Using Machine Learning to predict Customer Attrition in the Telecom Industry

Anish Mahapatra
Student ID 944563

Under the supervision of
Karthick Neelamohan

Research Proposal
Master of Science in Data Science
Liverpool John Moores University

JANUARY 2021

Contents

Abstract 4

1. Background 5

1.1 The need for Customer Churn Analysis 5
1.2 Flagging customers and retention policies 6
3. Problem Statement 7

4. Related Works 7

4.1 Sampling, balancing techniques and pre-processing 7
4.2 Feature engineering and selection of attributes 8
4.3 Ensemble methods 8
4.4 Machine learning techniques 9
5. Aim and Objectives 10

6. Significance of the research 11

7. Scope of the study 11

8. Research Methodology 12

8.1 Business Understanding 12
8.2 Data Understanding 12
8.3 Data Preparation 13
8.3.1 Data Cleaning 13
8.3.2 Feature Engineering 14
8.3.3 Data Formatting 14
8.4 Model Building 15
8.4.1 Model Selection Techniques 15
8.4.2 Test Designing 15
8.4.3 Model Iterations 15
8.4.4 Model Assessment 15
8.5 Model Evaluation 17
8.5.1 Model Evaluation 17
8.5.2 Process Review 17
8.5.3 Determine Next Steps 17
8.6 Model Deployment 17

8.6.1 Plan for Deployment 17
8.6.2 Monitoring and Maintenance 18
8.6.3 Reporting Results 18
8.6.4 Final Review 18
9. Required Resources 19
9.1 Hardware Requirements for the research 19
9.2 Software Requirements for the research 19
10. Research Plan 19
10.1 Gantt Chart for Research 19
11. Risk and Contingency Plan 20
References 21

Abstract

With the advent of increasing competition in various market segments, companies must retain customers to maximize profits. Customer retention policies can affect the annual turnover drastically depending on the rate of churn. The cost of customer churn to the Telecom industry is about $10 billion per year. Studies show that customer acquisition cost is 5-10 times higher than the price of customer retention. Companies, on average, can lose 10-30% of their customer annually. Developing processes and efficient consumer-centric policies to reduce customer churn can reduce spend on customer relations. For this, one would need to understand and track customer behaviour to understand the indicators that make a customer likely to churn.

Datasets for customer churn are quite large and saved in large data warehouses where many features are present. Not all attributes are significant for churn prediction. Hence, feature engineering requires not only computation but a substantial amount of time as well. Through this paper, we will find and the features that that will be significant for churn prediction. The aim is to predict churn accurately and showcase the variation in performance of various algorithms.

1. Background

With the increase in the number of options consumers in the telecom space have with the advent of the Digital Age, for a company to be successful, it is vital to keep costs low and profits high. One of the most effective ways to do this is to retain the existing customer base and focus the rest of the budget on acquiring new customers.

The retention of the existing customer base in a focused and systemic manner is to be done, or its bottom line can be affected. A targeted way to approach the end goal of customer retention is to flag customers that have a high probability to churn. Based on customer behaviour and attributes, if we can flag the customers that are likely to churn, we can run targeted campaigns to retain customers.

1.1 The need for Customer Churn Analysis

The ability to retain customers showcases the company's ability to run the business. With the digital age now, where everything is online, any business needs to understand customer behaviour and mentality. The cost of customer churn in the Telecom Industry is approximately $10 billion annually [(Castanedo et al., 2014)]. Customer acquisition costs are higher than customer retention by 700%; if we were to increase customer retention rates by just 5%, profits could see an increase from 25% to even 95% [(Hadden et al., 2006)]. For a company to be profitable, it is thus essential to take pre-emptive action to retain customers that may churn. Churn in telecom companies is defined as the customers who stop using their specific services and plans for long periods.

In this post-pandemic age, where virtual presence via calls and mobile data is the top priority, customers streamline their monthly expenditure. Competitors are employing strategies such as offering low prices or value-add services to get consumers to switch. After acquiring a significant customer base, the companies monetize their customer base and turn a quick profit. Companies that can identify the bracket of people that are likely to leave and run targeted campaigns to showcase more value in their current offerings at a minimal budget are the companies that will be successful in the long run.

1.2 Flagging customers and retention policies

As service providers contend for a customer's rights, customers are free to choose a service-provider from an ever-increasing set of corporations based on customer need. This increase in competition has led customers to expect tailor-made products at a fraction of the price [(Kuo et al., 2009)]. Churned customers those customers that move from one service provider to another [(Ahmad et al., n.d.) [(Andrews, 2019)]. Churn can be due to the non-satisfaction of current services, better offerings from other service providers and even lifestyle changes. Companies use retention strategies [(Jahromi et al., 2014)] to maximize customer lifetime value by increasing the associated tenure. For telecom companies to reduce churn, it is vital to predicting specific metrics such as the high-risk customers, estimated time to attrite and likelihood to churn.

The learnings from multiple such exercises have been introduced as deployable machine learning algorithms that have been iterated over and refined based on the evolving need to flag consumers more accurately. The selection of techniques to employ will depend on the model's performance on the selected dataset, be it meta-heuristic, data mining, machine learning or even deep learning techniques. In the customer's behaviour patterns, there is likely to be a few significant indicators as to why the customer is willing to take the active step of moving across service providers. We shall identify the attributes that can indicate churn in our methodology through this research.

3. Problem Statement

With the customer data acquired from the telecom company, we will accurately flag customers' bracket likely to churn. This research will help telecom companies leverage their database to predict and actively target campaigns to customers that might churn. The methodology can be a set standard in the industry where multiple machine learning algorithms can run on a newer dataset, we can monitor the accuracy of the model, and customers can be appropriately targeted.

The recommended model's primary users will be Telecom companies that wish to reduce customer attrition by leveraging what Data Science offers. Given that the model predicts customers that will churn accurately, this can be done with limited hardware and regular cadence.

4. Related Works

The utilization of meta-heuristic models and machine learning was done to get accurate predictions on the telecom datasets. Some literature focused on enhancing the data itself via coherent pre-processing and efficient feature engineering techniques [(Ahmed and Maheswari, 2017a)]. In contrast, the others focused on using more complicated algorithms such as Artificial Neural Networks, Support Vector Machine to get higher accuracy. With most papers focused on either data mining or modelling, some research employed novel techniques [(Ahmed and Maheswari, 2017b)] for prediction.

4.1 Sampling, balancing techniques and pre-processing

The literature that uses balancing techniques such as under-sampling, random sampling, gradient boosting and weighted random forests tend to have a higher accuracy of attrition prediction [(Burez and Van den Poel, 2009)]. Selected methods also decrease the strength of the model, such as random sampling. Some methods combine various balancing techniques such as Weighted Random Forest and sampling, including under-sampling and Synthetic Minority Oversampling Technique. The combination of specific sampling processes improves the value of F-measure and prediction strength [(Effendy et al., 2014)]. Using only under-sampling alone is not significant.

Boosting via AdaBoost or other boosting techniques was also proposed to improve customer churn prediction accuracy. Boosting combined with a basis learner such as logistic regression can help enhance model performance [(Lu et al., 2014)]. The application of a combination of Synthetic Minority Oversampling Technique and AdaBoost has been employed to process the imbalanced data. Post the Synthetic Minority Oversampling Technique on the imbalanced data, AdaBoost is used on the balanced data to predict attrition.

4.2 Feature engineering and selection of attributes

Feature selection using Support Vector Machine based on the profit model selects the top features based on the profit model. The focus is on selecting the appropriate kernel functions to perform customer attrition better.

4.3 Ensemble methods

Post combining the social and local features of the dataset, an ensemble model was designed. Data from a telecom operator that was used for testing. The model contained a spreading activation algorithm that spread the social and local variables to combine these features. The prediction did improve in the ensemble approach compared to the individual models [(Backiel et al., 2015)]. However, excluding non-customer nodes in the call graph leads to a reduction in the overall prediction of churn's effectiveness. The model's evaluation proves that the customer attrition prediction in terms of AUC, lift and Area under the curve is enhanced.

The next ensemble model proposed used consumer utilization of services and other behaviour patterns to predict churn. A binary classifier is built for attrition using decision trees and its ensembles, Gradient Boosted Trees and Random forest. Analyzing the research results showcases that the ensemble has better sensitivity and accuracy scores, especially for the improvement in the residual feedback[(Jayaswal et al., 2016)]. This approach was not tested on real-world streaming data, leading to limited reliability on the prediction model.

4.4 Machine learning techniques

Post-pre-processing of data with Principle Component Analysis, multiple machine learning models were applied on customer data to determine the customers that will churn. The models of neural networks, support vector machines, multi-layer perceptron and Bayesian networks were applied to the data. Support vector machine provides higher accuracy than the Bayesian network and Multi-Layer Perceptron [(Brandusoiu et al., 2016)]. The robustness of the model is under consideration as it employs individual models and not an ensemble model.

An improved balanced random forest was proposed to obtain more accurate customer churn numbers. This approach combines cost-sensitive attributes and sampling methods along with Random forests to predict attrition [(Xie et al., 2009)].

A combination of logistic regression and decision tree model was proposed to perform customer churn. To understand the impact of each feature on attrition, Logistic Regression was used, and the Decision Tree provides a visual representation of the strategies being employed for the same. This also reduces the time to predict churn and results in a restricted number of classes.

5. Aim and Objectives

The paper aims to develop a trustworthy and interpretable model that will predict the customers that will churn from a Telecom Company based on historical customer telecom data. The identification of the customers that churn will aid telecom companies in significantly reducing expenditure on customer relations.

The objectives of the research are based on the above aim and are as follows:

- To analyze the relationship and visualize patterns of customer behaviour to indicate to the telecom company if a customer is going to churn
- To suggest suitable feature engineering steps to extract the most value from the data including picking the most significant features
- To find appropriate balancing techniques to enhance the model performance on the dataset
- To compare the classification or predictive models to identify the most accurate model to determine the customers that will churn
- To understand the factors and behaviour that leads to customer attrition in the telecom industry
- To evaluate the performance of the models to identify the appropriate models

6. Significance of the research

The research is contributing to the explanation and interpretation of the prediction of various predictive models to support decision making and increase the bottom line of the company by flagging customers that are going to churn. This will help customer allocate budget and time to the customers that are likely to churn by running targeted campaigns. The sales team will be able to offer value-adds to the high-risk and high-value customers. This can help the company document the pain points faced by its customers and can ultimately help aid in fundamental policy changes that can increase the overall profit.

7. Scope of the study

Due to the limitation of the time frame in this research, the scope of the research will be limited to the below points:

- The data for the study has directly been obtained from the authorized source, and data validation will not be part of this research
- The research will include the development and evaluation of various machine learning algorithms. The latest algorithms such as Neural Networks and Deep learning will not be considered as a part of this study due to a lack of resources and time
- The study will limit the use of classification algorithms such as logistic regression, decision tree, K-nearest Neighbour as a part of interpretable models, whereas random forest, support vector machine, gradient boosting and XGBoost will be leveraged as black-box models for this study
- The research will be limited to the comparison of the latest technologies around global and local model agnostic methods

8. Research Methodology

8.1 Business Understanding

In this paper, we were able to identify that the telecom industry is an extremely competitive industry where customers have the free will to move across companies if they believe they are getting more value with another service provider. We also noted that based on the customer's behaviour patterns, we would have indicators to note if a customer might churn or not. Since the cost of retention is much higher than customer acquisition, it is vital to the company's survival to identify the customers likely to churn and run campaigns to retain the existing customer base. It was also observed that a reduction of customer attrition of 5% could lead to profit margins increasing from 25% to 95%[(Hadden et al., 2006)]. In the telecom industry where the approximated annual cost of customer attrition is $ 10 billion annually [(Castanedo et al., 2014)], and 30% customers churn on average, there is a substantial need to perform active targeting to retain the customer base.

8.2 Data Understanding

There are various data sources used to predict churn in the telecom industry through the literature survey. In this research, we shall be using the IBM Watson Telecom churn data found in the Kaggle website [(Kaggle)]. The telecom churn data consists of 7043 rows and 21 attributes at a customer id level. The data has a combination of numerical and categorical variables that can be used as feature variables to predict the target variable churn. Churn is indicated within the dataset as a "Yes" or a "No" indicating if a customer has churned or not churned respectively. This data presented is for the last month based on which predictions are to be made.

The given data consists of multiple factors about the customers regarding lifestyle, behaviour in a Yes or No format that can be leveraged post-processing. It is presented in a .csv format with customer attributes information as metadata.

The information obtained from the data can be broken down into four broad categories and is as follows [(Ebrah and Elnasir, 2019)][(Kaggle)]:

- Services that the customer may be using such as streaming movies and tv, technical support, device protection, online backup and service, broadband services
- Account Information of the customer such as customer tenure, total costing, monthly charges, paperless billing, payment method
- Demographic information such as age, gender, information about dependents and partners

8.3 Data Preparation

We shall carefully analyze the data, understand the data patterns through visualizations and proceed with the following steps in detail.

8.3.1 Data Cleaning

Data cleaning for the telecom churn dataset will occur by first doing a sense check if the data. Once it is verified that the data types of the data are as expected, we will check on the shape of the data to make sure the number of rows and columns is consistent with our expectations. We will then focus on the columns that have at least one missing value. Once we understand the attributes to consider, we will understand the percentage of missing values column-wise. This will help us to decide the strategies to take for the next steps. Post missing value analysis; we will decide if we can proceed with all the columns to the next step if we have to drop columns based on missing value percentage or employ methods such as mean imputation, mode imputation, deletion of rows and iterative imputation.

Looking at the percentage of missing values for each attribute after missing value analysis will help us understand the base dataset that we will be using when we go to the next step of feature engineering.

We will also perform outlier analysis and understand the skewness of the data to understand the feature's impact on customer churn. Post the understanding of each of the features' distribution; we will proceed to perform univariate analysis. This will help us understand and map out the inherent properties and distributions of each attribute. The bivariate analysis will then be performed on the data, ultimately followed by multivariate analysis to understand the features' direct and latent impact on the target variable.

8.3.2 Feature Engineering

Based on the cleaned dataset, we will not decide the next steps be taken to be able to extract the most value from the dataset. We can perform steps such as one-hot encoding on the features that of type object. Besides this, we shall also derive features from the existing dataset and feature engineer newer attributes. Based on the understanding of telecom's business, we will also apply business rules that make sense to the business and try to derive new features. Performing efficient feature engineering here will save us the hassle of running complicated models to get an accurate prediction. This will make the machine learning pipeline easier to deploy, thus saving the business expenditure on hardware.

Data visualization here will play a crucial part here to be able to draw insights that might help to be able to derive more from the data. Mapping out and understanding the relationship of each numerical and categorical variable with churn will help us start identifying the attributes that might have a high impact on customer churn. We shall perform multicollinearity and variance inflation factor tests to understand the data's inherent properties to understand the significant features to select for modelling. We will also look at the correlation scores for the numerical variables to identify the features that have a high positive or negative correlation with the target variable. We will also perform categorical analysis on the variables of type object to deep-drive into implicit and latent connections within the data.

8.3.3 Data Formatting

Based on the models we will apply, we will ensure the now cleaned data with the new features is formatted accordingly. This will help specific models converge at a faster rate as compared to if the data was not formatted. We can also apply feature selection techniques to understand the most significant features from the dataset.

8.4 Model Building

We shall now proceed to model building to choose the models that we would implement post the data cleaning, feature engineering and data formatting steps.

8.4.1 Model Selection Techniques

We shall now proceed to select the models that we will be working with to efficiently and accurately predict customer churn. From the literature, it has been seen that the supervised classifier models have given us good results. We shall use logistic regression, decision trees, Naïve Bayes, random forest, support vector machine and understand how the algorithms perform. Post analysis of the individual algorithms, we shall also attempt ensemble models with boosting such as XGBoost and Light GBM.

8.4.2 Test Designing

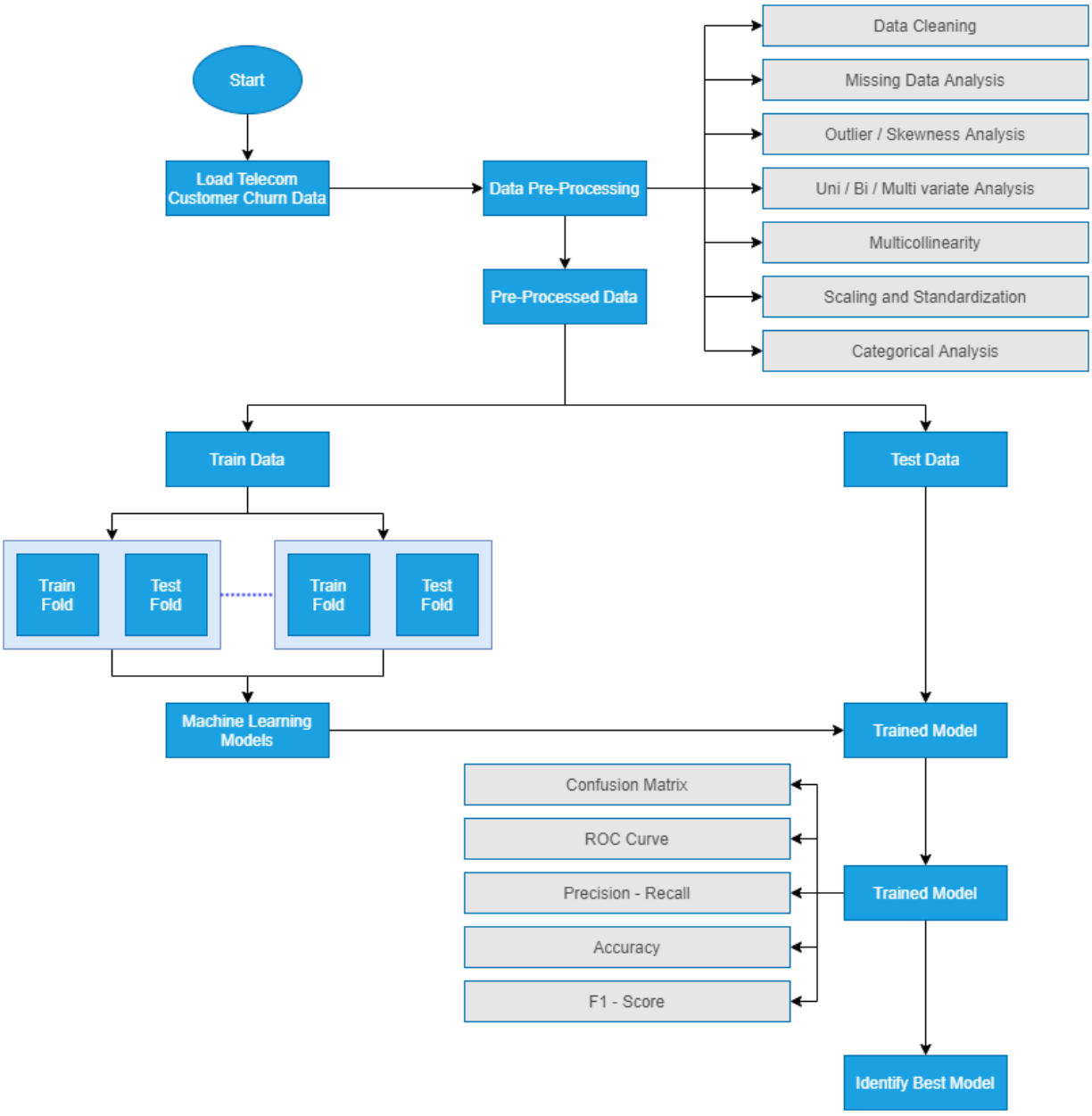
Another vital step to model building is to decide the train and test split strategically. If there were a larger dataset, we could have opted to go for a validation dataset as well. We will go for an 80-20 train-test split for the models. For the top-performing models with this design, we shall also attempt a 90-10 split as this was recommended in the literature review for a few research papers.

8.4.3 Model Iterations

After the model, as mentioned earlier, building steps are performed, we shall now proceed to perform more iterations on the models correspondingly analyzing model performance with each iteration. This can include monitoring p-values, the number of features, model performance, variance inflation factor scores which would differ across models. The top selected models will now be the challenger models based on which the best model will be decided. On the given models, we will perform hyperparameter tuning using previous learnings and methods such as Grid Search, Random Search, and Bayesian optimization depending on the model considered.

8.4.4 Model Assessment

For any models to be used by the business, model assessment is a critical part f the process. As we develop models from the eyes of a Data Scientist up until this point, for the business to leverage the model, we will need to take steps to ensure that the predictions are as expected.



Model Building Process by Author via draw.io

Model interpretability is vital to the functioning of the business as they would like to understand the customers that are likely to churn and gain insights as to why. Therefore, we are in the model assessment stage; we will need to focus on actionable insights and provide the business with the customer behaviour patterns linked with the high likelihood of churn.

8.5 Model Evaluation

We have now settled on the best model that we would like to showcase. This is the model on which extensive feature engineering has been carried out, and from a wide range of models, we have chosen the best. We will follow the below-mentioned steps to perform the model evaluation.

8.5.1 Model Evaluation

We will now proceed to compare the model results obtained with the other literature we have previously surveyed. Using the same metrics of accuracy, F-Score, the area under the curve, we will compare the performance of the new ensemble or individual models to the models' performance in the reviewed literature in the field. Once we evaluate the results and see if they are satisfactory, we will proceed to the next steps. Else, we shall analyze the results if they are not satisfactory and proceed to revaluate our approach to improve iteratively.

8.5.2 Process Review

We will list the final process post the different iterations we have carried out and carefully review the process. As compared to the other research done in this field, we will analyze if there are any potential mosses, flaws in approaches an address them.

8.5.3 Determine Next Steps

Based on the process review carried out in the above step, we will decide if we would like to finish our research project and move on to the next steps. If not, we shall initiate further iterations and refine the model. This is an essential step and will be based on the comparative analysis we will perform to benchmark our model.

8.6 Model Deployment

We will now decide the next steps for the business use that our model evaluation is satisfactorily completed.

8.6.1 Plan for Deployment

The model is to be utilized by telecom companies to reduce the rate of churn by targeting customers at a high likelihood of churn. There are certain factors to consider here based on which the

company's return on investment can be maximized. 80% of revenue is generated by 20% of the customer base [(Rajagopal, 2011)]. Based on the allocated budget for customer retention, we should filter out high-value customers with a high customer lifetime value and target those that are the most likely to churn. Allocating too much time to customers who are not generating as much revenue can be prioritized lower.

8.6.2 Monitoring and Maintenance

A cost-benefit analysis will be carried out to understand the actual cost of running the model in real-time. There might be potential data anomalies while new data comes in. Robust machine learning pipelines along with teams to monitor the same will be deployed. This will help us monitor the results and understand how we can make the deployment more efficient.

8.6.3 Reporting Results

For a machine learning model to improve with time, it is essential to create a feedback loop. Documentation of the research carried out, the results, and loopholes must be carefully documented to improve the model in the next iteration. If a similar accuracy can be obtained with lesser processing, this will also help the company save costs in operationalization expenditure.

8.6.4 Final Review

We will contemplate in the final review what are the things done right and what went wrong. There will be learnings from the entire process that we shall document and use in our next steps. We should also learn what was done well and what could have been avoided.

9. Required Resources

Following are the required hardware and software requirements to successfully and smoothly run the models.

9.1 Hardware Requirements for the research

Based on the defined scope of the proposed thesis, the following are the required resources:

The minimum hardware requirements for this project are:

RAM: Minimum 8 GB (16 GB recommended for optimum performance)

Disk space: Minimum of 4GB of free space needs to be allocated (Depends on the model iterations)

9.2 Software Requirements for the research

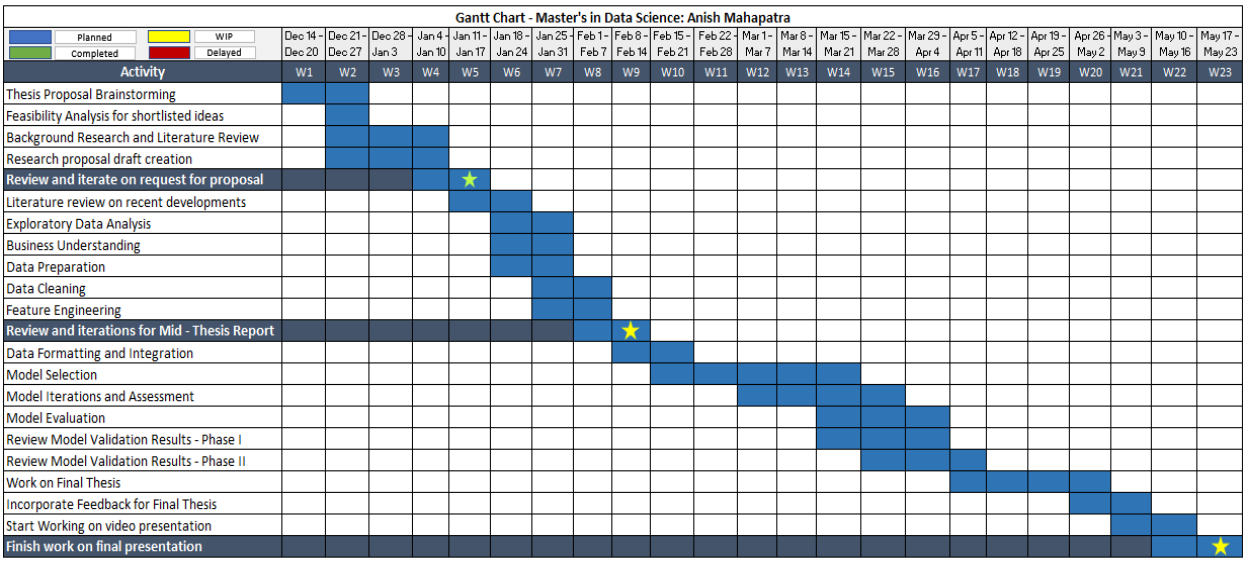
Software	Minimum Version
Python	>= 3.5
Jupyter Notebook	>= 6.0
Excel	>= 2007

10. Research Plan

The following GANTT chart proposes the timeline for the research and implementation of the project.

10.1 Gantt Chart for Research

Based on the complexity of the different phases, the timelines are subject to minor adjustments. Regardless, the candidate shall pledge to stick to the timeline as closely as possible.



11. Risk and Contingency Plan

Project Phase	Risk	Contingency Plan
Exploratory Data Analysis	Did not receive accurate data and the data is corrupt	Leverage a copy of the data stored on Google Drive for the same
Data Transformation	Jupyter Notebook on the local computer is corrupt	Use an online iPython service like Google Colab to perform the study
	Data transformation is time-consuming and does not result in the format	Revisit the exploratory data analysis phase and try alternate methods
Model Building	Might not be able to implement all the modelling techniques researched during the research space	Scale back by filtering modelling techniques prioritized on novelty and usage

References

1. Ahmad, A.K., Jafar, A. and Aljoumaa, K., (n.d.) Customer churn prediction in telecom using machine learning in big data platform. [online] Available at: https://doi.org/10.1186/s40537-019-0191-6.
2. Ahmed, A.A. and Maheswari, D., (2017a) Methods For Customer Retention In Telecom Industries. *2017 International Conference on Advanced Computing and Communication Systems*.
3. Ahmed, A.A.Q. and Maheswari, D., (2017b) Churn prediction on huge telecom data using hybrid firefly based classification Churn prediction on huge telecom data. *Egyptian Informatics Journal*, [online] 183, pp.215–220. Available at: http://dx.doi.org/10.1016/j.eij.2017.02.002 [Accessed 15 Jan. 2021].
4. Andrews, R., (2019) Churn Prediction in Telecom Sector Using Machine Learning. *International Journal of Information Systems and Computer Sciences*, 82, pp.132–134.
5. Anon (2021) *Kaggle: Your Machine Learning and Data Science Community*. [online] Available at: https://www.kaggle.com/ [Accessed 17 Jan. 2021].
6. Backiel, A., Verbinnen, Y., Baesens, B. and Claeskens, G., (2015) Combining local and social network classifiers to improve churn prediction. In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*. [online] New York, NY, USA: Association for Computing Machinery, Inc, pp.651–658. Available at: https://dl.acm.org/doi/10.1145/2808797.2808850 [Accessed 17 Jan. 2021].
7. Brandusoiu, I., Toderean, G. and Beleiu, H., (2016) Methods for churn prediction in the pre-paid mobile telecommunications industry. In: *IEEE International Conference on Communications*. Institute of Electrical and Electronics Engineers Inc., pp.97–100.
8. Burez, J. and Van den Poel, D., (2009) Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 363 PART 1, pp.4626–4636.
9. Castanedo, F., Valverde, G., Zaratiegui, J. and Vazquez, A., (2014) Using Deep Learning to Predict Customer Churn in a Mobile Telecommunication Network Federico. pp.1–8.
10. Ebrah, K. and Elnasir, S., (2019) Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms. *11Journal of Computer and Communications*,

[online] ``23df, pp.33–53. Available at: https://doi.org/10.4236/jcc.2019.711003 [Accessed 10 Jan. 2021].

11. Effendy, V., Adiwijaya, K. and Baizal, Z.K.A., (2014) Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. [online] Available at: https://www.researchgate.net/publication/266387792 [Accessed 17 Jan. 2021].
12. Hadden, J., Tiwari, A., Roy, R. and Ruta, D., (2006) Churn Prediction: Does Technology Matter. *International Journal of Intelligent Technology*, 1, pp.104–110.
13. Jahromi, A.T., Stakhovych, S. and Ewing, M., (2014) Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, [online] 437, pp.1258–1268. Available at: https://research.monash.edu/en/publications/managing-b2b-customer-churn-retention-and-profitability [Accessed 16 Jan. 2021].
14. Jayaswal, P., Prasad, B.R. and Agarwal, S., (2016) An Ensemble Approach for Efficient Churn Prediction in Telecom Industry. *International Journal of Database Theory and Application*, [online] 98, pp.211–232. Available at: http://dx.doi.org/10.14257/ijdta.2016.9.8.21 [Accessed 17 Jan. 2021].
15. Kuo, Y.-F., Wu, C.-M. and Deng, W.-J., (2009) The relationships among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. *Computers in Human Behavior*, 25, pp.887–896.
16. Lu, N., Lin, H., Lu, J. and Zhang, G., (2014) A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, 102, pp.1659–1665.
17. Rajagopal, D.S., (2011) Customer Data Clustering using Data Mining Technique. *International Journal of Database Management Systems*, [online] 34. Available at: http://arxiv.org/abs/1112.2663 [Accessed 17 Jan. 2021].
18. Xie, Y., Li, X., Ngai, E.W.T. and Ying, W., (2009) Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 363 PART 1, pp.5445–5449.