University of Hertfordshire

School of Physics, Engineering and Computer Science

MSc. Advanced Computer Science with Advanced Research

MSc. Computer Science Project

(7COM1039)

Final Progress Report

An Evaluation of Supervised Learning Models in the Prediction and Analysis of Customer Churn

Name: Simpson Chiwashira

SRN: 21017447

Level: 7

Academic Year: 2023-2024

Supervisor: John Sapsford

**MSc Final Project Declaration**

I certify that the work submitted is my own and that any material derived or quoted from the published or unpublished work of other persons has been duly acknowledged in the bibliography and references.

I certify that I did not use any human subjects in my MSc Project.

**Student Full Name:** SIMPSON CHIWASHIRA

**Student Registration Number:** 21017447

**Signed:** ………S. Chiwashira………………………………………

**Date:** …………09/11/2023…………………………………………

# **Abstract**

This study investigates the effectiveness of supervised learning models in predicting and analysing customer churn within the Telecommunications and Banking Marketing sectors. Utilising sector-specific datasets, the research explores imbalanced data handling techniques, employing both Random Undersampling and SMOTE to achieve balanced class representation. Evaluation of Support Vector Machines, Artificial Neural Networks, and Random Forests, post-hyperparameter tuning, identifies Random Forests as the optimal classifier across all datasets. The study emphasises the impact of resampling techniques on classifier performance pre- and post-model refinement, recommending the utilisation of Random Forests for predictive accuracy in such sectors.

# **Acknowledgements**

I would like to thank the Lord Almighty who has made this project a success. I would also like to extend my gratitude to the teaching team for guidance throughout the last two years at the university. My supervisor John Sapsford, you were amazing from the first day of this journey, Thank you. Lastly my family and fellow peers, you gave me the support I needed.

**Table of Contents**

[Abstract 3](#_Toc152615344)

[Acknowledgements 4](#_Toc152615345)

[1. Introduction 7](#_Toc152615346)

[1.1 Background of Research Area 7](#_Toc152615347)

[1.2. Project’s Aims and Objectives 8](#_Toc152615348)

[1.3. Project’s Advanced Aims 8](#_Toc152615349)

[1.4. Project Management 8](#_Toc152615350)

[1.5. Resources 9](#_Toc152615351)

[1.6. Legal, Ethical, Professional and Social Issues 9](#_Toc152615352)

[2. Literature Review 10](#_Toc152615353)

[2.1 Review Protocol 10](#_Toc152615354)

[2.2. Supervised Learning Models 10](#_Toc152615355)

[2.3. Algorithm Selection 11](#_Toc152615356)

[3. Methodology 16](#_Toc152615357)

[3.1. Project Plan 16](#_Toc152615358)

[3.2. Data Selection 16](#_Toc152615359)

[3.3. Datasets 16](#_Toc152615360)

[3.3.1. Banking and Marketing Dataset 16](#_Toc152615361)

[3.3.2. Telco Dataset 19](#_Toc152615362)

[3.4. Data Pre-Processing 21](#_Toc152615363)

[3.5. Data Imbalance 23](#_Toc152615364)

[3.6. Data Preparation and Transformation 24](#_Toc152615365)

[4. Model Implementation 25](#_Toc152615366)

[4.1. Model Architecture 25](#_Toc152615367)

[4.2. Model Training 26](#_Toc152615368)

[4.3. SVM – Implementation and Hyperparameter Tuning 26](#_Toc152615369)

[4.4. RF – Implementation and Hyper-Parameter Tuning 28](#_Toc152615370)

[4.5. ANN – Implementation and hyperparameter tuning. 29](#_Toc152615371)

[5. Evaluation and Analysis 30](#_Toc152615372)

[5.1 Evaluation Metrics 30](#_Toc152615373)

[5.2. The Confusion Matrix 30](#_Toc152615374)

[5.3. Evaluation of the Telco Dataset 30](#_Toc152615375)

[5.4. Evaluation of the Bank Marketing Dataset 31](#_Toc152615376)

[5.5. Feature Importance and Potential Predictors 32](#_Toc152615377)

[5.6. Research Question and Hypothesis 34](#_Toc152615378)

[6. Conclusions 35](#_Toc152615379)

[7. Limitations and Further Work 35](#_Toc152615380)

[8. Reflection 35](#_Toc152615381)

[Appendices 39](#_Toc152615382)

Table of Figures

[Figure 1: Decision Boundary for SVM Classifier (Pedregosa et al, 2011) 12](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540450)

[Figure 2: ANN Structure - Bekesiene, Smaliukiene and Vacantiene (2021) 14](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540451)

[Figure 3: Output of the Banking Dataframe 17](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540452)

[Figure 4: Pie chart showing data imbalance for Bank and Marketing Dataset. 18](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540453)

[Figure 5: Columns and data types for the Telco Dataset. 19](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540454)

[Figure 6: Output showing head of Telco Dataframe. 20](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540455)

[Figure 7: Pie plot showing the Telco data imbalance. 20](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540456)

[Figure 8: Output for loading the raw CSV file. 21](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540457)

[Figure 9: Showing the MS Excel view of the raw CSV file. 21](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540458)

[Figure 10: Error during the application of SMOTE 23](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540459)

[Figure 11: Iterative flowchart of the model implementation. 25](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540460)

[Figure 12: Application of SMOTE on Telco Dataset. 26](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540461)

[Figure 13: Application of the RandomUnderSampler on Telco Training Data. 26](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540462)

[Figure 14: Confusion matrix showing the performance. 30](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540463)

[Figure 15: Performance of Classifiers on the Telco Dataset. 31](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540464)

[Figure 16: Performance of Classifiers on the Bank Marketing Dataset. 31](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540465)

[Figure 17: Feature Importance of the RF (Telco Dataset). 32](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540466)

[Figure 18: Correlation Heatmap for Telco Dataset. 33](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540467)

[Figure 19: Feature importance of the RF (Bank Marketing Dataset). 33](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/21017447_Simpson_Chiwashira_Final_Progress_Report.docx#_Toc152540468)

Table of Tables

[Table 1: Feature description of the bank and marketing dataset. 18](#_Toc152540469)

[Table 2: SVM Hyper-Parameter Results for Telco-Dataset. 27](#_Toc152540470)

[Table 3: SVM Hyper-Parameter Results for Banking-Dataset. 27](#_Toc152540471)

[Table 4: RF-hyperparameter tuning for Telco Dataset 28](#_Toc152540472)

[Table 5: RF- Hyperparameter tuning for Bank Marketing Dataset. 29](#_Toc152540473)

# **1. Introduction**

## 1.1 Background of Research Area

As technology advances, a variety of sectors have been impacted negatively and positively. Technology has been instrumental in the decision-making process across many fields. Two major business sectors that have benefited from this advancement include, the telecommunication and banking sectors.

The rapid development and advancement in machine learning capabilities has been made possible by the availability of data. Historical evidence in the companies can help forecast which individuals could potentially leave the business (Dalvi et al, 2016). The authors (Dalvi et al, 2016). emphasised that data analysis tools are useful in identifying the characteristics of the customers who have left the business. Several investigations have been carried out to analyse customer churn within organisations.

Customer churn is the act of leaving a service or subscription provider for another provider for various reasons (Dalvi et al, 2016). Authors (Kaur. I and Kaur. J, 2020), define it as customer attrition where one would stop a relationship with a certain brand or company. Below is Equation 1 (Rudd, Huo, and Xu, 2021) to calculate the rate of churn.

CustomerChurnRate = x 100 (Equation 1)

Some of the approaches that can be used to retain customers include customer churn predictions and behavioural analysis methods. Customer churn analysis and prediction is a data driven approach that uses machine learning tools. Due to the rapid development of computer technology and GPU over the last few decades machine learning models have developed exponentially. The use of machine learning tools has seen a wide variety of applications and usage across many industries, including healthcare, business, and engineering. A good example is the use of machine learning techniques used in data mining (Karvana et al. 2019), which include classification and prediction have been utilised to analyse the customer churn problem. Machine learning algorithms are key in analysing the historical record of the churners to identify the probability that an individual will leave a service or business.

For example, Rudd, Huo, and Xu, 2021) conducted a causal analysis, which demonstrated the importance of predictive modelling as a method of identifying customer churn. In their research, (Rudd, Huo, and Xu, 2021), in agreement with Coussement, Lessmann and Verstraeten (2016), suggest that we should use prediction systems to identify specific churners beforehand. This investigation therefore builds on top of the aforementioned work and evaluates the implementation and the performance of the tools. A literature review on these methods will be conducted and documented including the investigation of the strategies employed in the current literature.

According to Wu et al. (2021), the development of 5G technology has created a competitive market where most telecommunication companies wish to retain their customers. (He, 2022) argues that the banking sector has seen a sizeable decrease in profit due to customer churn.

One of the causes for customer churn includes having a highly saturated market, hence creating competition. Wu et al. (2021) claims that looking for new customers is more expensive compared to retaining current customers. Marketing campaigns and efforts to lure customers into subscribing to a specific service has proven to be difficult, hence the reason for the importance for developing and using different retention methods.

According to Coussement, Lessmann and Verstraeten (2016), retention methods can keep track of customers who might discontinue their services, thus losing the company business growth opportunities. Wu et al. (2021) suggested that using the proactive approach in identifying the customers before they leave would be key to retaining the current customers.

## 1.2. Project’s Aims and Objectives

This research aims to investigate the methods that have been used and the results that they yielded. The goal of the project is to evaluate supervised learning models to predict and analyse the rate of customer churn in a business.

*RQ: Is there a difference in the performances of the supervised learning models when predicting customer churn?*

H0: There is no difference in the performances of the supervised learning models when predicting customer churn.

H1: There is a difference in the performances of the supervised learning models when predicting customer churn.

## 1.3. Project’s Advanced Aims

1. Develop models to analyse data on customer churn.
   * Support Vector Machine,
   * Artificial Neural Network,
   * Random Forests.
2. Finetune the models to obtain better scores.
3. Compare model performance using the following metrics:
   * F1 score
   * Precision.
   * Recall
   * Test Accuracy
4. Make recommendations as to why the model was chosen, and the parameters give better results and should be considered by future developers.

## 1.4. Project Management

A Gantt chart was designed and used as the project management tool responsible for important major tasks and milestones of the project. It is key to the completion and success of the project as it gives an overview of the project and the timeline by which the tasks and milestones are expected to be accomplished. A copy of the Gantt chart is available in the appendices section (see Appendix 1). The project management tool used is the Team Gannt software which is free and readily available for use to anyone. A copy of the proof of version control is available in Appendix 2.

## 1.5. Resources

* Python IDE / Jupyter Notebook – This is one of the most widely used IDEs by data scientists which means there are a lot of support and forums that can be helpful.
* Computer with Windows 11 Operating System (see Appendix 3 for full system specifications).
* Datasets to be used in the experiment - Bank and marketing dataset and Telco dataset.
* Microsoft Excel, Visio, and Word – Industry-standard utility tools.
* Github Repository and Bitbucket Repository - For version Control (see Appendix 6 for repository URL).
* VS Code - Text Editor for the frontend interface development.
* Team Gantt Software - Project Management Software tool.
* Flask-Web Framework – The program is written in Python to run our application on our local host (For deployment purposes).

## 1.6. Legal, Ethical, Professional and Social Issues

In this project, the experiments that are being done do not involve using human subjects because publicly available datasets have been used. This means we do not need ethical approval from the University of Hertfordshire (UH) Ethics Committee. However, it should be noted that we would like to believe the data when it was collected from the customers there was an ethical approval before it was collected.

That said, the project has taken into consideration legal issues relating to the collection and processing of data during our investigation. By using publicly available datasets, this investigation believes that the General Data Protection Regulation (GDPR) guidelines are being adhered to. The investigation also believes that whilst collecting the data from customers for analysis, privacy was maintained and the we understand failure to uphold the law will lead to prosecution and being fined. The datasets used in this project were obtained from public platforms, meaning the data is available for anyone to use.

When collecting data from customers, it is a must that one explains why the data is being collected for transparency and the data should be processed fairly. The algorithms and methods to be used will not discriminate. This is why in this project evaluates all the data features and work to try and de-identify (remove any customer’s sensitive details if available). This is why we need ethical approval to do this.

We subscribe to some professional and ethics bodies like the ACM Code of Ethics and Professional Conduct and the British Computer Society (BCS); therefore, we will adhere to the ACM and BCS code of conduct for professional issues. We will be responsible and disciplined throughout the investigation and hold the data with integrity on a professional level. Although social issues might not affect the investigation, we acknowledge how customer demographics may feel misrepresented or not treated unfairly and they will lose trust in the research team which results in a negative sentiment.

# **2. Literature Review**

## 2.1 Review Protocol

For us to begin our secondary study of the literature review, the University’s digital library was utilised for the review protocol where exclusion and inclusion criteria of the sources were applied. A literature review on the background of the research area was conducted and was able to make decisions based on how to move forward with the project. In the review protocol, the search string was defined with keywords that were highlighted earlier in the section. After the execution, we can begin the secondary research. The inclusion and exclusion criteria defined were as follows.

Inclusion Criteria

* Studies must be peer reviewed.
* The research should focus on supervised learning models in customer churn.
* Manuscript extracted must be written in the English language.

Exclusion Criteria

* The studies that are not based on empirical work.
* Papers that are duplicates.
* Research that does not relate to or focus on supervised models.

Machine Learning in Business Sector

Machine Learning (ML) is a subcategory field of the Artificial Intelligence (AI) domain that has evolved in the past decades and more changes have come with this development. According to Sharma, Kaur, and Semwal (2022), ML makes use of past data i.e., historical data to forecast future results. The authors also make note of how ML has had impacts in different fields like business intelligence. This study is focused on the customer churn on the business side of a Telco Company and the Banking and Marketing sector.

## 2.2. Supervised Learning Models

Supervised learning is a type of ML technique where new data is classified based on the training data (Gaur et al., 2021). The learning process is done under supervision which means the training data used will be labelled and put into classes to be observed. The input data will be added to the ML model and the weighted data will be accurately fit into the model. In their study, (Sharma, Kaur, and Semwal, 2022), the authors argue that there are two types of supervised learning which are classification and regression.

According to (IBM, 2023: Karvana et al., 2019: and Dalvi et al.,2016) classification is one of the data mining techniques where the chosen algorithm tries to distinguish classes based on the training data provided. The model classifier will predict or forecast which category or class the test data belongs to. Algorithms that fall under this category and are to be examined include Support Vector Machines (SVM), Random Forests and Artificial Neural Networks (ANN).

Regression is a ML model that tries to map the relationship between dependent and independent variables and the goal of this ML algorithm is to predict if the values are near or closer to the actual output (Gosha et al.,2021; IBM, 2023; Sharma, Kaur, and Semwal, 2022). It is mainly applicable in market prediction, price prediction etc. Algorithms that are used in this type include linear regression and logistic regression.

## 2.3. Algorithm Selection

According to our literature search of the different algorithms and models, we can note that most of the algorithms which appeared were supervised learning models and these were mostly deployed by researchers in their studies. According to Colreavy and Lewandowsky (2008) in their empirical work, they uncovered the differences between supervised and unsupervised learning where they concluded that the difference between these two learning models is the learning rate. Supervised models rely on labelled data hence optimised learning rate gives them the optimal performance.

However, (Love, Medin and Gureckis, 2004) make a strong argument in which they claim that supervised learning is easier to understand but unsupervised learning is what most researchers call, “*The Blackbox*” because there is very little explanation of how the model arrived at the outcome. Rudd, Huo, and Xu (2021) agree with this claim as they highlight their developed deep learning model as having weaknesses. To mitigate this, they applied the Bayesian Network and used LIME and SHAP to find the most contributing factors during prediction. In this research, we can observe the differences between these two and why most of the algorithms were selected in customer churn analysis. This is because supervised learning is mostly used for pattern recognition and customer churn analysis is all about recognising the patterns of the customers that churn whom we intend to forecast. A customer who shows the same behaviour is most likely going to be classified as a potential churner.

However, with the unsupervised, the model does not know who churned and who did not. Since it has no prior knowledge of the model will cluster data points which are similar considering the specific features which are potential predictors. Therefore, we can argue that these two models will perform differently considering the practical application of the model. We have decided to look at supervised learning models and uncover the best one to help us answer our RQ by validating or refuting the hypothesis.

One of the most popular data mining algorithms which was used by scientists in this field is the Support Vector Machine (SVM). It is part of a set of supervised learning models mostly used in instances of classification, and regression and to identify any outliers in the dataset (Pedregosa et al, 2011). The ability of this ML technique to generalise and solve both linear and non-linear data in high-dimensional feature spaces makes it ideal. According to (Piri, Delen and Liu, 2017), SVMs try to produce a decision boundary on the hyperplane as it cuts across the data points and separates them in data space. The classifier aims to minimise any errors while classifying the data and to maximise the margin of the decision boundary.

As shown in Figure 1 (Pedregosa et al, 2011), the dashed lines denote the margin of the decision boundary, and the middle solid line is the decision line or hyperplane which divides the two classes. The two brown dots on the dashed lines and the blue dot on the dashed line are what are called the support vectors from which the name of the model is derived.

In this discussion, it is clear that the SVMs solve problems of linear and non-linear data. This is made possible using Kernels. These enable the SVM to be mapped to a feature space. This is achieved using the mapping function which can be denoted by (He, 2022). The non-linear SVM makes use of convex quadratic programming to solve problems. The general formulation of the decision boundary is denoted by Equation 2 (He, 2022).

(Equation 2)

Where is the weight that is associated with the inputs or features that is denoted by and the is the bias term. The generalised version of the same formula while incorporating the kernel function can be denoted by Equation 3 (Hanyue He, 2022) below:

(Equation 3)

In this equation the term is the Lagrange multiplier that maps the inputs into a higher dimensional space while the which is the weight associated with the inputs in this case. The kernel mapping function is denoted by the term .

A diagram of a graph

Description automatically generated with medium confidence

Figure : Decision Boundary for SVM Classifier (Pedregosa et al, 2011)

This ML algorithm was proposed by the scientist Vapnik in 1992. Since the introduction of this classifier, we have seen different variations of the same model being applied in different areas. According to Cortes and Vapnik (1995), the classifier has performed better than most of the learning models available. We observed that (He, 2022) has also used two variations of the SVM model and compared them against each other in the analysis and prediction of bank customer churn.

After applying a 50:50 resampling split, the algorithm had the highest recall metric score which was a staggering 73.24%. The authors concluded that class sampling of the dataset was key to the model's performance (Kaur. I and Kaur. J., 2020; Karvana et al., 2019). We can see that the SVM classifier performs better when the data is balanced. In their report, Piri, Delen and Liu (2017) define the performance of the model when the data is imbalanced to have a “*dramatic deterioration*” in performance. This is why these authors, (Karvana et al., 2019; Wu and Chang, 2003) agree on the need to have balanced data split when training the model.

While the SVM classifier appears to be one of the most used methods in predicting customer churn in the search results, the Random Forest classifier was also used. Random Forest (RF) is part of the ensemble techniques used for both classification and regression tasks (Kaur I and Kaur J, 2020: Breiman, 2001). Since the introduction of the RF by Leo Breiman in 2001, a surge in the use of this algorithm has grown due to its nature of handling large data and generalisation ability (Breiman, 2001). Ensemble techniques in ML are usually a combination of different classifiers that are trained on a specific dataset and output different decisions with the sole goal of improving prediction and overall performance. Ensemble learning techniques use the two main methods which are voting and averaging (Kaur I and Kaur, 2020).

RF follows the ensembling techniques due to its nature and architecture, according to Breiman (2001), the RF classifier has a tree-like structure that consists of other classifiers of the same nature (decision trees) and each tree votes for the majority class. RF is generally a multiple of decision trees, hence it is key to understanding decision trees. Decision trees are part of supervised learning models that can compute classification and regression tasks. They are not shy of shortcomings as they are usually prone to overfitting (IBM, 2023).

The RF classifier can overcome overfitting by employing bootstrapping. According to Breiman (2001), randomness is key to avoiding model overfitting and this is the reason why the basis of this RF lies in the research of (Amit and Geman, 1997: and Ho, 1998) as they all agree on the random subspace required for the growth of each tree. Bootstrapping is a fundamental tool in RF to facilitate the establishment of random subsets of training samples. These training samples are randomly fit in each decision tree’s split.

Now that all the decision trees in the model have a random subset of the dataset, it means the risk of overfitting is avoided and instead, there will be diversity of the data in the model’s structure. The RF is robust to outliers and has a well-improved generalisation ability and mitigating the shortfalls of the model. According to the article by IBM (2023), the author claims that the RF uses the bagging method, and this is an additional step to ensure that the trees in the classifier are not correlated hence randomising the data subset.

In the summary of related works that were conducted by Kaur. I and Kaur. J (2020), the authors in question drafted a table where they concluded their findings and, in their quest, the RF proved to be one of the leading ensembling models that outperformed the other models in the conducted experiments. However, Kaur. I and Kaur. J (2020) aimed to conduct a churn analysis for the banking sector. They decided to do a hyperparameter tune of the n\_estimator and the max\_depth for better performance evaluation. The tuned parameters yielded good results which they evaluated using evaluation metrics like AUC-ROC, precision, recall and overall test accuracy. The RF showed that the sampling strategies employed achieved a higher accuracy of 85.23% with a stratified sampling strategy. They went on further to utilise ensembling methods where the RF was better than all the classifiers with averaging and voting techniques being applied. They attained 85.22% making it the best model to select when it comes to the banking sector as concluded by their research.

According to Wu et al (2021) in their research on customer segmentation and churn prediction, they applied a sampling strategy of oversampling the training set with the Synthetic Minority Oversampling Technique (SMOTE). In one of their datasets, the RF performed well with and without SMOTE applied where they achieved 92% and 95% respectively. The big question is what if the model was overfitting?

The literature review on the structure of RF suggests that bagging and bootstrapping would reduce the risk of overfitting. We know from the discussion that the RF classifier can handle imbalanced data and since every tree in the RF has random subsets of the features that means every class can be equally represented. The main goal now is to find the best hyper-parameter after the experiment which yields the best performance and find out feature importance and compare with the correlation heatmap.

Supervised learning models like Artificial Neural Network (ANN) have been used in pattern recognition tasks. This is from the results after the execution of the search string on supervised learning models. According to research by Park and Lek (2016), they define an ANN as a computational model inspired by biological systems. The model emulates the neurons in the brain and how they signal and fire.

A diagram of a network

Description automatically generated

Figure : ANN Structure - Bekesiene, Smaliukiene and Vacantiene (2021)

As previously discussed, the ANN mimics how the neurons send signals to each other, they are structured as illustrated by Figure 2 (Bekesiene, Smaliukiene and Vacantiene, 2021). The figure shows how all layers and nodes are interconnected to each other and the presence of an activation function to compute the output of each layer (IBM, 2023). The values are the inputs which are in the input layer (IL), and they are interlinked with the hidden layer (HL) and the output layer (OL).

The ability of the model to learn and try to improve its training and testing accuracy is the reason why they are used in pattern recognition applications. Wu et al (2021) in their research discovered how the ANN was mostly used in the prediction of customer turnover whilst employing SMOTE methods to resample the training sets. However, the authors used the Multi-Layer Perceptron (MLP) in their empirical work. The MLP was not outstanding except when they applied SMOTE to the training sample it performed better than other models with an F1-Score of 42.84%.

According to Rudd, Huo, and Xu (2021) in their pursuit to conduct research on the causal analysis of customer churn, they focused on the Deep Feed-Forward Neural Network (DFF-NN). The authors claim that DFF-NNs have room to handle massive data unlike the other traditional classifiers and in some instances, Convolutional Neural Networks (CNNs) have been used collaboratively with ANNs and DFF-NNs. The authors had skewed data distributions and turned to employ the oversampling method of SMOTE. For comparison’s sake, they went with ensembling methods and the proposed DFF-NN had the joint highest accuracy with the RF classifier however the XGBoost seemed to outperform all other models with an AUC-ROC of 80%.

In this review, we can conclude that each of the chosen models performed well after the hyperparameter-tuning and each model has its benefits and disadvantages that we just discussed. The most important factor we can take out of the review is the size of the data to be trained on, models like the ANN and RF can work with large data and will give better performance but there is a trade-off which is computational time. Data that has skewed distribution will need to be dealt with for the model to have enough representation for each class label.

# **3. Methodology**

In this investigation, we are using quantitative or objective methods to conduct the research whilst answering the posed RQ. We want to find out how much the performances of the chosen models differ considering the metrics of evaluation that will be discussed further in the report. The justification for the use of quantitative methods is that we will be using numerical figures to evaluate our models using metrics such as F1-score, Recall, Precision and Test Accuracy.

## 3.1. Project Plan

We will continue with data exploration and pre-processing as we have begun in the IPR. The next stage is to analyse the data and explore the columns with missing values and imbalanced data. Appropriate sampling techniques must be identified and applied. Features that are key predictors of the models will be identified and utilised in the modelling and training of the model. Appropriate resampling methods and tools will be used to resample the training sets before we can train and test the models.

The chosen algorithms will be applied and commence the training of our models. We can start analysing the results and fine-tuning the hyper-parameters of the model. The plan is to utilise the confusion matrix and make sure we reduce the False Positives and False Negatives. The pickle function in Python is to be selected for use to save the trained model so that it can be loaded again to be used again without having to retrain the model every time. This is also useful to load the ML model when one wishes to integrate it with User Interface (UI) like web applications and deploy it.

## 3.2. Data Selection

The initial project plan was to use only one dataset for the experiment which was the telecommunications dataset obtained from Kaggle. However, because we did not have the time and resources to compute data of high dimensionality and magnitude in size, we decided to use two datasets with enough data for our project from the banking and marketing sectors and the telecommunications sector with real-life business data. The initial dataset was used during the interim progress report, and it had 70,000 instances with 172 attributes (see Appendix 5). The column names were encoded, and a dictionary was attached to the folder for reference. Most of the data was encoded to binary (refer to Appendix 5) and there were missing values as well.

## 3.3. Datasets

### 3.3.1. Banking and Marketing Dataset

This dataset was obtained from the UCI Machine Learning Repository. The dataset can be easily accessed from the [link](https://archive.ics.uci.edu/dataset/222/bank+marketing) embedded in this document. According to the creators of this dataset Moro, Rita, and Cortez (2012), the dataset was collected from the data extracted by a Portuguese banking institution. The initial plan was to collect data to be used for classification purposes. The dataset repository consists of at least four datasets with data from May 2008 up to November 2010. The goal of this dataset was to use ML algorithms to predict whether a client would subscribe to a term deposit. However, this dataset has been cited by many researchers for different investigations and experiments than what it was originally set up for. We will be using the dataset to predict and analyse customer churn.

#### Exploratory Data Analysis (EDA)

We know the metadata for the dataset and that means we can begin the exploration of our data. In this investigation, the Pandas library was utilised to allow us to read the data which is in a Comma Delimited Value (CSV) format. We would like to understand the data types we have, the unique values we have and the shape of the data. After the data is loaded into the Data Frame the df.shape code snippet resulted in us knowing that the data has 45,211 instances and 17 features.

A table with black text

Description automatically generatedAfter loading the data, we want to know what the data looks like and how many unique values appear in a column. So the best option is to use the df.head() and that resulted in Figure 3 as shown below;

Figure : Output of the Banking Dataframe

We are aware of the shape of the data and have an idea of what it looks like. But what does the data mean? It is ideal to know the units of measurement and what the column names mean. We also want to know how many unique values appear in each feature. The table below shows the summary of each feature and the unit of measurement.

|  |  |  |
| --- | --- | --- |
| Column Name | Data type | Description |
| age | Integer | The length of time the customer has lived. |
| job | Categorical | If someone is employed or not or a student. |
| marital | Categorical | If the client is widowed, single or married, etc. |
| education | Categorical | How educated the client is? |
| default | Binary | Does the customer have any credit? |
| balance | Floating point | Yearly mean balance in Euros (€). |
| loan | Binary | Does the customer have any personal loans active? |
| housing | Binary | Is the customer on any housing loan? |
| contact | Categorical | How is the customer contacted by the bank? |
| day | Integer | On the last day, the customer was contacted. |
| month | Categorical | Last month, the client was contacted. |
| duration | Integer | Time the call lasted for in seconds. |
| campaign | Integer | Number of contacts made during the campaign. |
| pdays | Integer | Number of days since the last contact. |
| previous | integer | The number of contacts made prior. |
| poutcome | Categorical | Was the campaign successful or not? |
| Y (desired target) | Binary | The desired target. |

Table 1: Feature description of the bank and marketing dataset.

According to Moro, Rita, and Cortez (2012), features like duration were measured in seconds and it affects our desired target. For instance, when the duration is zero value then we know the output is going to be “No”. The prediction of the model is most likely not going to be accurate and therefore we can dismiss the duration as part of the potential predictor. However, we will be employing a correlation heatmap to find any highly correlated features, and these could potentially be our predictors.

A blue and orange pie chart

Description automatically generated

Figure : Pie chart showing data imbalance for Bank and Marketing Dataset.

According to the data, as shown in Figure 4, we can see the proportion of the customers who churned against non-churners. In a real-world setup, there are going to be fewer customers who churn. The data observed under the exploratory analysis backs that claim as we have 5, 289 which is 11.7% of customers who churned and 39,922 customers which is 88.7% of customers who did not churn after the application of Equation 1. However, there exists a data imbalance between the two classes which will be discussed in the pre-processing section in detail.

### 3.3.2. Telco Dataset

Telco dataset was obtained from Kaggle a public repository which has several datasets. The link to the repository is available [here](https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data). Originally this dataset belonged to IBM sample datasets and has been available online since 2017 and has seen several citations and references of this dataset in the data science community. The main aim of this dataset was for the development of customer retention models to be used to analyse customer behaviour and to build predictive models accordingly.

The repository contains one dataset with 7,043 instances and 21 features. The column labelled churn is our desired target column for classification. The rows of the data represent an individual’s attributes. The dataset can be divided into three categories: demographics, customer information and services. Services tell us which services the customer is subscribed to for instance streaming, tech support etc. The demographics tell us about information like gender, age etc. The customer information covers the contracts, payment information etc.

Exploratory Data Analysis

The dataset obtained from Kaggle did not have enough metadata about the dataset to understand the dataset. How data was collected and what was the intention of the dataset? However, we were able to make use of the little data we had from Kaggle and do an exploratory analysis using the pandas library in Python language. The figure below shows us the data types that we have, and we can classify the features into three distinctive categories which are demographics, personal information, and services. The figure also shows the number of data types of each feature and the number of columns and attributes.

A screenshot of a computer

Description automatically generated

Figure : Columns and data types for the Telco Dataset.

A screenshot of a computer

Description automatically generatedWe were able to look at the data that we have and look at what it would be like utilising the head function as mentioned earlier. The figure below shows what the data looks like before we begin the pre-processing and after being converted into a data frame.

Figure : Output showing head of Telco Dataframe.

In the next phase, the goal was to find out if we have data imbalance and what is the proportion of churners to non-churners. This is depicted by a pie chart which shows this in percentages in the figure below.

A pie chart with a number of percentages

Description automatically generated

Figure : Pie plot showing the Telco data imbalance.

We have attained this figure after the application of Equation 1, which was described earlier in the report. The value count of churners amounts to 1,869 which is 26.5% of the population of the dataset whilst non-churners are 5,174 with a higher percentage of 73.5%.

## 3.4. Data Pre-Processing

A screenshot of a computer

Description automatically generatedIn this phase, the two datasets were cleaned at the same time. They are not identical, however, there are cleaning strategies that can be applied to both and that has helped us to work on them much quicker. We faced some challenges, and the error messages were integrated, and different methods or approaches were employed.

Figure : Output for loading the raw CSV file.

A screenshot of a computer

Description automatically generatedOne of the major issues was that the datasets when downloaded were in the CSV format but not in each column when you opened them with the spreadsheet, in this case, the MS Excel application or when loaded using the pandas functions. Figure 8 shows the Python output using Pandas and Figure 9 shows the MS Excel output.

Figure : Showing the MS Excel view of the raw CSV file.

The above-mentioned problem was solved by loading the data using Python’s built-in function of reading external files and this allowed us to create a list where we stored the data after the splitting data. Then we were able to use the pandas library to convert the data into a data frame saved as a CSV to be loaded and for better inspection as the attribute was in its cell.

The two datasets have different data types and most of the features in the Telco dataset have object data types. The first step was to use the metadata to make sure the data types were correct, and we found out that the *TotalCharges* feature was an object data type. Financial balances usually are in the form of floating points, which can allow us to scale them. The banking and marketing had the balance as an integer data type. As mentioned above financial balances can be expressed as float data types and we used the pandas function to convert the data types.

The next phase was to drop the features that would not help in the prediction, these included the *customer ID* in the Telco dataset and the banking and marketing, we dropped the duration as explained in the EDA section and later dropped the *month*. Afterwards, we are left with features that have a unique value that is binary and the replace method was utilised to replace anything with the strings “*yes*” with a numeric value, “*1*” and “*no*” for a “*0*”. After that features that had more than two were transformed. The R Programming language uses the spread function for this operation however, in the pandas library, Python uses two methods that can be applied which are the pivot and one-hot-encoding (the chosen method).

Missing data

During the data preparation stage of the Telco Dataset phase, while applying the oversampling technique to the training data an error kept occurring. Figure 10 below shows the error message that appeared and prompted me to look back at my data and see if I had any missing values which did not appear the first time. However, after the conversion of all the data types that was when we realised the *TotalCharges* attribute had 11 missing values. The dataset was inspected to find the rows with the missing values using the pandas functions. The CSV file was inspected to confirm whilst using the row numbers and the fields were empty.

There are various ways to treat missing values which include filling them or dropping rows with missing values. In this case, we decided to drop the 11 values because the rows with missing values belong to the majority class which is 73% of the dataset. We can afford to drop these rows since we have enough data for training and testing the model.

A screenshot of a computer program

Description automatically generated

Figure : Error during the application of SMOTE

## 3.5. Data Imbalance

Class imbalance is a phenomenon where the distribution of data is not the same in each category or class. Real-world datasets usually have an imbalance, and this creates a problem of misclassification and inaccurate results. A class that has more data points than the other is the majority whilst the other one is a minority class.

According to Lee and Kim (2018), data from real-world applications will always be imbalanced with overlapping points. The problem that arises from data imbalance is usually the bias of the developed model. The model is always going to be biased towards the majority class. There is always a higher chance the algorithm or model will classify all data points to be belonging to the majority class. The models will score a higher accuracy in prediction which is not the true reflection of the data. Hence there is a need to look at our data and apply methods that can handle this problem.

Methods like oversampling, undersampling and weighted random sampling (in Pytorch) have been employed to tackle this problem. Oversampling also known as up-sampling is a method of creating new instances to oversample the minority class by matching the majority class. Undersampling also commonly known as down-sampling is a method that randomly reduces the majority class to match the minority. This technique can be a problem because we lose vital data since the resampling is done randomly and this will be normally reflected through the performance of the model. It still can be good if we have enough data sampled after the resampling since we will have enough data samples to train the model classifier. Weighted-Random-Sampler is a method that will assign a higher weight towards the minority class during the training. The only problem is that it does not solve the imbalance, but it gives enough representation.

There are many reasons why data could be imbalanced, and these include mistakes during data collection or capturing and data omitted through the selection process. One of the biggest issues is that the reason for data imbalance could be geographical or environmental for instance a certain race could have a higher representation than another because the race could be native to that location. The same concept applies to our case because we will never have a lot of customers churning at the same time, thus why the data can be imbalanced, but we can create synthetic data points. Chawla et al (2002) in their journal they developed a technique called Synthetic Minority Oversampling Technique (SMOTE) which creates synthetic data points from the existing data points. This is a method to be employed in our study whilst we will also be employing the Random Undersampler Technique for undersampling.

## 3.6. Data Preparation and Transformation

At this stage, the data is almost ready but needs to be prepared for the classifier. Data preparation comes after the data preprocessing is completed and it is very important because it reduces the input complexity while thriving to maximise the scalability of the model. The main aim is to process the classification whilst preserving the high performance (Coussement et al, 2016).

In both datasets, we managed to apply the MinMaxScaler to scale down the inputs to values between 0 and 1. This makes it easy for the classifier to run quicker with smaller numbers than actual. The features with more than two unique values had to be reduced, although this was done in the preprocessing stage it is still part of data preparation as well. The one-hot-encoding was employed to reduce the number of unique data values.

After the transformation, the data must be split into independent and dependent variables and then split into training data and testing data. The dependent variable is the treatment or desired target is dependent on the features and in our case, the “*Churn*” attribute. This is the dependent variable because its value is dependent on the independent variable which is all the other features.

The ratio of the train test split employed was 80:20 where 80% of the data was used for training and the remaining to be used for testing purposes. We want our model to learn as much as possible, the more we train it with more data the more it can learn different patterns that will help the prediction and make it more accurate.

The data is now ready to be fed into the classifiers. However, the data is still imbalanced, and in this experiment, the goal is to improve performance in predictions and evaluate it. So, three methods will be tested and the one that produces the best performance will be selected. The data is separated into oversampled train test splits, original train test split and the under-sampled train test split. The data is now ready to be trained and it is very important to note that the training data is the one that is oversampled or under sampled, this is because we want the maintain the distribution of the data.

# **4. Model Implementation**

## 4.1. Model Architecture

A computer screen shot of a diagram

Description automatically generatedThe project aims to analyse and predict the customer churn rate using the supervised learning models. The figure below is the flowchart of the project plan.

Figure : Iterative flowchart of the model implementation.

The above flowchart illustrated in Figure 11 shows the plan execution for both notebooks and datasets. In the first stage, we loaded our data and converted it into a data frame. After the data is loaded, data exploration and pre-processing are conducted. When the data is ready, the data is prepared and converted into numerical data because the Sklearn library requires the data to be in numerical format. Hence, we must scale our model data for easy classification. The prepared data was split into training and testing data samples.

As explained in the methodology section on data imbalance to tackle this problem we are going to run three different models with three different sampled training sets. We will apply the SMOTE algorithm in the training set and under-sample the training data using the random sampler. Lastly, we will use the imbalanced data to see any differences and how different classifiers will perform under those conditions. The next step is to apply the model classification without any tuning. The different models will be trained, and the results documented for analysis. The next stage is to tune the hyper-parameter of the three different models and evaluate the results. This is an iterative process as we must retrain the model until the results no longer change. Finally, we can deploy the model to be tested and used by the end user.

## 4.2. Model Training

Our datasets were all trained using the same models and with the same sampling techniques. During the resampling process, new variable names for training were created and the resampling methods were applied. The resampling methods did not specify ratios of the resampling but instead, the default parameters applied automatically. After applying SMOTE, the algorithm resampled the majority class of the Telco Dataset by reducing the majority class and increasing the minority class. The figure below shows the before and after the sampling.

A screenshot of a computer code

Description automatically generated

Figure : Application of SMOTE on Telco Dataset.

The same principle was applied with the Random Undersampler algorithm to down-sample the training data and it down-sampled both the majority and the minority classes. Figure 13 below shows the new resampled training data.

A screenshot of a computer code

Description automatically generated

Figure : Application of the RandomUnderSampler on Telco Training Data.

## 4.3. SVM – Implementation and Hyperparameter Tuning

SVM Classifier was utilised in both datasets. The first method was to use the unsampled data which was still imbalanced. We built our model by calling the SVC() method which is the Support Vector Classifier. We used the original train test split and fit the model. Following that we used the predict() method to test the pre-trained model. A classification report was generated, and the results were looking well. The same was applied throughout both datasets and for both resampling methods as discussed. The code for this is embedded in this report in the appendices section (See Appendix 6a).

Model fine-tuning

Many ways can be used to optimise the model to get better results and in this case, for the SVM the GridSearchCV() was explored. The algorithm itself takes multiple arguments and we were mostly interested in the classifier and the grid which has the parameters. The parameters to be used are the Kernel which was defined in the literature review, gamma values and cost. The cost denoted by “C” is the trade-off that we can afford for misclassification. The greater the cost value the lower the margin, it works inversely with the decision boundary margin. The gamma value influences the support vectors and if used together with the cost can prevent overfitting (Scikit-Learn, 2011).

The algorithm will take the parameters in the grid and fit all the permutations, and it will train multiple models using the given parameters. The best parameters are saved and used to fit the model and provide the classification report and output.

The table below shows the best parameters for the Telco dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Training Set | Cost (c) | Gamma | Kernel | Accuracy |
| Imbalanced Set | 10 | 0.01 | RBF (Radial Basis Function) | 79 % |
| Oversampled | 10 | 1 | RBF (Radial Basis Function) | 74 % |
| Under-sampled | 100 | 0.01 | RBF (Radial Basis Function) | 71 % |

Table 2: SVM Hyper-Parameter Results for Telco-Dataset.

The table below shows the best parameters for the Bank Marketing Dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Training Set | Cost (c) | Gamma | Kernel | Accuracy |
| Imbalanced Set | 1,000 | 0.01 | RBF (Radial Basis Function) | 89 % |
| Oversampled | 1,000 | 1 | RBF (Radial Basis Function) | 74 % |
| Under-sampled | 1,000 | 0.001 | RBF (Radial Basis Function) | 72 % |

Table 3: SVM Hyper-Parameter Results for Banking-Dataset.

We can see that the accuracy dropped drastically in both datasets for all the sampling methods. When we evaluate our models, we will arrive at why the models performed the way they did. While also looking at different parameters we tried to look at different kernels. Most real-world datasets do not have linear data points that can easily be separated by the hyper-plane. This is why the idea of using the linear kernel was not considered. However, the polynomial function was another kernel that could be explored, and the following are the results we got from the Telco dataset.

|  |
| --- |
| ========================================  SVM - Original (Unsampled Dataset)  ========================================  {'C': 1, 'gamma': 0.1, 'kernel': 'poly'}  SVC (C=1, gamma=0.1, kernel='poly') |

The overall accuracy was 78% which was 1% less than using the RBF, where we had unsampled training data. This kernel did not perform badly but due to the computational time needed the idea of using it was dismissed. Other kernels like sigmoid were not explored but could be in future because there is limited time to complete the project.

## 4.4. RF – Implementation and Hyper-Parameter Tuning

The RF was applied on all datasets and all sampling methods were applied. The RF classifier was called and initially, we started with the n\_estimator parameter to be a value of 100, the max\_feature to be “sqrt” and the max\_depth to be 10 with a random state of 42. This was applied to the data, and we got the classification. The results were reasonable, but we felt like they could be optimised further to get a higher accuracy score.

Model fine-tuning

In the SVM Classifier, the grid search algorithm took a lot of time because generally, the SVMs are not good with large datasets and the grid search algorithm given the parameter grid would take so much time to compute and train. The Telco dataset has fewer instances than the Bank Marketing and it took time to run compute results and RFs are slightly slower in training. Because time is limited another hyperparameter optimisation tool was explored which is the RandomSearchCV.

Time is an important resource in this investigation and hence finding ways to save while getting the best outcome is key. The RandomSearchCV gives us the ability to explore the best parameters whilst also saving time. The algorithm randomly picks defined parameters within a certain given range. The trade-off is that because it is randomly picking parameters the one that is not chosen might be the best one but if the randomisation process is done well that could be avoided.

The RF has many parameters that contribute to the success of the model, but the two most important are the n\_estimator and the max\_feature. According to (Scikit-Learn, 2011), the bigger the n\_estimator the better the model because that is the number of trees in the forest but also at the cost of being slow since that means more time to compute. The max\_feature is the size of the random subset of training features. These are the parameters that were tuned in RF and additionally, the other parameters that were tuned to reduce the size of the model as shown in Equation 4 (Scikit-Learn, 2011) below.

*O (M.N.log (N))*  Equation 4

Equation 4 shows the space complexity that also affects the time complexity taken by the algorithm where M is the number of trees and N is the number of samples (Scikit-Learn, 2011). The authors suggest that we should tune other parameters which were added to the parameters list. The tables below show the results from fine-tuning the RF models.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | n\_estimator | min\_sample\_split | min\_sample\_leaf | max\_feature | max\_depth | bootstrap | accuracy |
| Imbalanced | 200 | 5 | 4 | auto | 10 | True | 79 % |
| SMOTE | 1800 | 2 | 1 | auto | 20 | False | 76 % |
| Undersampler | 400 | 2 | 4 | sqrt | 10 | True | 73 % |

Table 4: RF-hyperparameter tuning for Telco Dataset

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | n\_estimator | min\_sample\_split | min\_sample\_leaf | max\_feature | max\_depth | bootstrap | accuracy |
| Imbalanced | 400 | 2 | 4 | sqrt | 10 | True | 89 % |
| SMOTE | 1400 | 2 | 1 | auto | 40 | False | 86 % |
| Undersampler | 200 | 5 | 4 | auto | 10 | True | 76 % |

Table 5: RF- Hyperparameter tuning for Bank Marketing Dataset.

The results will be evaluated and analysed in the next section, but optimal parameters were achieved from the random search algorithm.

## 4.5. ANN – Implementation and hyperparameter tuning.

We began with building a base neural network which has the IL, this translated to the number of columns or features that the ML classifier takes. The HL and OL had dense layers with 15 neurons and 1 neuron respectively. The reason for having one OL is because our output is either 1 or 0. In between the three layers are the activation functions that will compute the outcome of the layer based on the inputs and corresponding weights. The activation functions were Rectified Linear Unit (ReLu) and Sigmoid. The sigmoid function is mainly used in the output layer. The Adaptive Moment Estimation (Adam) is the optimizer that is mainly used to update the weights iteratively. For our loss function, we went with the binary cross\_entropy since it is a binary classification problem for pattern recognition. Our epochs were set to 100 and then from there, we built our base model that would be fine-tuned and get better results.

Model fine-tuning

In the previous classifiers, we used some methods that would find the best parameters for us. However, for this case, we decided to manually tune the algorithm. For the model fine-tuning, we used the oversampled dataset because there was no point in working with imbalanced data that would give us high accuracies without telling the full picture. For the first model, we added an extra HL to see the results. Adding an extra layer enables the ML to extract features from the input layers, in turn, may or may not improve the accuracy. In this case, the accuracy went down, and it did not improve. The second model was to change the learning rate of the Adam optimiser. There were some slight changes but not significant. The third model had more changes as we added dropout and regularisation. These are mainly used for large networks with large training data and help with overfitting as it drops the nodes that are noisy and there were little to no changes. However, in the next model, we manipulated the dropout rate and we saw some improvements in the performance. The next model saw us change the IL’s activation function from ReLu to the Tanh function. There was a change in the accuracy as well as other metrics which was positive. The model that gave us one of the best results was after we removed on HL that we had added.

# **5. Evaluation and Analysis**

## 5.1 Evaluation Metrics

As highlighted in the advanced aims and other sections leading up to this point in our report, different metrics were used alongside the overall classification accuracy to measure the effectiveness of the model classification. Precision was used as one of the key metrics in evaluating the performance of the model classification as it generally measures how correctly the model accurately classifies the labels. Recall or sensitivity metric focuses on the effectiveness of the positive predictions. When the precision is very high and the sensitivity is low, the overall classification accuracy is high. But this does not tell the full story hence the need for a balance between these two metrics. F1-Score is a combination of precision and recall, and this is an important metric to consider in the performance of the models.

## 5.2. The Confusion Matrix

This is a matrix which shows the predicted labels against the actual predictions. The values that are established here are the ones that compute the metrics discussed above. This is why it is important to reduce the False Positives (FP) and False Negatives (FN) whilst maximising as many True Positives (TP) and True Negatives (TN) as possible. The figure below shows the confusion matrix of the best model for Telco Data showing the model performance.

A screenshot of a graph

Description automatically generated

Figure : Confusion matrix showing the performance.

## 5.3. Evaluation of the Telco Dataset

Throughout our journey, we have seen three different classifiers being implemented under three different sampling methods. We have seen the best results achieved with visual confirmation after further fine-tuning. According to our findings, the group with the best models chosen was the oversampled group. We chose the group because it has a good representation of data in each class making sure the model was not biased. Figure 15 below shows the plot of the three chosen and best models.

A graph of different colored bars

Description automatically generatedFrom the results, we can visualise the performance of the algorithms. However, the two outstanding models were ANN (model 6) and the RF (model 2). The Accuracy for these models was tied at 76% however, even though the ANN got the better recall and f1-score, the RF has a better balance of the precision and recall. The RF was chosen as the best-performing model to be used. The individual performances of these plots are available in the appendices section with different plots (see Appendix 6b).

Figure : Performance of Classifiers on the Telco Dataset.

## 5.4. Evaluation of the Bank Marketing Dataset

A graph with different colored bars

Description automatically generatedThe best three classifiers chosen did well in terms of the overall classification accuracy although the RF (model 2) stood out. However, all models struggled with precision and the f1-score. The metrics scored less than 50% and the RF had the lowest recall. There was a balance between precision and sensitivity which was a trade-off we accepted to conclude that RF was our model of choice. The figure below shows a plot of the models and performances.

Figure : Performance of Classifiers on the Bank Marketing Dataset.

## 5.5. Feature Importance and Potential Predictors

The RF is a powerful tool for classification, and it can compute the important features of the model. we managed to plot the feature importance of each chosen RF model in all datasets. Below are the figures showing the feature importance and respective weights.

A graph with blue and white bars

Description automatically generated

Figure : Feature Importance of the RF (Telco Dataset).

Figure 17 above shows us how tenure has the highest weight of 0.15 which contributes to the prediction. It would make sense that time (*tenure*) spent with a company would determine if one churned or not. This is also followed by other factors such as *MonthlyCharges*, *TotalCharges*, *Contracts* (*month-to-month and two\_year*) and *InternetService\_fibre\_optic.* This data is further supported by the correlation heatmap which was plotted before the model training or application of any data preparation.

In Computer Science, a correlation coefficient is considered strong when it is . The *tenure* and *churn* attribute had a correlation coefficient of 0.83 and this is one of the predictors. The second-best attribute was *MonthlyCharges* which has a weight of 0.145 and in the heatmap, the *InternetService\_fibre\_optic* and *MonthlyCharges* had a weight of 0.79 which is still higher. The *TotalCharges* and *Churn* showed a stronger correlation with a coefficient of 0.83. The *TotalCharges* feature was the third most important feature with a weight of 0.13.

The figure shown on the next page is the correlation heatmap showing the strongly correlated features.

A screenshot of a chart

Description automatically generated

Figure : Correlation Heatmap for Telco Dataset.

The illustration below (Figure 19) shows the feature importance of the Bank Marketing dataset and just like it was discussed in the other dataset the important features were the *campaign*, *balance*, and *age*. The correlation heatmap had a weak correlation and hence did not further support it.

A graph with blue and white bars

Description automatically generated

Figure : Feature importance of the RF (Bank Marketing Dataset).

## 5.6. Research Question and Hypothesis

We posed our RQ in which we wanted to answer in the customer churn predictions and analysis. We can validate and answer using the results reported. Generally, the RQ needed us to focus on the best-performing model. According to our results discussed in the evaluation section, for both datasets, the RF Classifier was the best supervised learning model as it produced the best performance. As a result, we rejected the null hypothesis hence validating the alternate hypothesis (H1).

H1: There is a difference in the performances of the supervised learning models when predicting customer churn.

We accept the alternate hypothesis (H1) for both datasets used in the investigation. However, the performances are not marginally different. The ANN was closer to the RF in all datasets and the SVM was still a long way to compete with the two classifiers.

# **6. Conclusions**

This investigation aimed to answer the RQ and make the best recommendations for future data analysts. The main findings were the effectiveness of the resampling techniques and the RF classifier and other reviewed ensemble techniques in customer churn. The preprocessing techniques of resampling the imbalanced data did play an active role in yielding unbiased results.

We observed the effects through the classification reports where the overall accuracy of the models where significantly different in some instances. As expected, the imbalanced would be biased towards the majority class hence having higher accuracy, however, the results for oversampled and under-sampled data showed a gap between the two performances. The expectation is that both resampling techniques will create equally represented classes. Due to the randomness of the process of removing the data in the majority class, crucial data is lost, and this would be a possible reason that affected the classification performance as the models were not trained for all possible instances.

There was a balance in the metrics classification report between the minority and majority classes after model training and cross-validation. The RF proved to be better than the SVM and ANN in this experiment due to the nature of the data and size. For the feature extraction by RF, we can look at the features that contributed less to the prediction and segmentation of those features can help the business to isolate the characteristics or attributes of a churner.

# **7. Limitations and Further Work**

The major limitation was the time factor to implement and test all different scenarios. However other factors like computational resources would have made the investigation thorough and would have helped to achieve even more. The investigation achieved reasonable results, however, the hyperparameters can be further tuned to produce better metrics for a more accurate classification. In the RF we could reduce the model in the event of deployment by dropping the less important features.

We applied the oversampling and undersampling of the training data. We could try to use other oversampling techniques like ADASYN (Adaptive Synthetic) and other variations of SMOTE because the traditional has limitations such as over-generalisation and variance. Visualisation tools and dashboards can be used to help the management make informed decisions and target the customers before unsubscribing from them. A model prototype of the form that can be filled and predict the customer churn with confidence level was designed and deployed (See Appendix 7).

# **8. Reflection**

Engaging in this project provided valuable insights into the practical application of data science and analysis. The hands-on experience of developing models and implementing and testing procedures was particularly beneficial. The ability to create a prototype UI deployed on our local host and Heroku added a practical dimension to the project. The Utilisation of authentic datasets for this endeavour was incredibly rewarding, especially considering my limited exposure to data science, primarily through a data mining module in my background. While the module lacked substantial coding components, this project compelled me to enhance my Python proficiency and gain a deeper understanding of the strategic use of various libraries. Overcoming numerous challenges was an integral part of the process and leveraging data science forums and documentation proved invaluable in exploring potential solutions. Importantly, the experiment successfully fulfilled its aims and objectives, significantly contributing to my skill set and professional development in the field of data science.

#

**References and Bibliography**

* [Bekesiene](https://sciprofiles.com/profile/921176?utm_source=mdpi.com&utm_medium=website&utm_campaign=avatar_name), S., [Smaliukiene](https://sciprofiles.com/profile/908270?utm_source=mdpi.com&utm_medium=website&utm_campaign=avatar_name), R. and [Vaicaitiene](https://sciprofiles.com/profile/author/RU9PS2doSFFBQllDYm1LUkU1VEtNS1VzdFVaSlVqRGc5d25YK3lxWHVqVT0=?utm_source=mdpi.com&utm_medium=website&utm_campaign=avatar_name), R. (2021) “Using Artificial Neural Networks in Predicting the Level of Stress among Military Conscripts”. *[Online]*. Available at: <https://doi.org/10.3390/math9060626>.
* Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique.". *Journal of Artificial Intelligence Research*. 16, pp. 321-357.
* Colreavy, E., Lewandowsky, S. (2008) “Strategy development and learning differences in supervised and unsupervised categorization”. *Memory & Cognition.* 36, pp. 762–775.
* Cortes, C. and Vapnik, V., (1995) “Support-vector networks”. *Machine learning*. 20, pp.273-297.
* Coussement, K., Lessmann, S. and Verstraeten, G. (2017) “A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry”. *Decision Support Systems*. 95, pp.27–36, Available at: https://doi.org/10.1016/j.dss.2016.11.007.
* Dalvi, P., Khandge, S., Deomor, A., Bankar, A., and Kanade, V. (2016) “Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression”.
* Gaur, L., Singh, G., Solanki, A., Jhanjhi, N.Z., Bhatia, U., Sharma, S., Verma, S., Petrović, N., Muhammad, F.I. and Kim, W., (2021) “Disposition of youth in predicting sustainable development goals using the neuro-fuzzy and random forest algorithms”. *Human-Centric Computing and Information Sciences*. 11.
* Geurts, P., Ernst, D. and Wehenkel, L. (2006) “Extremely randomized trees”. *Machine Learning*. 63(1), pp. 3-42
* He, H. (2022) “Bank Customer Churn Prediction Analysis Based on Improved WOA-SVM”. *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*. Available at: https://doi.org/10.1109/iwecai55315.2022.00093.
* IBM. (2023) “What is Random Forest?”. *IBM [online].* Available at: https://www.ibm.com/topics/random-forest [Accessed: 14/11/2023].
* IBM. (2023). “What are Neural Networks?”. *IBM [online].* Available from: <https://www.ibm.com/topics/neural-networks#:~:text=Neural%20networks%2C%20also%20known%20as> [Accessed: 14/11/2023]
* IBM. (2023). “What is Supervised Learning?”. *IBM [online]*. Available from: https://www.ibm.com/topics/supervised-learning [Accessed: 14/11/2023].
* Karvana, K., Yazid, S., Syalim, A. and Mursanto, P. (2019) “Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry.”.
* Kaur, I. and Kaur, J. (2020) “Customer Churn Analysis and Prediction in Banking Industry using Machine Learning.”. *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. Available at: https://doi.org/10.1109/pdgc50313.2020.9315761.
* ‌Lee, H.K. and Kim, S.B. (2018) “An overlap-sensitive margin classifier for imbalanced and overlapping data”. *Expert Systems with Applications [online].* 98, pp.72–83. Available at: https://doi.org/10.1016/j.eswa.2018.01.008.
* Liaw, A., & Wiener, M. (2002). "Classification and Regression by randomForest.". *R News.*  2(3), pp. 18-22.
* Love, C., Medin, L. and Gureckis, M. (2004). “SUSTAIN: A network model of category learning”. *Psychological Review*. 111, pp. 309-332.
* ‌Moro, S., Rita, P., and Cortez, P. (2012) “Bank Marketing”. *UCI Machine Learning Repository [online].* Available at: https://doi.org/10.24432/C5K306.
* Park, Y., -S and Lek, S. (2016) “Artificial Neural Networks: Multilayer Perceptron for Ecological Modelling”. *Developments in Environmental Modelling [online]*. 28 (ISSN 0167-8892), pp.123–140. Available at: https://doi.org/10.1016/B978-0-444-63623-2.00007-4.
* Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Varoquaux, G., Gramfort, A., Thirion, B., Dubourg, V., Passos, A., Brucher, M., and Perrot Andédouard Duchesnay, M. (2011) “Scikit-learn: Machine Learning in Python”. *Journal of Machine Learning Research* *[online].* 12, pp.2825–2830.
* Piri, S., Delen, D. and Liu, T. (2018) “A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets”. *Decision Support Systems*. 106, pp.15–29.
* Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). "Rotation Forest: A new classifier ensemble method.". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 28(10), pp. 1619-1630.
* Rudd, D. H., Huo, H., and Xu, G. (2021) "Causal Analysis of Customer Churn Using Deep Learning,". *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*. pp. 319-324, Available at: https://doi:10.1109/DSInS54396.2021.9670561.
* Scikit-Learn (2011) “RBF SVM parameters”. *[online].* Available at: <https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html#:~:text=The%20gamma%20parameters%20can%20be>. [Accessed: 14/11/2023]
* ‌Sharma, A., Kaur, A. and Semwal, A. (2022) “Supervised and Unsupervised Prediction Application of Machine Learning”. *International Conference on Cyber Resilience (ICCR)*. pp. 1-5.
* Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007) "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.". *BMC Bioinformatics*. 8, 25.
* Wu and Chang. (2003) “Adaptive feature-space conformal transformation for imbalanced-data learning”. *International Conference on Machine Learning*. pp.816–823.
* Wu, Lin and Weng. (2004) [“Probability estimates for multi-class classification by pairwise coupling”](https://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/svmprob.pdf). *Journal of Machine Learning Research*. 5, pp. 975-1005.
* Wu, S., Yau, W. -C., Ong, T. -S. and Chong, S. -C. (2021) “Integrated Churn Prediction and Customer Segmentation Framework for Telco Business.”. *IEEE Access*. 9, pp.62118–62136, doi: <https://doi.org/10.1109/access.2021.3073776>.

# **Appendices**

**A screenshot of a project

Description automatically generatedAppendix 1a**: Gantt Chart for the IPR

**A white screen with multiple colored lines

Description automatically generated with medium confidenceAppendix 1b**: Gant Chart for the Final Report

**A screenshot of a computer program

Description automatically generatedAppendix 2a**: Version Control

**Appendix 2b**: Version Control Final Report

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated**Appendix 3**: System Specifications

A screenshot of a computer program

Description automatically generated**Appendix 4a**: Installation of Libraries

**Appendix 4b**: Installation of Libraries

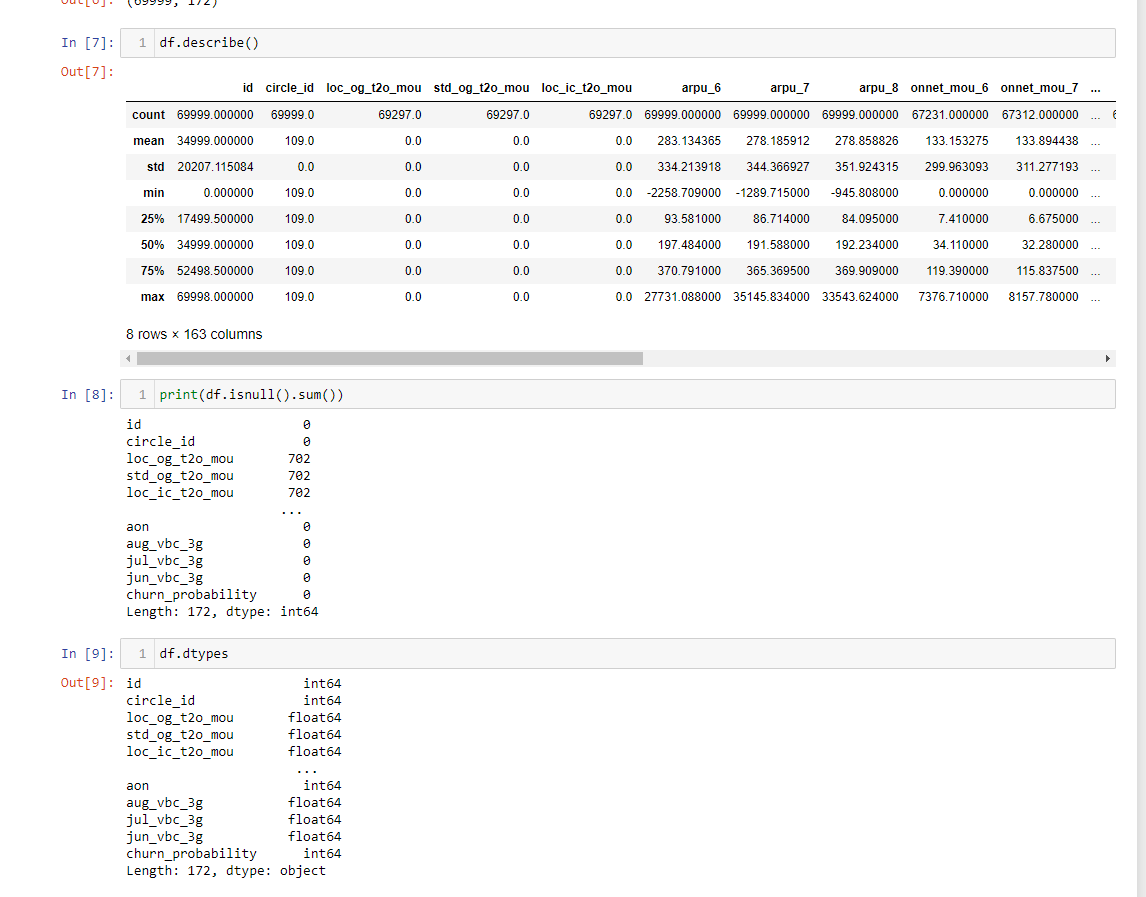
A screenshot of a computer program

Description automatically generatedIn this appendix, we initially wanted to install the Graphviz library to visualise the first decision trees of the RF and the version downloaded was not stable, so we installed another version of NumPy which is dependent on TensorFlow. That created a mess of conflicts which I was able to do by installing the stable version of NumPy and other dependencies required to continue with the project.

**Appendix 5**: Relevant Code and Visualisations for IPR.

A screenshot of a computer

Description automatically generated



:

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generatedAs we can see from the appendix, shows the code where we aggregate data and create a new aggregated column and the shape has changed from 172 attributes to 155. The goal was to reduce the size of the data frame as much as possible.

**Appendix 6a**: The Final Progress Report Relevant Code and Visualisations

Because there are many lines of code in the notebook, I am attaching the repository’s URL for bitbucket cloning it and previewing the code and output. (<https://sc21aea@bitbucket.org/simpsonchiwashira/msc-acs-project.git>)

A screenshot of a computer program

Description automatically generated

**Appendix 6b**: Performances of other models for Telco

**A graph of different colored bars

Description automatically generated**

A graph of different colored bars

Description automatically generated

A graph of different colored bars

Description automatically generated

A graph of different colored bars

Description automatically generated

A graph of different colored bars

Description automatically generatedA graph of different colored bars

Description automatically generated

**Appendix 7:** Design and deployment of the model on a UI with Heroku

A screen shot of a computer program

Description automatically generatedThe app file will preprocess the data we get from the user on the web browser and then the saved model will be utilised to predict with confidence level for the likelihood of the customer to leave. Below is the snippet of the Python code which utilises the flask framework for this integration.

A screen shot of a computer

Description automatically generatedThe HTML file will receive the information from the user and send it to the app.py for backend preprocessing. The following shows a code Snippet of HTML Code.

A computer screen with many small white text

Description automatically generatedAfter all was designed, we used the Heroku CLI to push the developed code although we face some challenges that include dependencies needed by the environment. Below is a snippet of the Terminal during the deployment.

A screenshot of a computer

Description automatically generatedThe following is a snippet of the model after being deployed and ready for use.

**Appendix 8: Evidence of Supervisory Meetings**

A close-up of a document

Description automatically generatedThe below shows an introductory meeting and the meeting notes. This is where we discussed what is expected of the project.

A close-up of a letter

Description automatically generatedThe below figure shows the next meeting after the initial meeting.

Meeting with Supervisor: John Sapsford

Date: 21 November 2023

Timeslot: 10:00hrs – 10:30hrs

Venue: Online Meeting, MS Teams

Main points

The main aim of the meeting was to get feedback on the Final Progress Report as the submission date was closer. The main points that were touched are the following.

* Clarity in the introduction, the main part had ideas jumbled around and they needed to be rearranged for cohesion.
* There are two RQs which are similar, and a decision must be made to rephrase them and make sure they are asking the right question that can be answered by the investigation.
* The literature review has a sentence which is incomplete and not clear.
* Explain what some of the jargon used means in the lit review.
* Cite and reference all the equations used in the report.
* Explain why I change datasets and adopt the notion of using two instead.
* Make sure all appendices are referenced and cited.

Meeting with Supervisor: John Sapsford

Date: 17 October 2023

Timeslot: 10:00hrs – 10:30hrs

Venue: Online Meeting, MS Teams

Main points

The meeting aimed to discuss the progression of the report to look at the literature review and methodology.

* The major point was to deal with the resampling techniques to be applied. The idea was not to dwell on one point and not answer the RQ.
* There was also a point raised to make sure challenges like these are incorporated into the report.
* To make sure the methodology section has cohesion.

Meeting with Supervisor: John Sapsford

Date: 03 October 2023

Timeslot: 10:00hrs – 10:30hrs

Venue: Online Meeting, MS Teams

Main points

The main aim of the meeting was to talk about the FPR Draft with the lit review beginning to shape up.

* The equations were structured as if they were referenced using the number system and it had to be changed to match the referencing system used at UH.
* In the background section had to be clear if I justified the reason for the need for the project or found the importance of the need for the project.
* References to be in Harvard format and in the bibliography section to remove any numbering and use bullet points to avoid any confusion with the markers when referencing.
* Explain why the need for the resources mentioned in the IPR.
* The legal, professional, and social issues had to be addressed and I made sure I understood the need by touching on each aspect. GDPR, BSC and ACM code of conduct to be looked at and what they say in terms of the project.

Meeting with Supervisor: John Sapsford

Date: 20 October 2023

Timeslot: 11:30hrs – 12:00hrs

Venue: In-person – Café Rore Main Reception, College Lane Campus

Main points

This was a routine check-up soon after the IPR and a discussion of the journey ahead.

* Discussed the IPR what it lacked and what could be used for the FPR.
* The discussion of the datasets if I had made up my mind and what the plan was.
* Showed the supervisor some of the progress with the notebooks shaping up.
* Talked about time not being enough with the number of algorithms chosen for evaluation.