Churn target field. The gradient boosted model is the predicted model, which helps improve the performance of the weak learner model, which in this case, is the decision tree. The model was trained on the training set and then evaluated for its performance using the test set. The performance evaluation report of the model is shown below in figure 11. The accuracy of the model was recorded as 73.8 %, classification error that the model made was found to be 26.2%, AUC is recorded as 83.2%, the precision of the model was 68.2%, the recall for the model was found to be 90%, the f-measure also known as F1 Score was recorded at 77.5%, the model had 90% sensitivity and 57.4% specificity.

Gradient Boosted Trees - Performance

Performances

Criterion	Value	Standard Deviation
Accuracy	73.8%	± 2.4%
Classification Error	26.2%	± 2.4%
AUC	83.2%	± 1.4%
Precision	68.2%	± 3.2%
Recall	90.0%	± 2.5%
F Measure	77.5%	± 1.7%
Sensitivity	90.0%	± 2.5%
Specificity	57.4%	± 6.4%

Figure 11: Gradient Boosted Trees Model Performance

(Source: Generated by RapidMiner)

Figure 12 below shows the confusion matrix of the model. The true 'yes' predictions in that the model are 386, and false 'yes' that the model made are 43. Similarly, the true 'no' predictions in that the model are 244, and false 'no' that the model made is 181.

Confusion Matrix

	true No	true Yes	class precision
pred. No	244	43	85.02%
pred. Yes	181	386	68.08%
class recall	57.41%	89.98%	

Figure 12: Gradient Boosted Trees Model Confusion matrix

(Source: Generated by RapidMiner)

Deep Learning model is implemented on the data to predict the class of the data into 'yes' or 'no' for the Churn target field. Deep Learning is the implementation of an artificial neural network in the RapidMiner platform. The model was trained on the training set and then evaluated for its performance using the test set. The performance evaluation report of the model is shown below in figure 13. The accuracy of the model was recorded as 73.8 %, classification error that the model made was found to be 26,2%, AUC is recorded as 82.9%, the precision of the model was 69.8%, the recall for the model was found to be 84.4%, the f-measure also known as F1 Score was recorded at 76.3%, the model had 84% sensitivity and 63.1% specificity.

Deep Learning - Performance

Performances

Criterion	Value	Standard Deviation
Accuracy	73.8%	± 1.5%
Classification Error	26.2%	± 1.5%
AUC	82.9%	± 1.6%
Precision	69.8%	± 3.1%
Recall	84.4%	± 3.6%
F Measure	76.3%	± 1.7%
Sensitivity	84.4%	± 3.6%
Specificity	63.1%	± 4.5%

Figure 13: Deep Learning Model Performance

(Source: Generated by RapidMiner)

Figure 14 below shows the confusion matrix of the model. The true 'yes' predictions in that the model are 362, and false 'yes' that the model made are 67. Similarly, the true 'no' predictions in that the model are 268, and false 'no' that the model made is 157.

Confusion Matrix

	true No	true Yes	class precision
pred. No	268	67	80.00%
pred. Yes	157	362	69.75%
class recall	63.06%	84.38%	

Figure 14: Deep Learning Model Confusion Matrix

(Source: Generated by RapidMiner)

The logistic Regression model is implemented on the data to predict the class of the data into 'yes' or 'no' for the Churn target field. Logistic regression is a simple and most popular classification model for binary classification. Binary classification means that the model can work on two classes' maximum for the classification task. The model was trained on the training set and then evaluated for its performance using the test set. The performance evaluation report of the model is shown below in figure 15. The accuracy of the model was recorded as 73.2 %, classification error that the model made was found to be 26.8%, AUC is recorded as 82.9%, the precision of the model was 67.7%, the recall for the model was found to be 89.1%, the f-measure also known as F1 Score was recorded at 76.9%, the model had 89.1% sensitivity and 57.4% specificity.

Logistic Regression - Performance

Performances

Criterion	Value	Standard Deviation
Accuracy	73.2%	± 1.1%
Classification Error	26.8%	± 1.1%
AUC	82.9%	± 1.1%
Precision	67.7%	± 1.6%
Recall	89.1%	± 2.2%
F Measure	76.9%	± 0.9%
Sensitivity	89.1%	± 2.2%
Specificity	57.1%	± 2.7%

Figure 15: Logistic Regression Model Performance

(Source: Generated by RapidMiner)

Figure 10 below shows the confusion matrix of the model. The true 'yes' predictions in that the model are 382, and false 'yes' that the model made are 47. Similarly, the true 'no' predictions in that the model are 243, and false 'no' that the model made is 182.

Confusion Matrix

	true No	true Yes	class precision
pred. No	243	47	83.79%
pred. Yes	182	382	67.73%
class recall	57.18%	89.04%	

Figure 16: Logistic Regression Model Confusion Matrix

(Source: Generated by RapidMiner)

4.2. Python

For data preparation and predictive modeling using the H2O framework in Python, we have used the Jupyter notebook as a tool. The data is loaded into the environment. Figure 17 shows the data.

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSup
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	...	Yes	
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	...	No	
6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	...	No	
7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	...	Yes	
6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	...	No	

Figure 17: Data

(Source: Code generated output screen)

The data columns are shown in figure 18 below. These are the names of the columns that we have in the data.

```
Index(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',
       'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling',
       'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn'],
      dtype='object')
```

Figure 18: Data Columns

(Source: Code generated output screen)

The statistics are generated for the data. We have tenure, 'SeniorCitizen,' and 'MonthlyCharges' as a numeric field, so the statistics about the mean, standard deviation, count, first, second and third quartiles, minimum and maximum values is generated only for these columns.

	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

Figure 19: Data Statistics

(Source: Code generated output screen)

The data is now analyzed for its overall structure. As shown in Figure 20 below, the data has an index range from 0 to 7042, which means the total number of rows in data is 7043. There are 21 columns in data. The data has integer values in column tenure, and 'SeniorCitizen,' float type in column 'MonthlyCharges,' and all the remaining columns are object type, which is either string or date type.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

Figure 20: Data Information

(Source: Code generated output screen)

The data is now checked for the missing or null values. The count is generated for all the columns in the data. Figure 22 shows the code output for this missing values count. Total-Charges column is the only one with missing data, having 11 points missing. The treatment of these missing values is done by dropping the rows with missing values.

```
customerID           0
gender               0
SeniorCitizen        0
Partner              0
Dependents           0
tenure               0
PhoneService         0
MultipleLines        0
InternetService      0
OnlineSecurity       0
OnlineBackup         0
DeviceProtection     0
TechSupport          0
StreamingTV          0
StreamingMovies      0
Contract             0
PaperlessBilling     0
PaymentMethod        0
MonthlyCharges       0
TotalCharges        11
Churn                0
dtype: int64
```

Figure 22: Null value count

(Source: Code generated output screen)

The class balance is checked next to see the total number of data points in churn, target field. The yes churn records are more 5163 than no-churn records 1869; this concludes that the data is not balanced. The output of the count generated for the classes is shown in Figure 23 below.

```
No     5163
Yes    1869
Name: Churn, dtype: int64
```

Figure 23: Class balance

(Source: Code generated output screen)

Figure 34 below shows the insight from the data based on the gender of the customers churning. We can see that the percentage difference, based on gender, is not much, so there the customers churn irrespective of their gender, so the plans made for customer retention should not be based on gender.

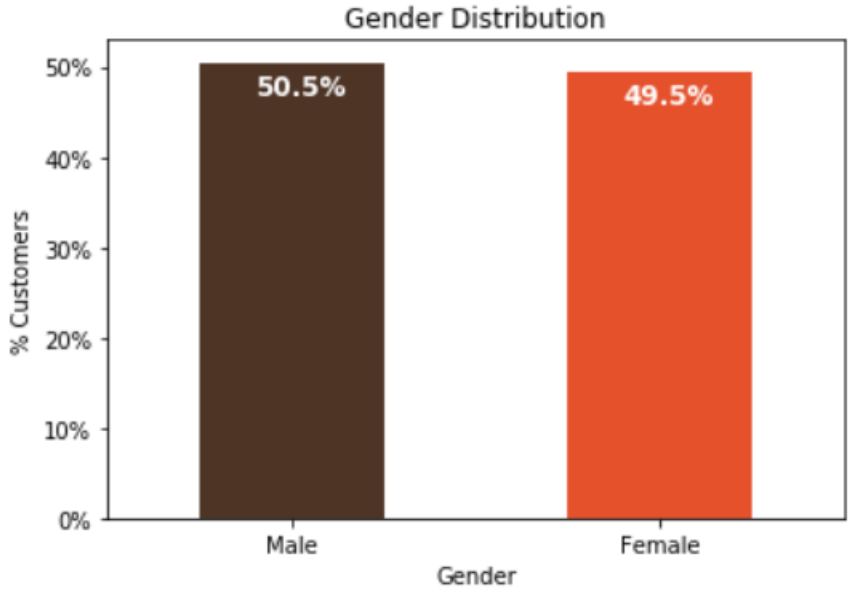


Figure 24: Gender Distribution plot

(Source: Code generated output screen)

Figure 25 below, shows the insight about the data that Senior Citizens have 16.2% churn records as positive and 83.8% churn records as negative. The senior citizens are less likely to churn, so the telecom industry should make plans for the younger generation.

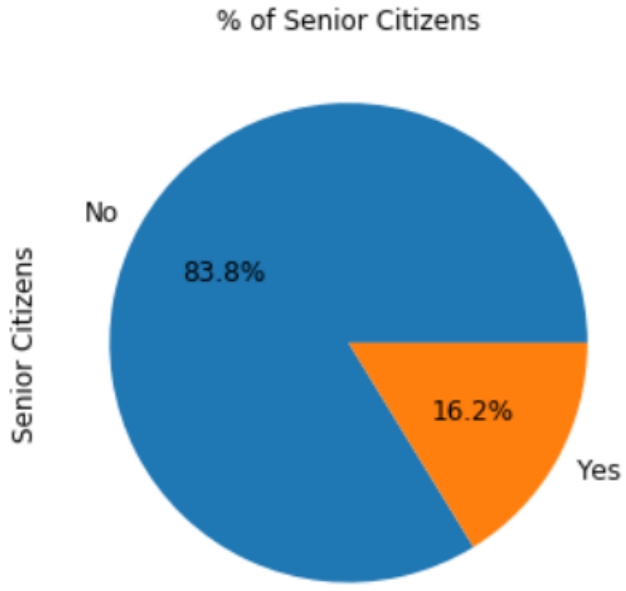


Figure 25: Senior Citizen Churn plot

(Source: Code generated output screen)

Label encoding is performed in the data as the classification models don't perform well when data has qualitative values. By qualitative values, we mean the values in the column being string or character in type. The qualitative data is then encoded into numeric data, which is the quantitative data. Figure 26 below shows the data after the label encoding is performed on the data.

SeniorCitizen	Partner	Dependents	tenure	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	Contract	PaperlessBil
0	1	0	1	1	0	0	2	0	0	0	
0	0	0	34	0	0	2	0	2	0	1	
0	0	0	2	0	0	2	2	0	0	0	
0	0	0	2	0	1	0	0	0	0	0	
0	0	0	8	2	1	0	0	2	0	0	
0	0	1	22	2	1	0	2	0	0	0	
0	0	0	10	1	0	2	0	0	0	0	
0	1	0	28	2	1	0	0	2	2	0	
0	1	1	13	0	0	2	0	0	0	0	
0	0	0	16	0	2	1	1	1	1	2	

Figure 26: Data after Label Encoding

(Source: Code generated output screen)

The class count was not balanced as we saw above in Figure 23; the resampling of the class is performed using the oversampling method. This made the count of the data points in both the classes to be equal as 5163.

```
Random over-sampling:
1    5163
0    5163
Name: Churn, dtype: int64
```

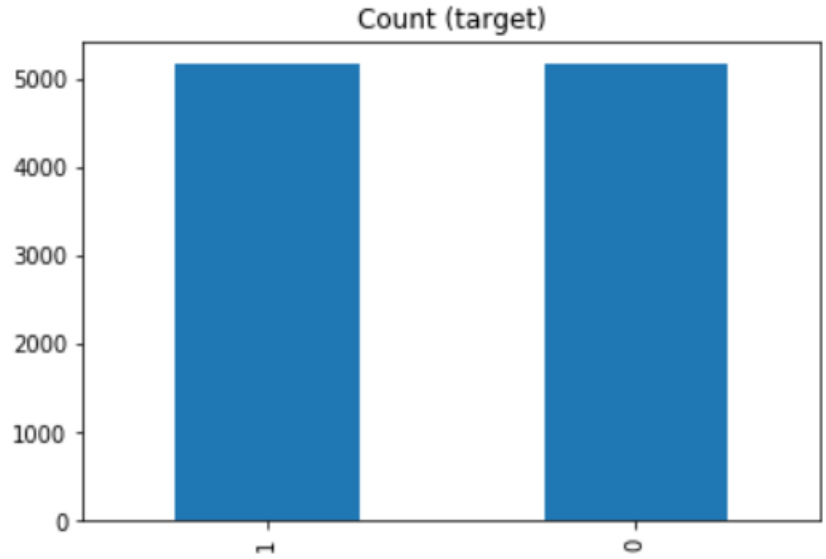


Figure 27: Balanced out classes after resampling

(Source: Code generated output screen)

A Chi-square test is performed to check for the hypothesis testing of the features where their significance level is recorded as scores, and the probability score of the features is generated for the level of insignificance.

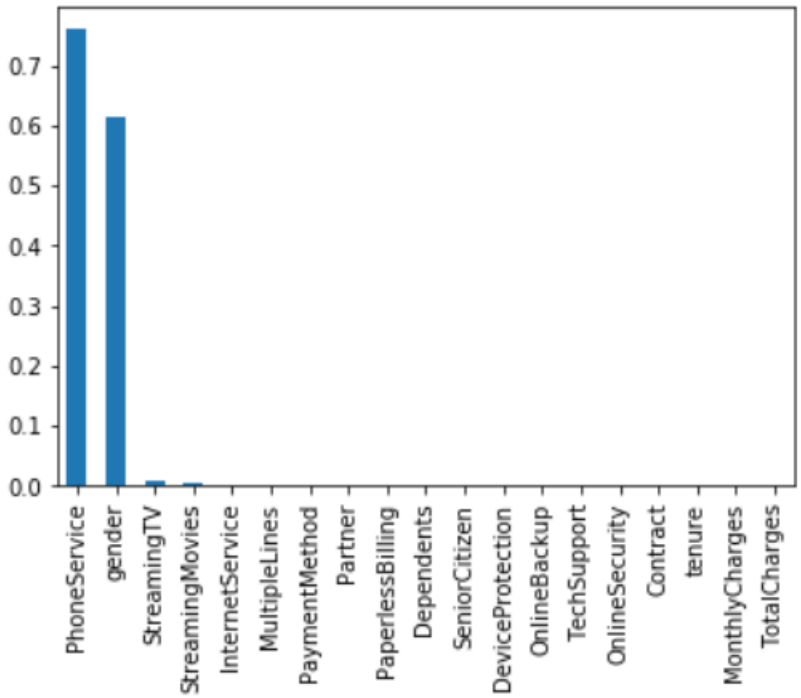


Figure 27: Chi-Square Probability plot

(Source: Code generated output screen)

From the graph, as shown in figure 27 above, its visible that 4 variables have bars, which means their probability of being insignificant is high. These four features are then dropped from the data.

The data is now split into training and testing set; the split ratio is used similarly as with RapidMiner. The model performance gets affected by the split of the data, so the split ratio was kept the same as we wanted similar scenarios for both the tools.

The stacked Ensemble model is implemented on the data to predict the class of the data into 'yes' or 'no' for the Churn target field. The stacked ensemble model is generated by stacking two models, one model on top of another. In this research, the Generalized Linear Model and Gradient Boosted Machine model are clubbed as the models to be stacked. The reason behind this is that the GLM's performance was better in RapidMiner, by stacking GLM with GBM we aiming to see if the stacking helps to improve the overall performance of the model. The model was trained on the training set and then evaluated for its performance using the test set. The performance evaluation report of the model is shown below in figure 29. The accuracy of the model was recorded as 80 % which show that stacking has improved the performance as in RapidMiner the performance of models separately was very less comparatively; classification

error that the model made was found to be 14%, AUC is recorded as 82%, the precision of the model was 88%, the recall for the model was found to be 5%, the f-measure also known as F1 Score was recorded at 73%, and the model had 88% specificity. Figure 28 below shows the confusion matrix of the model. The true 'yes' predictions in that the model are 285, and false 'yes' that the model made are 64. Similarly, the true 'no' predictions in that the model are 325, and false 'no' that the model made are 703.

```
ModelMetricsBinomialGLM: stackedensemble
** Reported on test data. **

MSE: 0.14117101495019865
RMSE: 0.3757273146181931
LogLoss: 0.4349752632196161
Null degrees of freedom: 1376
Residual degrees of freedom: 1374
Null deviance: 1560.689700673565
Residual deviance: 1197.9218749068227
AIC: 1203.9218749068227
AUC: 0.8270349971569687
AUCPR: 0.6271257053529439
Gini: 0.6540699943139374

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.17522386684809632:
```

		0	1	Error	Rate
0	0	703.0	325.0	0.3161	(325.0/1028.0)
1	1	64.0	285.0	0.1834	(64.0/349.0)
2	Total	767.0	610.0	0.2825	(389.0/1377.0)

Figure 28: Stacked ensemble Model confusion Matrix

(Source: Code generated output screen)

Maximum Metrics: Maximum metrics at their respective thresholds

	metric	threshold	value	idx
0	max f1	0.175224	0.594369	286.0
1	max f2	0.094029	0.738180	345.0
2	max f0point5	0.517696	0.603265	137.0
3	max accuracy	0.552913	0.801743	122.0
4	max precision	0.887776	1.000000	0.0
5	max recall	0.051729	1.000000	398.0
6	max specificity	0.887776	1.000000	0.0
7	max absolute_mcc	0.433378	0.455179	172.0
8	max min_per_class_accuracy	0.219487	0.734436	261.0
9	max mean_per_class_accuracy	0.175224	0.750236	286.0
10	max tns	0.887776	1028.000000	0.0
11	max fns	0.887776	348.000000	0.0
12	max fps	0.051482	1028.000000	399.0
13	max tps	0.051729	349.000000	398.0
14	max tnr	0.887776	1.000000	0.0
15	max fnr	0.887776	0.997135	0.0
16	max fpr	0.051482	1.000000	399.0
17	max tpr	0.051729	1.000000	398.0

Figure 29: Stacked ensemble Model Performance

(Source: Code generated output screen)

Figure 30 below shows the group cumulative response rate.

	group	cumulative_data_fraction	lower_threshold	lift	cumulative_lift	response_rate	score	cumulative_response_rate	cumulative_score	capture_r
0	1	0.010167	0.838989	3.945559	3.945559	1.000000	0.858256	1.000000	0.858256	0.040
1	2	0.020334	0.817694	3.381907	3.663733	0.857143	0.824808	0.928571	0.841532	0.034
2	3	0.030501	0.810078	3.100082	3.475849	0.785714	0.814436	0.880952	0.832500	0.031
3	4	0.040668	0.797292	2.536431	3.240995	0.642857	0.802600	0.821429	0.825025	0.025
4	5	0.050109	0.788393	2.731541	3.145011	0.692308	0.792495	0.797101	0.818896	0.025
5	6	0.100218	0.688558	2.573190	2.859101	0.652174	0.738984	0.724638	0.778940	0.128
6	7	0.150327	0.588016	2.172916	2.630372	0.550725	0.633553	0.666667	0.730478	0.108
7	8	0.200436	0.509977	1.944188	2.458826	0.492754	0.547497	0.623188	0.684732	0.097
8	9	0.299927	0.335937	1.382386	2.101750	0.350365	0.422284	0.532688	0.597673	0.137
9	10	0.400145	0.202569	1.229413	1.883270	0.311594	0.262434	0.477314	0.513712	0.123
10	11	0.500363	0.140359	0.972094	1.700771	0.246377	0.170769	0.431060	0.445023	0.097
11	12	0.599855	0.098026	0.863991	1.561983	0.218978	0.118461	0.395884	0.390860	0.085
12	13	0.700073	0.072159	0.371683	1.391587	0.094203	0.083260	0.352697	0.346826	0.037
13	14	0.799564	0.060182	0.057599	1.225596	0.014599	0.065627	0.310627	0.311836	0.005
14	15	0.899782	0.054094	0.114364	1.101827	0.028986	0.057242	0.279257	0.283479	0.011
15	16	1.000000	0.051369	0.085773	1.000000	0.021739	0.052273	0.253450	0.260308	0.008

Figure 30: Stacked Ensemble group cumulative response

(Source: Code generated output screen)

Deep Learning is implemented on the data to predict the class of the data into 'yes' or 'no' for the Churn target field. Deep learning is the implementation of artificial neural networks in the H2O framework. The model was trained on the training set and then evaluated for its performance using the test set. The performance evaluation report of the model is shown below in figure 32. The accuracy of the model was recorded as 79.5 %, classification error that the model made was found to be 14%, AUC is recorded as 82%, the precision of the model was 80%, the recall for the model was found to be 2%, the f-measure also known as F1 Score was recorded at 73%, the model had 80% specificity. Figure 31 below shows the confusion matrix of the model. The true ‘yes’ predictions in that the model are 268, and false ‘yes’ that the model made are 87. Similarly, the true ‘no’ predictions in that the model are 281, and false 'no' that the model made is 747.

```
ModelMetricsBinomial: deeplearning
** Reported on test data. **

MSE: 0.14266037768403092
RMSE: 0.3777040874600524
LogLoss: 0.43562941966370783
Mean Per-Class Error: 0.2564678960453993
AUC: 0.820843042377889
AUCPR: 0.608844781991838
Gini: 0.6416860847557779

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.2539941055829337:
```

		0	1	Error	Rate
0	0	747.0	281.0	0.2733	(281.0/1028.0)
1	1	87.0	262.0	0.2493	(87.0/349.0)
2	Total	834.0	543.0	0.2672	(368.0/1377.0)

Figure 31: Deep Learning model confusion matrix

(Source: Code generated output screen)

Maximum Metrics: Maximum metrics at their respective thresholds

	metric	threshold	value	idx
0	max f1	0.253994	0.587444	233.0
1	max f2	0.131458	0.734476	305.0
2	max f0point5	0.523960	0.591133	118.0
3	max accuracy	0.554882	0.795933	103.0
4	max precision	0.800094	1.000000	0.0
5	max recall	0.026815	1.000000	397.0
6	max specificity	0.800094	1.000000	0.0
7	max absolute_mcc	0.413088	0.438631	165.0
8	max min_per_class_accuracy	0.264866	0.733463	229.0
9	max mean_per_class_accuracy	0.211660	0.743532	255.0
10	max tns	0.800094	1028.000000	0.0
11	max fns	0.800094	348.000000	0.0
12	max fps	0.025749	1028.000000	399.0
13	max tps	0.026815	349.000000	397.0
14	max tnr	0.800094	1.000000	0.0
15	max fnr	0.800094	0.997135	0.0
16	max fpr	0.025749	1.000000	399.0
17	max tpr	0.026815	1.000000	397.0

Figure 32: Deep Learning Model Performance

(Source: Code generated output screen)

Distributed Random Forest Trees is implemented on the data to predict the class of the data into 'yes' or 'no' for the Churn target field. Distributed Random Forest Trees generates a forest which is like a flowchart for the classification of the data. The tree is generated by training the model on the training set and then evaluated for its performance using the test set. The performance evaluation report of the model is shown below in figure 34. The accuracy of the model was recorded as 80 %, classification error that the model made was found to be 14%, AUC is recorded as 82.4%, the precision of the model was 78%, the recall for the model was found to be 100%, the f-measure also known as F1 Score was recorded at 73%, the model had 78% sensitivity specificity. Figure 33 below shows the confusion matrix of the model. The true ‘yes’ predictions in that the model are 218, and false ‘yes’ that the model made are 131. Similarly, the true ‘no’ predictions in that the model are 861, and false 'no' that the model made is 167.

```
ModelMetricsBinomial: drf
** Reported on test data. **

MSE: 0.14002188531283988
RMSE: 0.37419498301398946
LogLoss: 0.429125575924416
Mean Per-Class Error: 0.2575298518279019
AUC: 0.8249542885174986
AUCPR: 0.624097942139924
Gini: 0.6499085770349973

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.4091859961104405:
```

		0	1	Error	Rate
0	0	861.0	167.0	0.1625	(167.0/1028.0)
1	1	131.0	218.0	0.3754	(131.0/349.0)
2	Total	992.0	385.0	0.2164	(298.0/1377.0)

Figure 33: Distributed Random Forest Trees Confusion matrix

(Source: Code generated output screen)

Maximum Metrics: Maximum metrics at their respective thresholds

	metric	threshold	value	idx
0	max f1	0.409186	0.594005	144.0
1	max f2	0.149429	0.730890	293.0
2	max f0point5	0.507808	0.597510	90.0
3	max accuracy	0.507808	0.800290	90.0
4	max precision	0.787041	1.000000	0.0
5	max recall	0.011498	1.000000	397.0
6	max specificity	0.787041	1.000000	0.0
7	max absolute_mcc	0.421366	0.448480	136.0
8	max min_per_class_accuracy	0.306126	0.738327	199.0
9	max mean_per_class_accuracy	0.284671	0.742470	209.0
10	max tns	0.787041	1028.000000	0.0
11	max fns	0.787041	348.000000	0.0
12	max fps	0.009581	1028.000000	399.0
13	max tps	0.011498	349.000000	397.0
14	max tnr	0.787041	1.000000	0.0
15	max fnr	0.787041	0.997135	0.0
16	max fpr	0.009581	1.000000	399.0
17	max tpr	0.011498	1.000000	397.0

Figure 34: Distributed Random Forest Trees Performance

(Source: Code generated output screen)

Chapter 5: Conclusion and Future Work

This section presents the conclusions of the research and the future enhancements that can be made into the system to make it better.

5.1 Conclusion

In the literature review done on this topic of Customer Churn Prediction, it can be concluded that classical machine learning models are already in use prediction and are performing with good accuracy. But it is never enough as the organization are still losing customers.

In the RapidMiner implementation, we found that the Generalized Liner model has an accuracy score of 74.2% and has 373 as the number of true predictions for the class 'yes.' Gradient Boosted trees model the accuracy score of 73.8% and have 386 as the number of true prediction for the class 'yes'. Deep Learning models the accuracy score of 73.8% and has 362 as the number of true prediction for the class 'yes'. Logistic Regression model the accuracy score of 73.2% and has 382 as the number of true prediction for the class 'yes.' From this, we can conclude that the Generalized linear model has the best accuracy score of 74.2%. But the maximum number of 'yes' predictions are made by the gradient boosted tree, which is 386.

The H2O implementation we found that the stacked ensemble model where we stacked generalized linear model and gradient boosted machine has the accuracy score of 76%. Deep Learning models the accuracy score of 74% and Distributed random forest trees model the accuracy score of 74%. From this, we can conclude that the stack ensemble model has the best accuracy score of 76% as we used the best scoring model from RapidMiner as the base model to be stacked.

H2O stacked ensemble has a maximum accuracy score of 76%. For churn prediction, we can conclude that H2O stacked ensemble model has a higher accuracy score and also the maximum number of 'yes' predictions.

5.2 Future Work

For the future expansion of this churn prediction, we can run different models in stack ensembles in the H2O platform. The combinations that could be worked on are logistic

regression and generalized linear model, deep learning, and distributed random forest trees. The deep learning models like artificial neural networks and deep neural networks can be implemented in Python for future works.

The big data approach on these models can also be taken into consideration for future works. The big size of data will also aid in improving the model's accuracy.

As discussed, our approach is able to predict churn with 76% of accuracy; therefore, this can further be industrialized to provide recommendations for a churned customer to make him a loyal customer again. For example, a customer identified to be a churned customer, can be given attractive offers and discounts for the purpose of retention.

Reference List

A Hammoudeh, M. Fraihat, and M. Almomani, "Selective Ensemble Model for Telecom Churn Prediction," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 485-487, doi: 10.1109/JEEIT.2019.8717406.

Amin et al., "Customer churn prediction in the telecommunication sector using a rough set approach", *Neurocomputing*, vol. 237, pp. 242-254, 2017. Available: 10.1016/j.neucom.2016.12.009

Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty", *Journal of Business Research*, vol. 94, pp. 290-301, 2019. Available: 10.1016/j.jbusres.2018.03.003.

Dahiya and S. Bhatia, "Customer churn analysis in telecom industry", 2015 4th International Conference on Reliability, Infocom Technologies, and Optimization (ICRITO) (TrendsandFutureDirections),2015.Available:ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7359318 [Accessed 29 June 2020].

Debjyoti Das Adhikary, Dr. Deepak Gupta "Ensemble Learning Models for Churn Prediction", *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 2, pp. 159-164, 2019. Available: 10.35940/ijitee. b1091.1292s19 [Accessed 29 June 2020].

G. Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems", pp. pp. 3146-3154, 2017. [Accessed 29 June 2020].

G. Xia, H. Wang and Y. Jiang, "Application of customer churn prediction based on weighted selective ensembles", *2016 3rd International Conference on Systems and Informatics (ICSAI)*, pp. pp. 513-519, 2016. [Accessed 29 June 2020].

J. Xiao, L. Huang and L. Xie, "Cost-Sensitive Semi-Supervised Ensemble Model for Customer Churn Prediction," 2018 15th International Conference on Service Systems and Service Management (ICSSSM), Hangzhou, 2018, pp. 1-6, doi: 10.1109/ICSSSM.2018.8465099.

K. Mishra and R. Rani, "Churn prediction in telecommunication using machine learning," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 2252-2257, doi: 10.1109/ICECDS.2017.8389853.

Krawczyk, L. Minku, J. Gama, J. Stefanowski and M. Woźniak, "Ensemble learning for data stream analysis: A survey", *Information Fusion*, vol. 37, pp. 132-156, 2017. Available: 10.1016/j.inffus.2017.02.004.

Lessmann, Stefan & Coussement, Kristof & De Bock, Koen & Haupt, Johannes. (2018). Targeting Customers for Profit: An Ensemble Learning Framework to Support Marketing Decision Making. SSRN Electronic Journal. 10.2139/ssrn.3130661.

M. Saghir, Z. Bibi, S. Bashir and F. H. Khan, "Churn Prediction using Neural Network based Individual and Ensemble Models," 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 2019, pp. 634-639, doi: 10.1109/IBCAST.2019.8667113.

Mishra and U. Reddy, "A comparative study of customer churn prediction in telecom industry using ensemble-based classifiers", *International Conference on Inventive Computing and Informatics (ICICI)*, pp. pp. 721-725, 2017. [Accessed 29 June 2020].

Mumin Yildiz, Songul Varli **"**Customer churn prediction in telecommunication", 2020. [Online].Available:https://www.researchgate.net/publication/282981765_Customer_churn_prediction_in_telecommunication. [Accessed: 29- Jun- 2020].

Pretam Jayaswal, Bakshi Rohit Prasad "Approach for Efficient Churn Prediction in Telecom Industry", *Article.nadiapub.com*, 2016. [Online]. Available: http://article.nadiapub.com/IJDTA/vol9_no8/21.pdf. [Accessed: 29- Jun- 2020].

R. Andrews, "CHURN PREDICTION IN TELECOM SECTOR USING MACHINE LEARNING", *International Journal of Information Systems and Computer Sciences*, vol. 8, no. 2, pp. 132-134, 2019. Available: 10.30534/ijiscs/2019/31822019.

U. Ahmed, A. Khan, S. Khan, A. Basit, I. Haq and Y. Lee, "Transfer learning and meta classification based deep churn prediction system for telecom industry", 2019. [Accessed 29 June 2020].

Ullah, B. Raza, A. Malik, M. Imran, S. Islam and S. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector", *IEEE Access*, vol. 7, pp. 60134-60149, 2019. Available: 10.1109/access.2019.2914999.

V. Umayaparvathi and K. Iyakutti, "A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics", *International Research Journal of Engineering and Technology (IRJET)*, vol. 3, no. 04, 2016. [Accessed 29 June 2020].

W. Bi, M. Cai, M. Liu and G. Li, "A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn", *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1270-1281, 2016. Available: 10.1109/tii.2016.2547584.

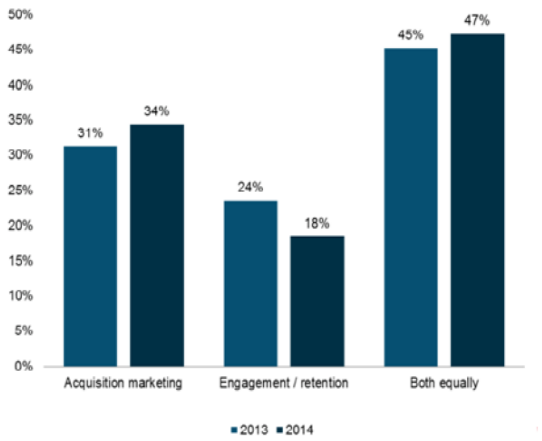
Y. Khan, S. Shafiq, A. Naeem, S. Ahmed, N. Safwan and S. Hussain, "Customers Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry", *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 9, 2019. Available: 10.14569/ijacsa.2019.0100918.

Yu, R. Zhou, L. Tang and R. Chen, "A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data", *Applied Soft Computing*, vol. 69, pp. 192-202, 2018. Available: 10.1016/j.asoc.2018.04.049.

Appendices

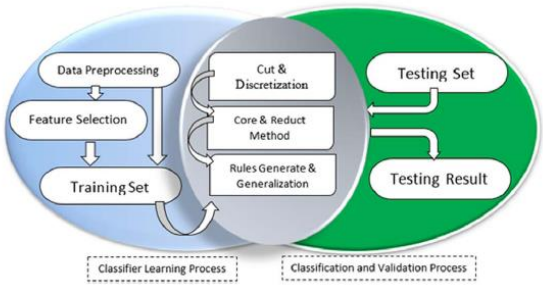
This section will lead you through the material of the Report, Artifacts for the dissertation project entitled "A Study of Ensemble Machine Learning to Enhance Consumer Churn Telecommunication Prediction."

Appendix 1: Customer Acquisition and Retention



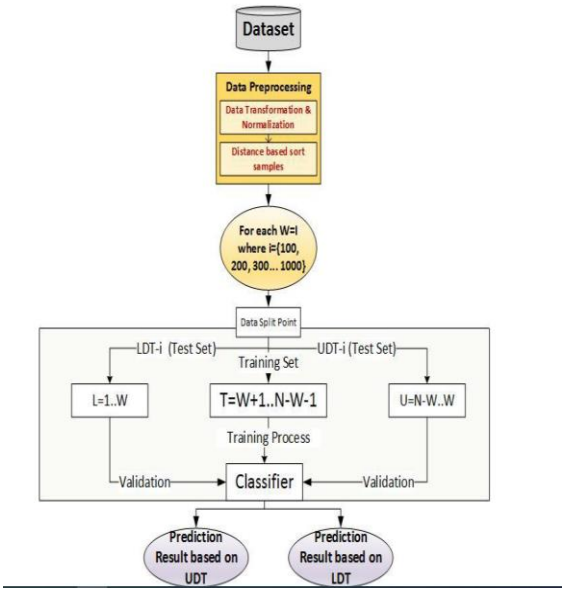
(Source: article.nadiapub.com, 2020)

Appendix 2: RST based classification Based



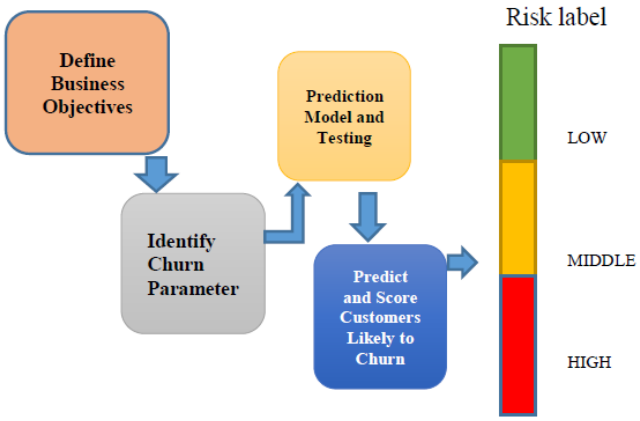
(Source: Amin et al. 2017)

Appendix 3: Visualization of the benchmark framework



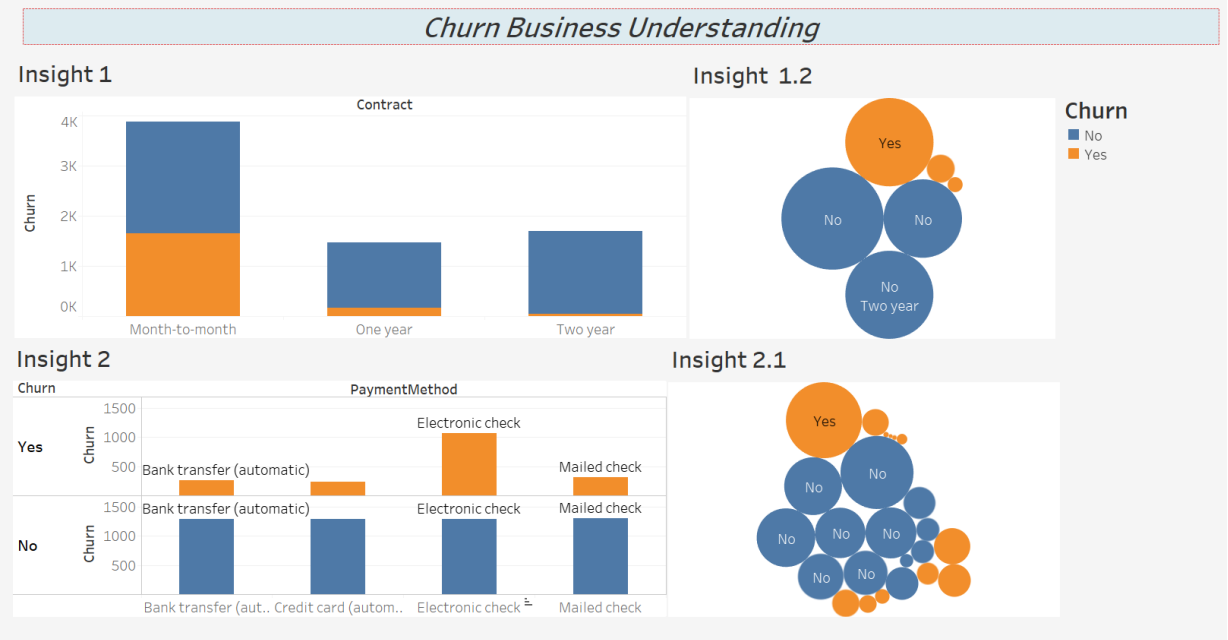
(Source: Amin *et al.* 2019)

Appendix 4: Churn Prediction Model



(Source: Mishra and Reddy, 2017)

Appendix 5: Tableau Dashboard



(Source: Generated by Tableau)

Appendix 6: Artifacts

1. Datasets

- Original_Data_WA_Fn-UseC_-Telco-Customer: Original data
- Telco_churn_balanced_data: Processed and balanced data
- 80_percent_training_data: Training data
- 20_percent_testing_data: Testing data
- Telco-Customer-Churn_Tableau: CSV for tableau

2. Tableau_Business_Understanding

- Churn_Business_Understanding: tableau project file
- Tableau_screenshots: Word file containing tableau screen shots

3. RapidMiner

- Rmp_Processes_files: Folder contains all the algorithms RMP files
- Auto_model_process: Folder contains automodel process for GLM and GBT

4. Python Code

- Churn_Prediction_Python_Ensemble_H2O: Code for Python