Prediction of Customer Churn in the Telecom Industry (Using Machine Learning)

Khubim Kumar Chhetri
Student ID D00251757

Under the supervision of

Dr. David O'Keeffe

Master of Science in Data Analytics

Dundalk Institute of Technology, Dundalk

DUNDALK
INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA
DHÚN DEALGAN

School of Informatics and Creative Arts
Department of Computing Science and Mathematics

ABSTRACT

With the advent of increasing competition in the telecom industry, companies must retain customers to maximise profits. With an average rate of churn of 30%, customer retention policies affect the annual turnover drastically. The cost of customer churn to the telecom industry is about $10 billion per year globally. 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 effective customer relationship management processes and consumer-centric policies can help reduce spend on customer relations. Thus, one would need to understand and track customer behaviour to understand the indicators that make a customer likely to churn.

Harnessing valuable data for business intelligence to develop churn management strategies is a proven data-driven strategy. Machine learning models require modest computation power and can deliver high accuracy when it comes to predicting attrition. Accurate predictions coupled with business understanding from interpretable machine learning can revolutionize the telecom industry.

This research intends to build a predictive framework that can predict churn accurately and identify behaviour patterns using interpretable machine learning models that predict customer churn. The paper will showcase the performance of various machine learning algorithms and how the process can be optimised. The dataset to be used for this research paper is the IBM Watson Dataset on customer churn in the Telecom industry. Extensive feature selection, processing, model tuning, and interpretable machine learning can help identify churn accurately.

Keywords: Machine Learning, Churn, Classification

Contents

ABSTRACT..

ACKNOWLEDGEMENTS ....

LIST OF TABLES ....

LIST OF FIGURES ....

LIST OF ABBREVIATIONS ....

CHAPTER 1: INTRODUCTION..

1.1 Background of the Study ....

1.1.1 Churn Analysis in the Telecom Industry ....

1.1.2 Customer Retention policies..

1.2 Struggles of the Telecom Industry...

1.3 Problem Statement...

1.4 Aim and Objectives...

1.5 Research Questions...

1.6 Scope of the Study ....

CHAPTER 2: LITERATURE REVIEW ....

2.1 Introduction and Related Research Publications ....

2.1.1 Feature Engineering for Telecom Datasets...

2.1.2 Summary of Literature Review...

2.2 Discussion...

2.3 Summary...

CHAPTER 3: RESEARCH METHODOLOGY ....

3.1 Introduction and Data Understanding ....

3.2 Research Methodology ....

3.2.1 Data Selection ....

3.2.2 Data Preprocessing ....

3.2.3 Data Transformation ....

3.2.4 Data Visualization ....

3.2.5 Model Building ....

3.2.5.1 Model Selection Technique……………………………………………………

3.2.6 Model Evaluation ....

3.2.6.1 Evaluation Metrics…………………………………………………………….

CHAPTER 4: ANALYSIS ....

4.1 Introduction aand Dataset Description..

4.2 Exploratory Data Analysis ....

4.2.1 Analysis of variables ....

4.2.2 Outlier Analysis ....

4.3 Univariate Analysis……………………………………………………………..

4.3.1 Relation with Target Variable……..……………………………………………

4.3.2 Distribution of Variables with respect to Churn……………………….……………

CHAPTER 5: RESULTS AND DISCUSSIONS ....

5.1 Interpretation of Visualisations..

5.4 Model Results ....

5.5 Model Interpretation ....

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS..

6.1 Discussion and Conclusion...

6.2 Future Recommendations…………………………………………………………………
Error! Bookmark not defined.

REFERENCES...

ACKNOWLEDGEMENTS

I would like to acknowledge Dundalk Institute of Technology for the opportunity to learn and obtain a renowned degree. I want to express my heartfelt gratitude to my thesis supervisor, Dr.David o’Keeffe, for his invaluable guidance. He has guided and encouraged me to be professional even when the going gets tough, and I am fortunate to have him as a mentor.

I would like to thank my committee members and mentors from Dundalk Institute of Technology,Dundalk,Ireland for their patient advice and guidance through the research process.

Finally, I thank my family, who supported me with love and understanding. Without you, I could have never reached this current level of success. Thank you all for your unwavering support.

LIST OF TABLES

Table 1: Literature Review for IBM Watson Telecom Dataset……………………………………

Table 2: Model Results of Individual and Ensemble Models……….……………………………….

LIST OF FIGURES

Figure 1: Most significant challenges faced by the industry ....

Figure 2: Types of Churners ....

Figure 3: Visual Data Exploration ....

Figure 4: Model Building Process ....

Figure 5: Distribution of variables (by percentage) ....

Figure 6: Boxplots of Churn versus Total Charges and Churn versus Monthly Charges ....

Figure 7: Scatter plot of Monthly Charges versus Total Charges ....

Figure 8: Univariate Analysis of numerical features of IBM Teleco Data ....

Figure 9: Churn analyzed with Internet, Streaming and Contract ....

Figure 9.1: Distribution of Demographic Attributes with respect to Churn ....

Figure 9.2: Distribution of all features with respect to Churn ....

Figure 9.3: Feature Selection using Gradient Boosting Classifier ....

Figure 9.4: Feature Selection using Gradient Boosting Classifier and Light GBM ....

LIST OF ABBREVIATIONS

AdaBoost	Adaptive Boosting
ANOVA	Analysis of Variance
AUC	Area under ROC Curve
CRM	Customer Relationship Management
CV	Cross Validation
EDA	Exploratory Data Analysis
GBM	Gradient Boosting Machine
KNN	K Nearest Neighbour
SVM	Support Vector Machine
SVC	Support Vector Classification
XGBoost	Extreme Gradient Boosting

CHAPTER 1: INTRODUCTION

With the increase in the number of consumers in this Digital Age, a company needs to keep costs low and profits high to be successful. One of the most effective ways to do this is to retain the existing customer base and focus the remaining budget on acquiring new customers.

Churn in telecom companies is defined as the customers who stop using their specific services and plans for long periods . The Telecommunication sector has becoming one of the main industries in developed and developing countries . The Technical progress and the increasing number of operators has raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies. Three main strategies have been proposed to generate more revenue , acquire new customers and upsell the exists.Getting new customers is much more expensive than retaining existing ones. Some studies have shown that it costs six to seven times more to acquire a new customer than to keep an existing one.

In this ongoing pandemic age, where virtual presence via calls and mobile is on high priority, customers maintain their expenditure for the cause. All the industries tries to offer minimal prices and add-on services to customers to switch . Companies turn quick profit after acquiring customer base ,which later identify the people bracket which are likely to leave and runs campaigns to offers a more value to the customers in the minimal budget that would be very profitable in the long run.

1.1 Background of the Study

In the competitive telecommunication space, if companies do not adapt to optimize existing resources to increase profits, it is tough to thrive in the future. One of the most effective ways to do this is to retain the existing customer base and focus the remaining budget on acquiring new customers. The retention of the existing customer base in a focused and systemic manner is to be done, or the 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 of churn. Based on customer behaviour and attributes, the likelihood of customers to churn can be flagged, and targeted campaigns can be run to retain customers (Jain et al., 2021).

1.1.1 Churn Analysis in the Telecom Industry

The ability to retain customers showcases the company's ability to run the business. In the digital age, where everything is going online, any business needs to virtually 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 customer retention rates were increased by just 5%, profits could see an increase from 25% to even 95% (Hadden et al., 2006). For a company to be profitable and increase the capital, it is essential to take pre-emptive action to retain customers that may churn. Churn is defined as customers who stop using their specific services and plans for long periods. Churn can occur due to various reasons and can be broadly classified as voluntary and involuntary churn.

In this ongoing pandemic age and post pandemic period, where virtual presence via calls and mobile is on high priority, customers maintain their expenditure for the cause or try to reduce their monthly expenditure. All the industries or competitors tries to offer minimal prices and add-on services to customers to switch telecom operators. Companies turn quick profit after acquiring a significant customer base, the companies monetise their customer base and profit in the long term (Jain et al.,2021). The companies that identify the customer bracket which are likely to leave and runs targeted campaigns to offers a more value to the customers in the minimal budget that would be profitable in the long run.

1.1.2 Customers 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. This increase in competition has led customers to expect tailor-made products at a fraction of the price (Kuo et al., 2009). Churned customers move from one service provider to another (Ahmad et al., n.d.) (Andrews, 2019). Customer churn can be due to the non-satisfaction of current services, better offerings from other service providers, new industry trends and lifestyle changes. Companies use retention strategies (Jahromi et al., 2014) to maximise customer lifetime value by increasing the associated tenure. For telecom companies to reduce churn, it is vital to analyse and predict key performance indicators to identify high-risk customers, estimated time to attrite and likelihood to churn.

1.2 Struggles of the Telecom Industry

The telecom industry has been struggling for years now for the survival. Telecom businesses have struggled to launch their numerous services or products annually. The Huthwaite study shows that telecom companies have at least a new product failure annually, costing millions of dollars annually. Rather than developing strategies that meet evolving customer needs, telecom operators follow the traditional cycle of setting up networks, building cross-channel presence, and offering revamped plans. The losses, as seen by the industry, highlights the fundamental flaw in the approach. A study by Capgemini showed that most companies showed a Net Promoter Score between zero and negative (Why is the telecom industry struggling with product success?, 2021). The telecom industry is rife with disruption in all areas. The pandemic has changed how everyday communication supplements and enhances discussion between customers and brands.

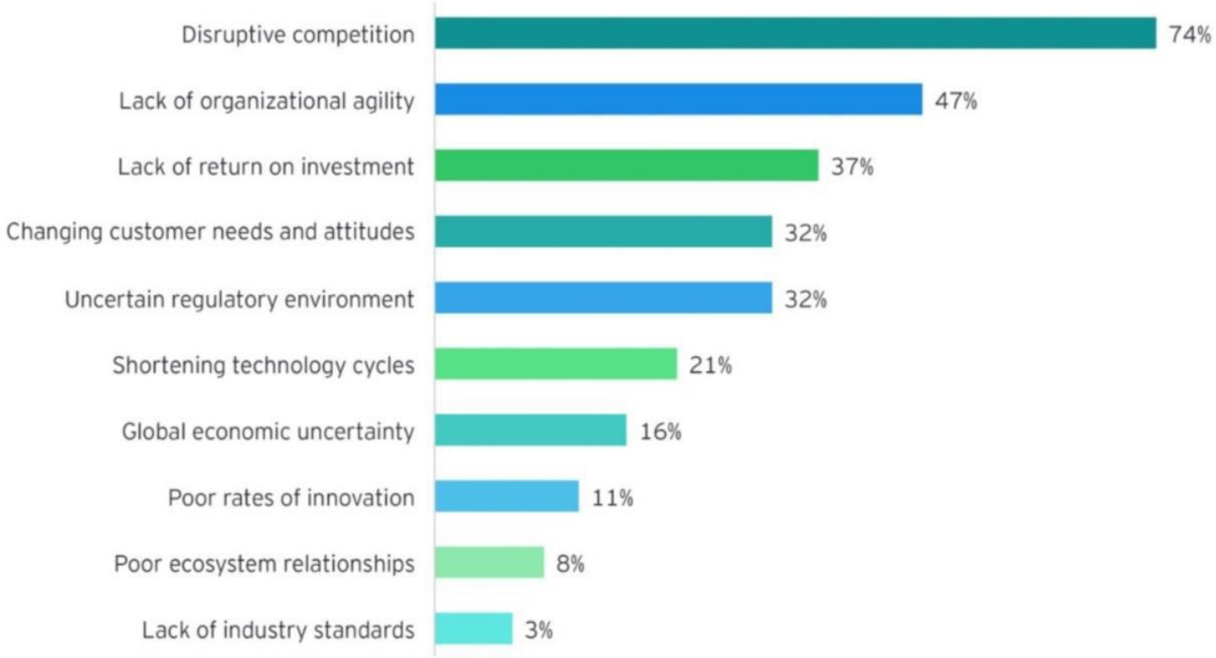


Figure 1: Most significant challenges faced by the industry,
Source: (Digital transformation for 2020 and beyond eight telco considerations, 2021)

Disruptive competition is the primary reason why telecom operators are struggling globally. Customer attrition is the main reason to track at-risk customers that may churn and target programs to retain them. This targeted effort will help retain customers and ultimately increase the telecom company's profits by employing churn prediction strategies.

1.3 Problem Statement

The reduction of attrition of the customers is vital to the company's Reputation and the features & the services they offering. To maintain a good market share in the competitive telecommunication industries must understand and tackle the root cause of why a customer might shift their service provider. This research will help telecom companies leverage their existing consumer database to predict and actively target campaigns to customers likely to churn. The machine learning methodology employed can be personalized to the use case based on the operator. When a suitable set of machine learning algorithms run on a newer dataset, the model's evaluation metrics can be monitored, and high-risk customers can be appropriately targeted.

The recommended model's primary users will be telecom conglomerates that wish to reduce customer attrition and improve their profitability in the market. This needs to be done, keeping in mind the overhead costs.

1.4 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 analyse the relationship and visualise patterns of customer behaviour to make aware 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 of consumers that leads to customer attrition in the telecom industry
- To evaluate the performance of the models to identify the appropriate models

1.5 Research Questions

The following research questions have been formulated based on the literature review done so far in the field of customer churn:

- Is there a clear conclusion regarding the best overall modelling approach, be it classical machine learning or more complicated algorithms?
- Does the presence of multicollinearity, outliers, or missing values in the training data impact customer churn prediction accuracy?
- Do techniques such as hyperparameter tuning result in significantly better models?
- Can balancing techniques be suggested to increase the accuracy of the model?
- Are the results obtained from interpretable models reliable?
- Do statistically significant features mean that the business can take actionable insights directly?

1.6 Scope of the Study

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

- The data for the study has directly been obtained from the authorised 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 focus of the research is on interpretable models. If time permits, an attempt to use other models to perform customer attrition analysis can be made .

CHAPTER 2: LITERATURE REVIEW

A thorough survey of the research and work done in customer attrition in the telecom industry will help us understand more about the telecom industry's nuances. This literature review will set the baseline to understand the expected standard to implement a robust classification model to predict customers' high risk of churn in the telecom industry. The approaches used by the authors range from using single machine learning models, meta-heuristic models, hybrid models, data mining techniques and even social methods (Oskarsdottir et al. 2016).We have given as much to conventional methods that have solved the problem to churn and given weightage to the novel methods that deal with churn.

With the advent of the massive investments from telecom operators in this internet age, the market has become the most competitive in the decades. This Literature Review would more focus on the telecom industry’s ongoing trends and how data analytics affect the telecom industry. Customers have moved from expecting just the cheapest plans, the average customer now expects to have tailor made plans and solutions at a fraction of the cost that their monthly bill used to be (Umayaparvathi and Iyakutti 2012).

Customers no longer need to stick to a monthly commitments of a subscribed plan, they can quickly get the benefits of the company’s infrastructure within minimal commitments using a prepaid plan rather than a postpaid plan. On average. A telecom company loses 30% of its customer base annually, not all customers can be stopped from churning (Umayaparvathi and Iyakutti 2016). There are categories of customers that leave voluntarily and involuntarily,also among the churners that leave voluntarily there is a further division of those that attrite deliberately and incidently.

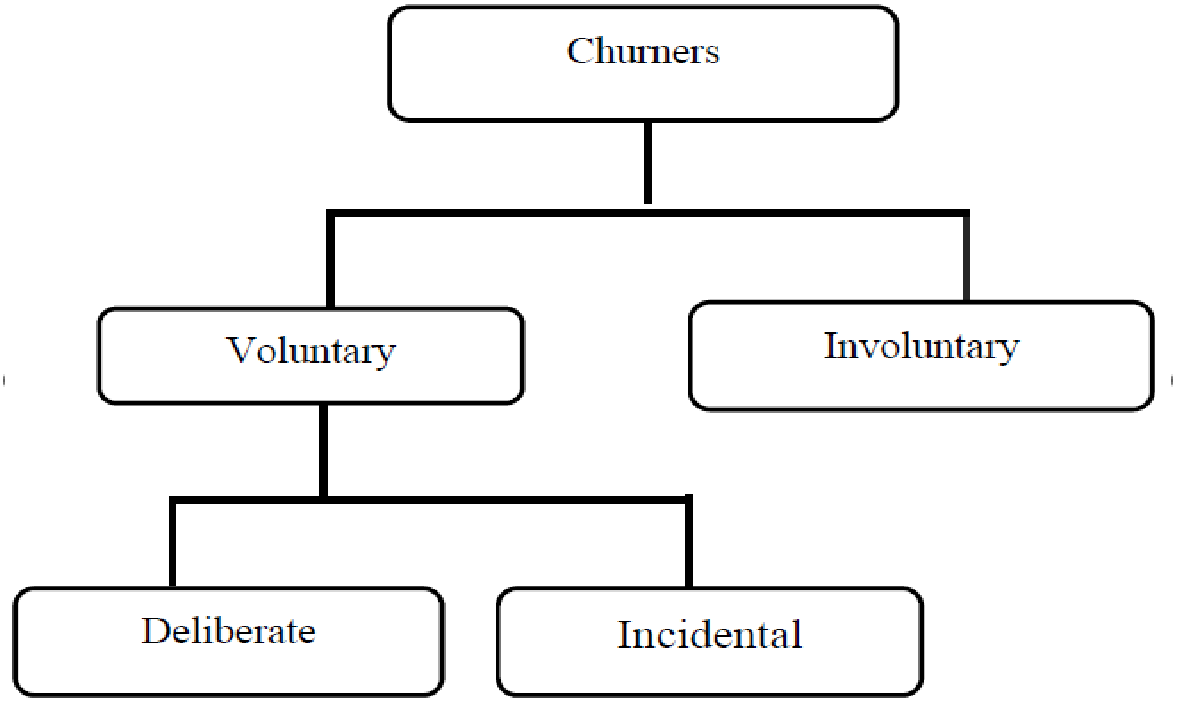


Figure 2: Types of Churners (Saraswat, S. & Tiwari, 2018)

The telecom industry might seem like it is booming with the internet age, but that is not the case for most telecom operators. The telecom industry has a heavy dependency on external factors riddled with serious debt complications in the industry. The investments range from building infrastructure that can carry lines across the country, investments in the latest technologies that will help enable the latest in voice and internet technology like 5G, money spent on buying bandwidth frequencies. Additionally, the cost of upkeep and maintenance of a vast network can be grossly expensive as operators have to pay rents, keep up the set infrastructure, lobby the government, provide customer service, and deal with the unexpected changes in the ecosystem. For all of these risks that telecom operators take to run effectively, various business models can ensure a steady income.

Since a Business to Consumer (B2C) model is high-risk and high-reward, ensuring that there are guaranteed paying customers at the end of the month can be crucial to maintain market share. The telecom sector's riskiest customers are the prepaid customers, as it is challenging to flag if they are active or not because different segments of customers have different behavioural patterns. The telecom industry has truly earned its place as the backbone of our country and even the economy. It is exceedingly difficult to imagine a world in which a call, message or communication with someone at a fraction of the cost paid for the same service about just a decade ago. The rate of mobile and internet penetration in third-world countries is increasing exponentially; this leads to a whole host of some of the largest companies in the world backing up telecom operators to be able to acquire a customer base as loyal and dedicated as possible so that this cash-burn can be leveraged to profit in the future.

To have a higher stake in the Industrial Revolution 4.0, telecom operators need to move away from a conventional customer retention approach. A customer is no longer associated with a company because only one service provider exists in the area. The telecom operators should improve their CRM infrastructure to move from merely fulfilling an internal need to a full-fledged ecosystem with value-proposition for the end-customers and all stakeholders involved telecom pipeline. A happy customer is a loyal one. Attracting new customers might seem like an attractive way to grow market share. However, the experienced players in the market know that the secret to being profitable in the long run is two-fold, first, focusing on the retention of customers, especially the high-value customers and second, being able to leverage the existing database that is a trove of customers who are likely to come back to the company if courted aptly. Gaining new customers is 5 to 10 times more expensive than keeping existing customers loyal (Wassouf et al., n.d.; Ebrah and Elnasir, 2019). The recommended method to effectively implement a data science predictive framework is to scale and leverage it to make a robust and effective model as a custom-designed use case. A custom solution in terms of strategy is one where leadership would invest less effort on a proof of concept and leverage the long-term benefits for the company - if the project can help increase the profits in the long term. The idea of investing in the future to move from a model that reduces loss to increases profit is a game-changer. Several low-code or no-code tools are being used to build proof of concept projects; the reality is that implementation is vital. Models need to focus on explainability and usage of metrics rather than a black-box approach.

This is critical to build a solid data science exercise within the organization because it may be easier and even faster to build a proof of concept with a ready-made tool or technology. However, when it comes to scaling the exact implementation at an organization-wide level whilst keeping the overhead costs minimal, it can get complicated. Implementation on a large scale has two problems. First, it may be costly to get multiple licences or pass large amounts of data in the tool. Secondly, there may be a black-box approach for the data problems, so modifying the code may not be feasible. Tools such as RapidMiner that can leverage explainable models that can be understood by senior management can be a good starting point (Halibas et al., 2019) for proof of concept implementations. Developing an in-house custom analytics solution is the long-term aim of a company and building data science competencies. Most companies require a custom setup for churn analysis on different datasets, technology stacks, databases and overall requirements (Fonseca Coelho, n.d.). Understanding the requirement for the cadence of forecasting based on the model selected is also a vital area of research to move from a batch-processing system to a more real-time system (Tamuka and Sibanda, 2021). Depending on the complexity of requirements and budget, a cloud-based flexible architecture can also be set up.

A predictive modelling framework for data science involves a list of tasks that can be understood through the literature survey. In this section, let us understand the details of the supervised machine learning techniques. Customer churn analytics in the telecom industry aims to flag the segment of customers likely to churn. This classification problem predicts one of two things; if a customer will churn or not. There are different methods to do this, and in the literature review below, an understanding of supervised machine learning algorithms will be given. The fusion of multilayer features uses a framework of complementary fusion by employing feature construction and feature factorisation to improve churn prediction accuracy. This approach resolved the problem of high dimensionality and imbalance of data. Feature selection was also attempted, which led to the reappearance of imbalanced data (Ahmed and Linen, 2017). Novel methods of engineering the data was also used in the research where tokenisation was used for categorical attributes and standardisation was used to standardise numerical attributes (Momin et al., 2020).

Novel methods for feature selection, such as gravitational search algorithm (Lalwani et al., 2017), have been used. Gravitational Search Algorithm helps reduce the dimensionality of the data and improves the data's accuracy by optimising the search for significant features (Lalwani et al., 2021). Methods for pre-processing data tasks such as missing value imputation have developed well over the last few years. A method used to explore and perform multiple missing value imputations to fill up quantitative variables that suffer from an uneven distribution is Predictive Mean Matching (Mahdi et al., 2020). While some methods are agnostic to the data type, specific methods assess numeric variables' uneven distribution using a logarithmic transformation (Tamuka and Sibanda, 2021). Categorical variables used in telecom datasets are also converted to numeric variables using techniques such as label encoding or one-hot encoding (Agrawal, 2018). The popular methods used to handle categorical variables are label encoding and one-hot encoding. With larger datasets, high dimensionality is a problem – for this, some of the authors with large datasets have worked with sparse matrices or have leveraged dimensionality reduction techniques such as principal component analysis. Some of the authors have leveraged modelling techniques that work with categorical variables, continuous and discrete variables.

For data of any form to be leveraged, understanding the dataset is fundamental. One of the fastest ways to perform exploratory data analysis is to visualise the data. Figure 2 illustrates the relationship between data, visualisation and models with the intermediary knowledge gained from visual analytics (Yuan et al., 2021).

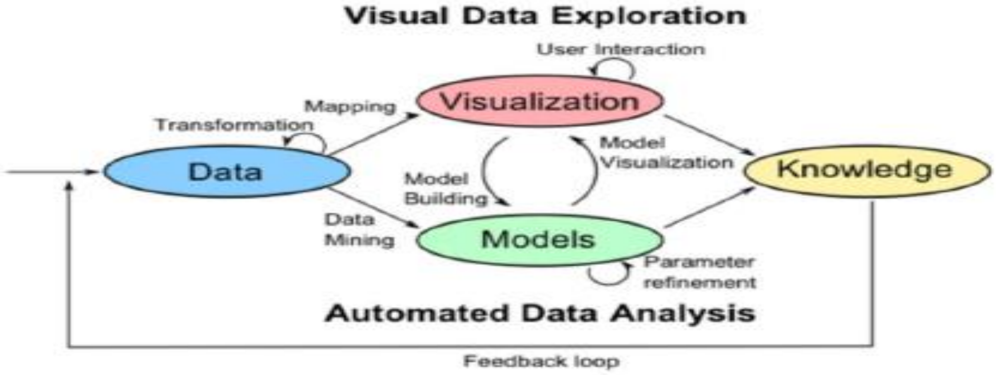


Figure 3: Visual Data Exploration

Being able to perform automated data analysis is the essence of visual data exploration. Based on the visualisations formed, further understanding of row-level data is developed. When data transformation is performed, visualising the data post-processing helps understand if further data manipulation is required before the modelling phase. For instance, feature importance using a method out of advanced regression, XGBoost or random forest has been calculated.

At the same time, the visualization and sum of feature importance scores obtained for features visually, the identification of the top feature using a bar chart with an indication of the top features to choose from for the next steps. Using multiple methods of visualising the features' distribution, the variance of the data points and the other analysis helps us make decisions for the next steps.

2.1 Related Research Publications

This section will review how data analytics is used in the telecom industry to identify customers at a high risk of attrition and the data-driven processes followed to set the baseline of the techniques carried out in the industry far. So the below sections will focus on feature engineering for the data and handle class imbalance. Efficiently carrying out data pre-processing will help us obtain better results in the following stages of implementing machine learning and validation via k-fold cross-validation. In the literature review, an understanding of the evaluation methodology used to assess the models' performance will be analysed. Section 2.6.3 will review the evaluation metrics used for classification (Karimi et al., 2021).

2.1.1 Feature Engineering for Telecom Datasets

Feature engineering is a critical step in the data science flow. Based on the analysis of the existing techniques implemented by authors, the significant features from the dataset that can affect churn are picked or generate new features from the existing set of attributes that can help predict churn better. When the authors have set out to perform feature engineering, keeping the dataset and the predicted model's accuracy in mind is only done. When performing feature engineering on a dataset, another critical task is identifying the attributes that have the highest impact on the target variable. This can be done by leveraging rigorous algorithms or even RapidMiner and Azure ML Studio (Thontirawong and Chinchanachokchai, 2021).

Feature selection is made using methods such as random forest, xgboost and advanced regression, based on which the less significant values are discarded and the effect on the accuracy of churn prediction is observed. Techniques that leverage the correlation with the target variable are also used; the correlation matrix operator (Halibas et al., 2019) performs feature selection, and less significant features were discarded. The scoring of features based on their relation to the target variable indicates the variable's feature importance in consideration. Since the data has been generated from various sources and periods, standardisation of the data to compare different sets effectively helps the author decide the essential features based on the correlation matrix operator. The operator produces a pairwise table of correlation coefficients.

2.1.2 Summary of Literature Review

The telecom industry is a competitive space, and authors have been trying to solve customer attrition for years. There are multiple ways to tackle churn and as machine learning advances, so do the methods by which a customer that may leave is flagged. The data present within a company is a golden opportunity to build a robust model that can be leveraged to increase profitability. There has been some stellar research in classification, from single machine learning models to hybrid models (Induja and Eswaramurthy, 2015). Recent literature has a significant impact on the modelling of customer attrition in the telecom industry. Being able to view all of the work in the form of the below table gives us an overview of the significant work that has been done to support the same. More importance can be given to feature engineering from the above section, as most papers have used more conventional methods. Similarly, for class balancing, instead of opting for simple random oversampling techniques, other structured oversampling techniques can be leveraged for the next steps.

A whole host of machine learning models can be used for the use case of solving the classification of high-risk customers. An excellent approach to focus on the machine learning approach and the data pre-processing. A few authors implemented class balancing techniques, and better accuracy was observed. Our approach will be made on all of the steps mentioned above of data pre-processing, missing value analysis, outlier analysis, variance analysis, k-fold cross-validation and class balancing techniques for phase 1. This will be followed by single machine learning algorithms and hybrid machine learning models in phase 2. Once the best models can be found for our use case, k-fold cross-validation will be performed to get the best generalised and robust model. This thorough literature review of the best the academic community offers has provided us with a baseline understanding before deciding the appropriate research methodology for our use case.

2.2 Discussion

From the above literature review, there are various ways to identify the customers at a high risk of churn. The problem's approach varies from data mining techniques to selecting the right set of attributes, valuable data pre-processing, and efficient feature selection. This effort to obtain the right set of data to feed results in choosing a simpler model to perform classification; thus, saving computation time and keeping the overall computational requirements minimal, saving companies' overhead costs.

The other approach is to rely on the machine learning model to flag the customers that are likely to churn effectively. The data size plays a considerable role; if the data's size is limited, focusing on the machine learning algorithm is more sensible, whereas a hybrid approach can be experimented with for larger datasets. The literature on deep learning suggests that even though a neural network approach works for some cases, the model's performance is not significantly better to opt-in for deep learning models exclusively. It is a common misconception that deep learning models perform better than machine learning models in all use-cases. From the literature review, the understanding for telecom use cases studied where a predictive framework based on a machine learning or deep learning framework has been made, hybrid machine learning models and a balancing technique have given the best results.

Different feature selection techniques, in turn, have resulted in a different set of features being selected for different algorithms. Exploring more feature engineering techniques and summarising our results so the observed and latent relationships of the features with the target variables will aid future implementation. Imputation of the data is also a step where some authors have taken advanced methods such as logarithmic transformations and predictive mean matching for imputing missing data rather than the conventional methods to impute the missing values with mean, median or mode (Tamuka and Sibanda, 2021). This approach, along with oversampling techniques, has given some of the best results per the literature survey. In Chapter 3, this is the approach to take inspiration for and more advanced feature selection methods.

Authors	Year	Feature Engineering	Model
(Tamuka and Sibanda, 2021)	2021	Feature Importance, Logarithmic Transformation	**Accuracy**: Logistic Regression - 97.8%, Decision Tree - 78.3%, Random Forest - 79.2% **F1-Measure**: Logistic Regression - 97.8, Decision Tree - 77.9, Random Forest - 77.8
(Lalwani et al., 2021)	2021	*Phase 1*: Variance Analysis, Correlation Matrix, Outliers Removed *Phase 2*: Cleaning & Filtering *Phase 3*: Feature Selection using Gravitational Search Algorithm, Feature Importance	**AUC**: Logistic regression - 0.82, Logistic Regression (AdaBoost) - 0.78, Decision Tree - 0.83, Adaboost classifier - 0.84, Adaboost Classifier (Extra Tree) - 0.72, KNN classifier - 0.80, Random Forest - 0.82, Random Forest (AdaBoost) - 0.82, Naive Bayes (Gaussian) - 0.80, SVM Classifier Linear - 0.79, SVM Classifier Poly - 0.80, SVM (Adaboost) - 0.80, XGBoost - 0.84, CatBoost - 0.82
(Momin et al., 2020)	2020	Tokenisation, Standardisation	**Accuracy**: Logistic Regression - 78.87%, Naïve Bayes - 76.45%, Random Forest - 77.87%, Decision Trees - 73.05%, K-Nearest Neighbor - 79.86%, Artificial Neural Network - 82.83%
(Oka and Arifin, 2020)	2020	Label Encoding Binary Columns, Scaling Numerical Columns, Feature Importance	**Accuracy**: Random Forest - 77.87%, XGBoost - 76.45%, Deep Neural Network - 80.62% **AUC**: Random Forest 0.83, XGBoost 0.84, Deep Neural Network - 0.84
(Mahdi et al., 2020)	2020	PMM - Predictive Mean Matching for imputation	**Accuracy**: PPForest with LDA - 72%, PPForest with SVM - 75% **AUC**: PPForest with LDA - 0.67, PPForest with SVM - 0.73

Table 1: Literature Review for IBM Watson Telecom Dataset

2.3 Summary

A whole host of machine learning models can be used for the use case of solving for the classification of high-risk customers. An excellent approach to try would be to focus on the machine learning approach and the data pre-processing. A few authors implemented class balancing techniques, and better accuracy was observed. Our approach will be made on all of the steps mentioned above of data pre-processing, missing value analysis, outlier analysis, variance analysis, k-fold cross-validation and class balancing techniques for phase 1. This will be followed by single machine learning algorithms and hybrid machine learning models in phase 2. Once the best models can be found for our use case, k-fold cross-validation will be performed to get the best generalised and robust model. This thorough literature review of the best the academic community offers has provided us with a baseline understanding before deciding the appropriate research methodology for our use case.

CHAPTER 3: RESEARCH METHODOLOGY

This chapter is dedicated to the research methodology that will be used with the IBM Watson Telecom dataset. The customers at a high risk of churn will be flagged based on the literature reviewed and understanding of the telecom business. The literature review will be applied in this research methodology in data pre-processing, feature engineering, predictive framework, evaluation metrics and interpretable machine learning.

A baseline understanding of how to tackle the customer churn problem in the telecom industry from the literature review will help decide the improvements and modifications that can be made. This section will set up the research methodology for tackling the use-case for our study.

3.1 Data Understanding

There are various data sources used to predict customer churn in the telecom industry through the literature survey. This research shall be using the IBM Watson Telecom churn data found on the Kaggle website derived from the IBM Cognos Analytics Community (Cognos Analytics - IBM Business Analytics Community, 2021). The telecom churn data consists of 7043 rows and 21 attributes at a customer-id level. The data combines 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.

Each row in the telecom churn represents customer attributes used to describe the customer's behaviour. The data is unique at a Customer ID level with a high cardinality of 7043. The Total Charges column is uniquely distributed. There is an equal 50-50 distribution of male and female customers. As one would expect in the Churn column, there is an imbalance, with 27% of customers churning and 73% retention. This dataset has been collected over a month with a Kaggle Usability Score of 8.8 based on the provided metadata and various other factors, as mentioned in the website (Kaggle, 2018).

Let us understand the descriptive dataset statistics in detail. Here, let us analyse and understand the dataset better by deep driving into the statistics of each column:

- Customer ID: Unique Customer Id assigned to each customer (7043 unique values)

- Dependents: Indicative of whether the customer has dependents or not
- Tenure: Number of months the customer has stayed with the company
- Phone Service: Indicative of whether the customer uses the phone service or not
- Multiple Lines: Whether the customer has multiple lines or not
- Internet Service: Information regarding the internet service provider (DSL, Fiber optic, No)
- Online Security: Whether the customer has online security or not
- Online Backup: Whether the customer has opted in for Online Backup
- Device Protection: Whether the customer has open in for Device Protection Plan
- Technical Support: Whether the customer has requested Technical Support
- Streaming T.V.: Whether the customer has opted in for T.V. Streaming services
- Streaming Movies: Whether the customer has opted in for Streaming Movies services
- Contract: Whether the customer has opted for a monthly, annual or two-year plan
- Paperless Billing: Whether the customer has opted in for paperless billing
- Payment Method: Electronic check, Mailed check, Bank Transfer or Credit Card
- Monthly Charges: Monthly Charges of the customer
- Total Charges: The total charges of the customer
- Churn: Whether the customer has churned or not
- Gender: Indicative of whether a customer is male or female
- Senior Citizen: Binary of whether the customer is a senior citizen or not
- Partner: Information on whether the customer has a partner or not

3.2 Research Methodology

The following section contains the steps to perform predictive modelling to predict the customers with a high attrition risk. The steps followed are data selection, pre-processing, data transformation, data visualisation, class balancing, model building, model evaluation and model deployment.

3.2.1 Data Selection

There were a few datasets to choose from when it comes to telecom data. The data selected is the IBM Watson Telco Customer Churn Data.

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

- 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
- 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

3.2.2 Data Preprocessing

Discussion on the data pre-processing steps ensures that the data is standardised when used as input to various models for numerous iterations. A sense check of the telecom churn dataset is performed to understand if the import of the data and the dataset's encoding are per expectations. Once the data types of the features are noted, the shape of the data is checked to ensure the number of rows and columns is consistent per expectations. Focus is then directed on the columns that have at least one missing value. Once the attributes are selected, the percentage of missing values column-wise is analysed. This will help us to decide the next steps. Missing value analysis will determine if all the columns or selected columns will be carried forward to the next step. Suppose columns must be dropped based on absent value percentage methods such as mean imputation, mode imputation, deletion of rows, and iterative imputation can be used.

3.2.3 Data Transformation

Data transformation will be carried out to extract the most value from the dataset where pre-processing steps have been performed. Steps such as one-hot encoding are applied to the categorical features. Besides this, features derived from the existing dataset and feature engineering will occur. Based on the understanding of telecom's business, business rules and heuristics are applied to the business and derive new features.

3.2.4 Data Visualization

Data visualisation is an integral part of exploratory data analysis to be able to understand the data. Visualisation packages to analyse and understand the data such as pandas profiling, sweetviz and data prep can be leveraged. This will help us understand the distribution of the columns, the variance, and the data profile.

3.2.5 Model Building

Model Building is one of the more crucial components of this study. The following steps will help identify the right set of models and appropriate techniques to leverage to get optimal results. This model building is followed by choosing the models to implement after the data cleaning, feature engineering, and data formatting steps.

3.2.5.1 Model Selection Techniques

The best performing models are selected based on multiple factors ranging from accuracy to interpretability. From the literature review, it has been observed that the supervised classifier models have given good results. Single algorithm models are implemented to pick out the models that have the best performance.

3.2.5.2 Model Assessment

For any models to be used by the business, model assessment is a critical part of the process. As models are developed from the perspective of a Data Scientist, the following steps will also ensure that the predictions are as expected for the company to leverage the model. Model interpretability is vital to the business's functioning as they would like to understand the customers that are likely to churn and gain insights. Therefore, in the model assessment stage, the focus needs to be on actionable insights and provide the business with customer behaviour patterns linked to churn's high likelihood. The diagram below highlights the stages to use for the model building process, from the data loading to the final model output.

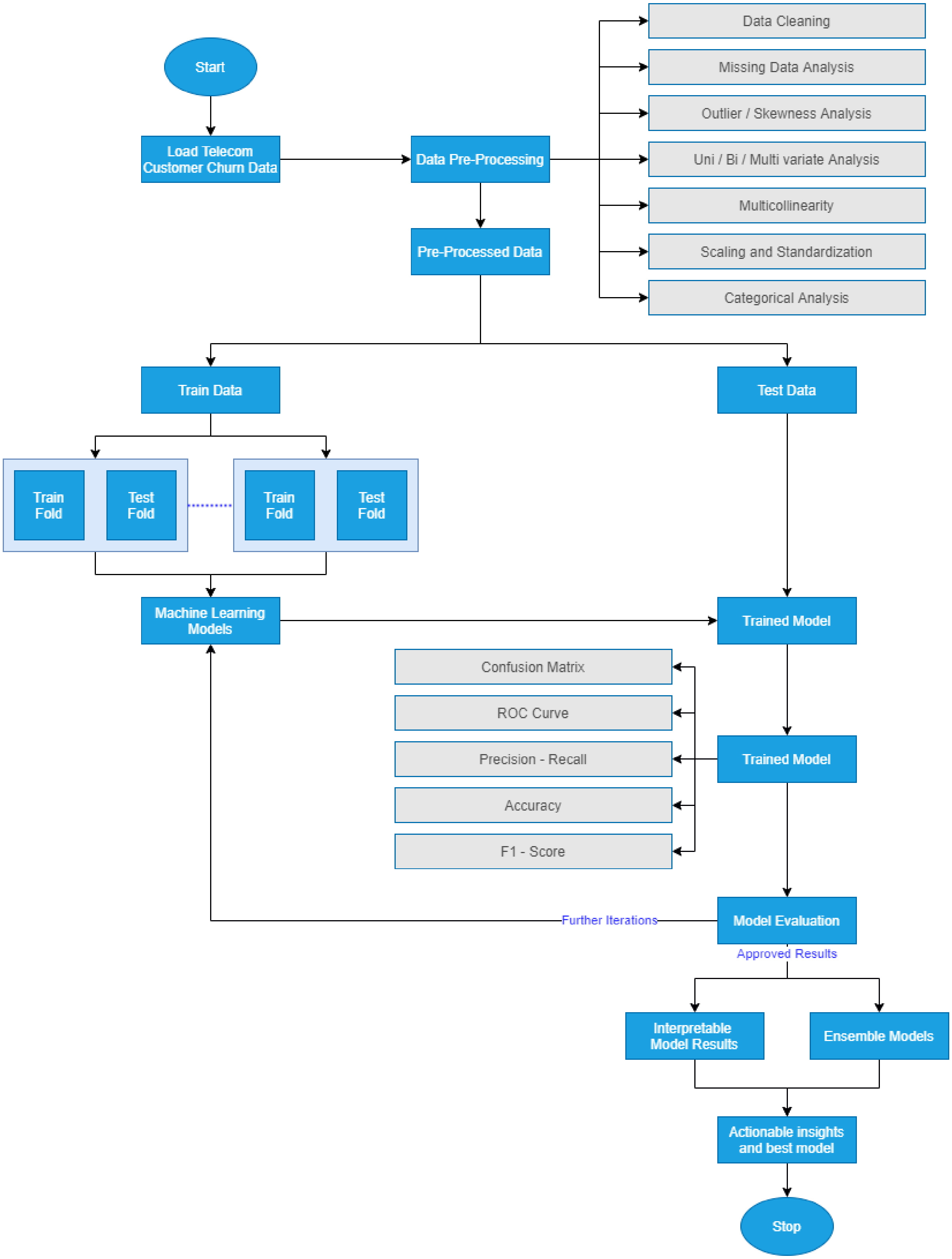


Figure 4: Model Building Process

3.2.6 Model Evaluation

The evaluation of the models will be done using metrics such as F-Measure, AUC, and accuracy, as mentioned in the literature review.

3.2.6.1 Evaluation Metrics

A comparison of the model results will be made based on the metrics obtained from the literature previously surveyed. They used the same accuracy metrics, F-Score, the area under the curve, and the new ensemble or individual models' performance to the models' performance in the field's reviewed literature. There are standard metrics that can be used and can be visually compared to select a model that can excel in most of the evaluation metrics chosen for classification.

CHAPTER 4: ANALYSIS

In this chapter, an in-depth analysis of the steps that can be taken to perform customer churn analysis will be explained with a business explanation and technical justification.

4.1 Dataset Description

The dataset used is sourced from IBM (Kaggle, 2018). The dataset will be analyzed to understand customer behaviour to predict the likelihood to churn customers.

The target variable is the attribute Churn. There are 21 attributes, and the Churn column is the variable that is being predicted. 7043 data points capture customer level data along with their metadata in the form of attributes. This data has been sourced from the Cognos Analytics Team at IBM. It contains information about a telecom company that provided telecom and internet services to 7043 customers. The data indicate that the customers that have stayed left or signed up for the service. It contains 18 categorical attributes and three numerical attributes, including the target variable.

4.2 Exploratory Data Analysis

In this section, the details of the IBM Telecom dataset will be understood. The focus will be on the data details and how the data can be used as input for the various models. Analysis of the data in the form of analysis, both univariate and bivariate, will be presented. The distribution of the variables will also be analysed along with missing value analysis and outlier analysis.

4.2.1 Analysis of Variables

In this section, the percentage distribution of variables and absolute distribution will be analyzed through the visualizations. The chart shown below helps us get an overall understanding of the distribution of each of the variables being considered. The distribution of gender suggests that the distribution of males and females is almost equal in the customer base of the telecom data.

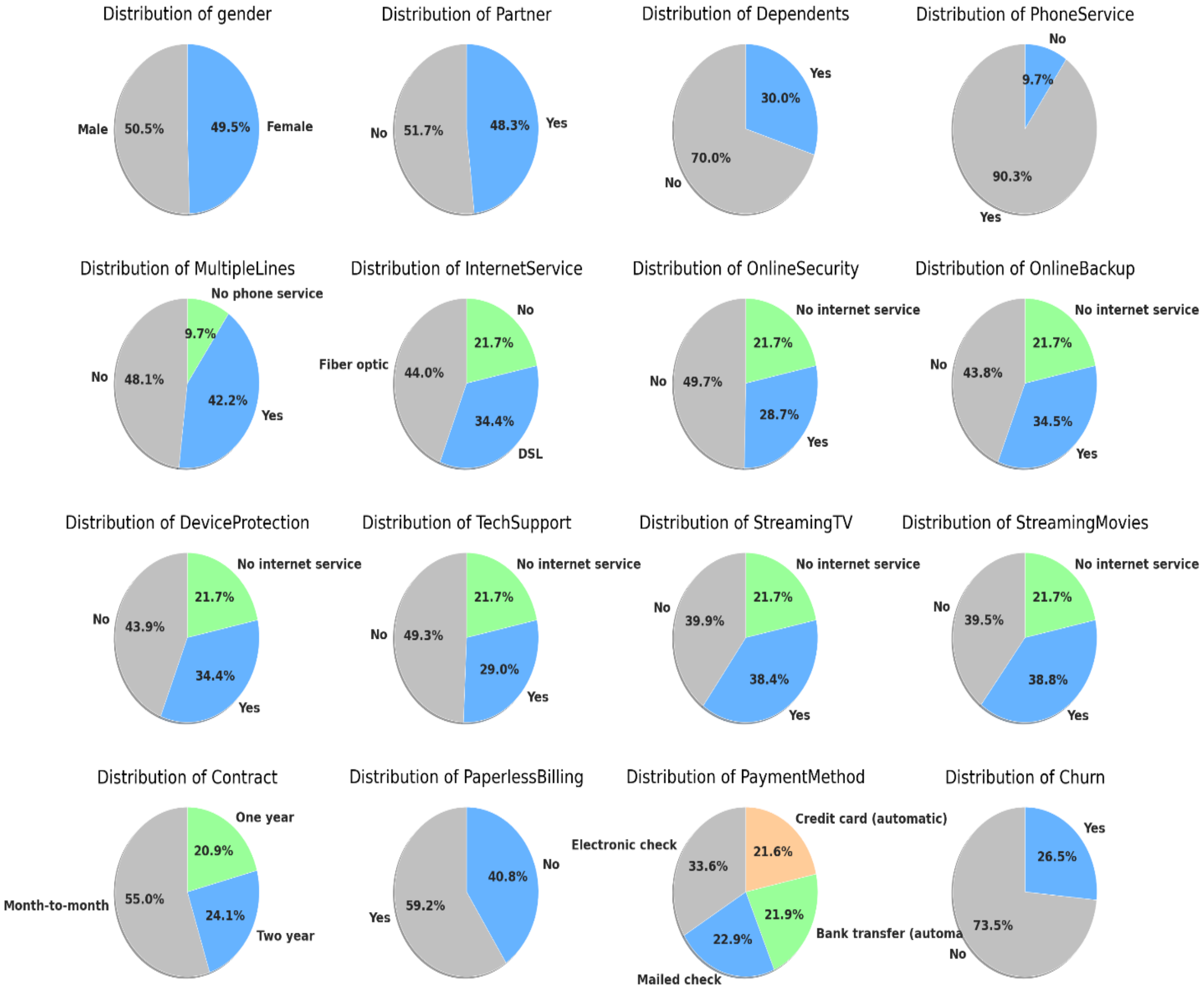


Figure 5: Distribution of variables (by percentage)

Based on the distribution of Phone Service, it is also understood that 90.3% of customers use the phone service and 9.7% of customers use other services such as the internet from the company.

4.2.2 Outlier Analysis

The dataset has categorical variables as metadata for each customer. There are two attributes – Monthly Charges and Total Charges-numerical values on which outlier analysis can be performed. The study will be using a boxplot with an inter-quartile range of 1.5 x (Interquartile Range) as the upper and lower whiskers for the two attributes. The attributes will be plotted against churn, where 0 indicates that the customer did not churn and one indicates that the customer did churn.

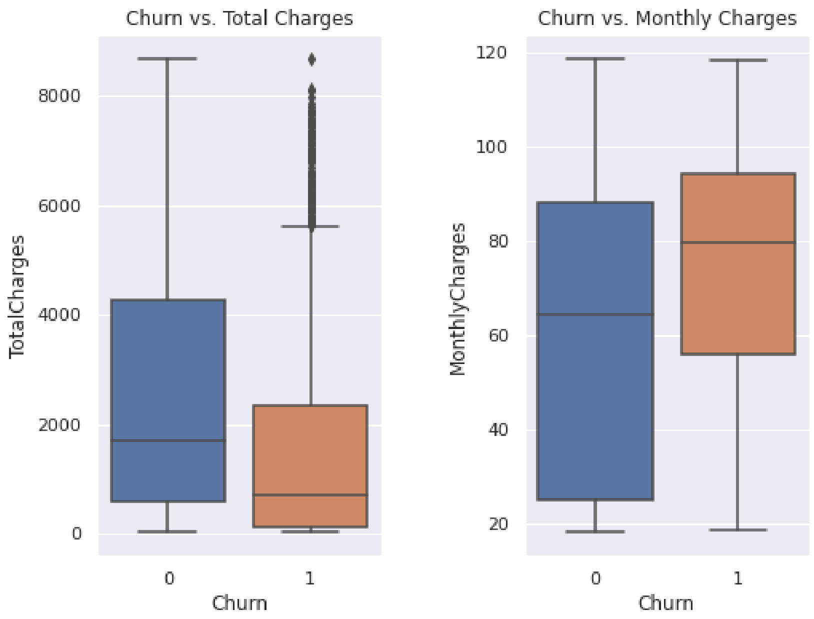


Figure 6: Boxplots of Churn versus Total Charges and Churn versus Monthly Charges

The distribution in Figure 4.1 shows that for Total Charges, most customers have a customer lifetime value of less than 2000, which indicates that customers who have a lower tenure with the company are likely to churn. Whereas, in the boxplot of Monthly Charges, the distribution of customers that churn is populated between 60 to 90.

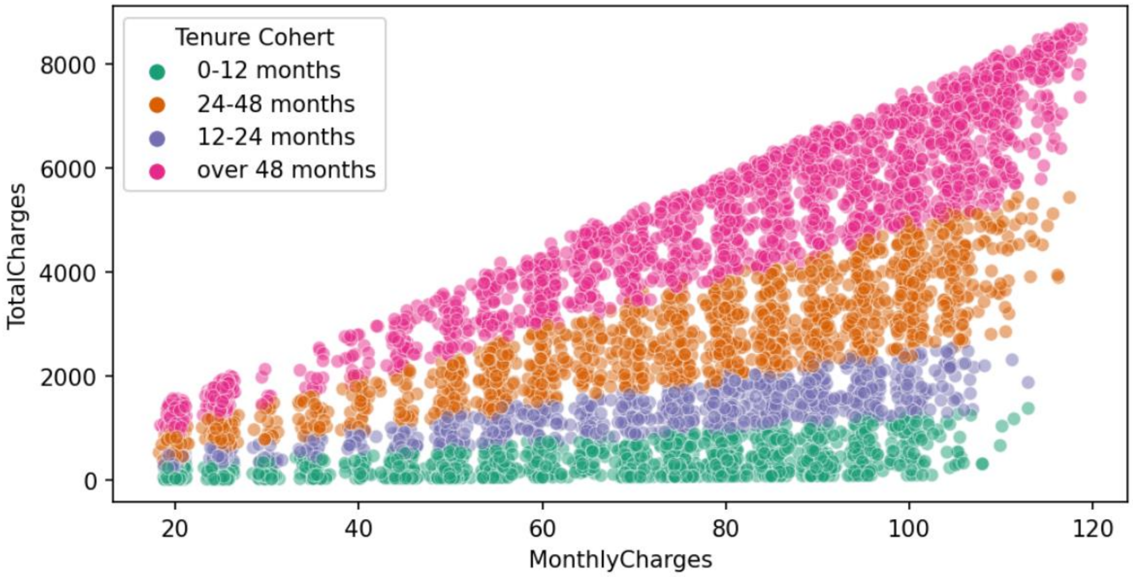


Figure 7: Scatter plot of Monthly Charges versus Total Charges

There is a significant correlation between Monthly Charges and Total Charges, as expected. As expected, Figure 4.2 illustrates that as the monthly charge per customer increases, the total charges or the customer lifetime value to the company increase.

4.3 Univariate Analysis

In this section, the numerical attributes of the dataset will be analyzed in greater depth. Understanding the distribution of the three numerical features – Monthly Charges, Total Charges, and Tenure is how univariate analysis can be performed. Most customers have a monthly charge of around 20. The histogram of Monthly Charges suggests that high-value customers peak around 80 and gradually taper off around 120. The frequency of customers based on tenure suggests that after spending around 15 months with the Telecom company, the number of customers with a high tenure decreases.

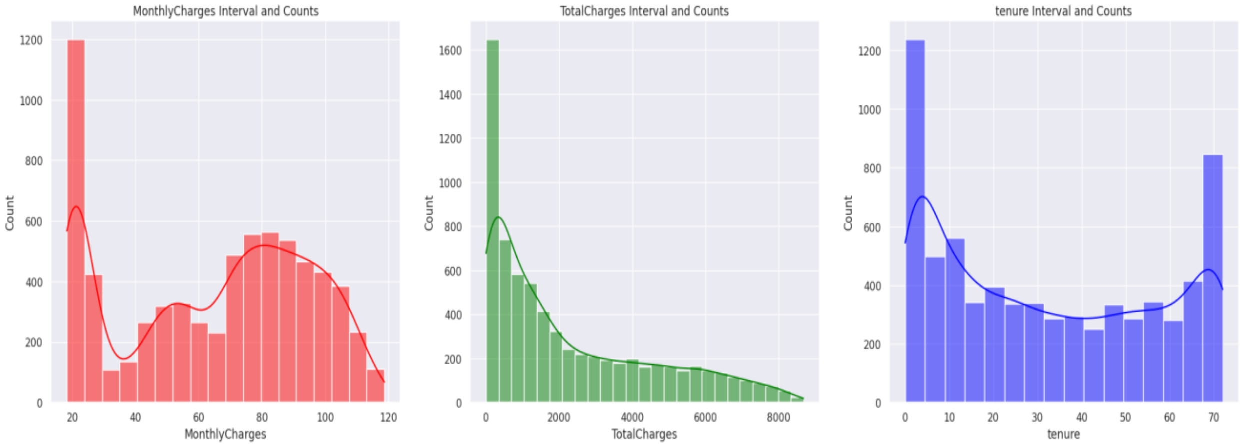


Figure 8 : Univariate Analysis of numerical features of IBM Teleco Data

The visualization in Figure 4.3 showcases the distribution of the numerical values, where Monthly charges seem to have an uneven distribution.

4.3.1 Relation with Target Variable

In this section, the relation of multiple attributes concerning churn is observed. Based on whether the churn is marked as Yes or No, the distribution of multiple features is observed. A deeper understanding of the behaviour or churn is observed when visualizations are used. The relation of the demographic variables with churn can be seen in Figure 4.6. It is observed that customers that do not have a partner or dependents are more likely to churn. This churn indicates that customers who have a family might take more services from the company and are more likely to stick to them.

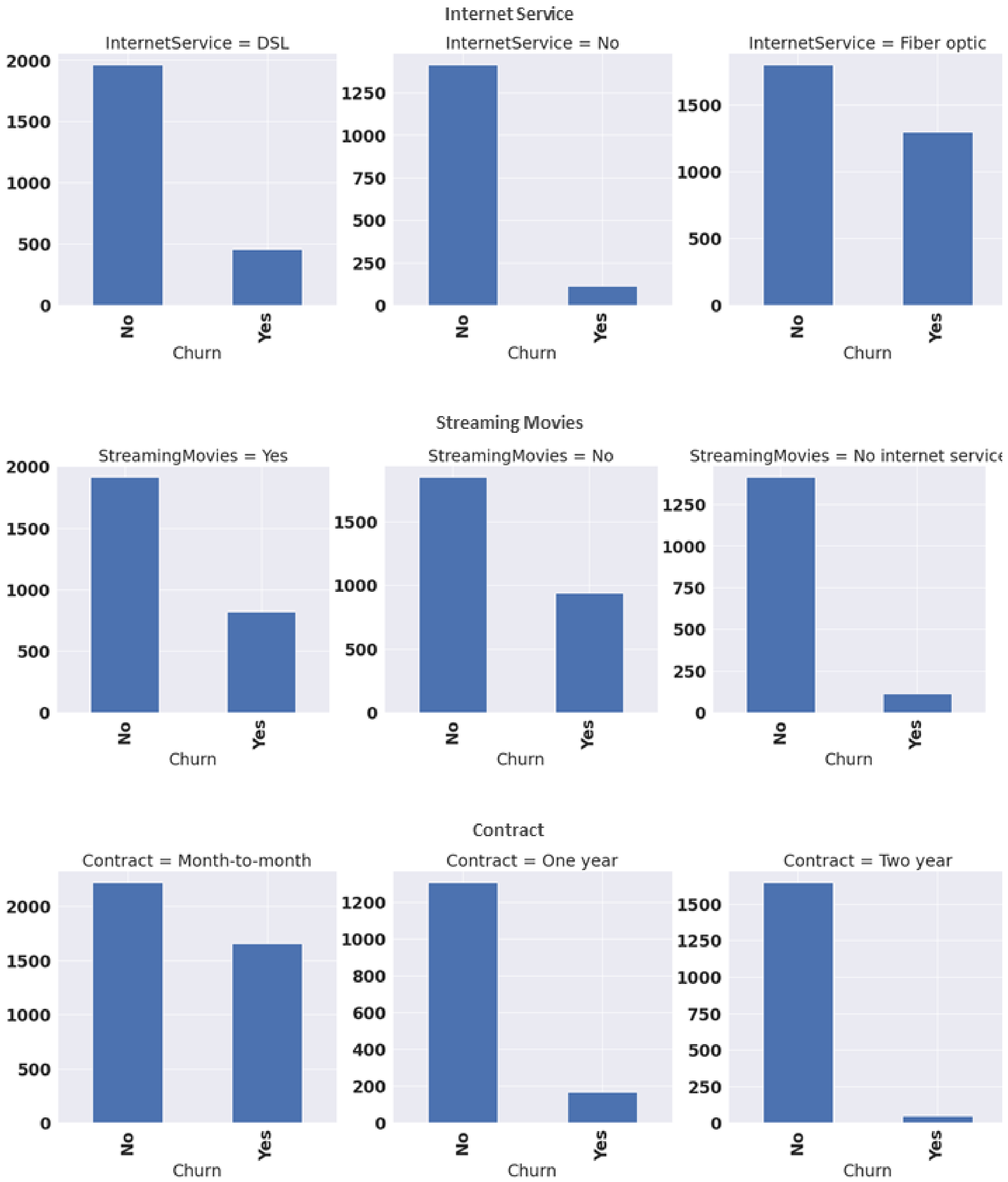


Figure 9 : Churn analyzed with Internet, Streaming and Contract

Correlation of the various attributes will also be noted, where the relationship between quantitative variables and the correlation between qualitative/ categorical variables will also be plotted. The correlation between the variables in the dataset can be understood using these plots.

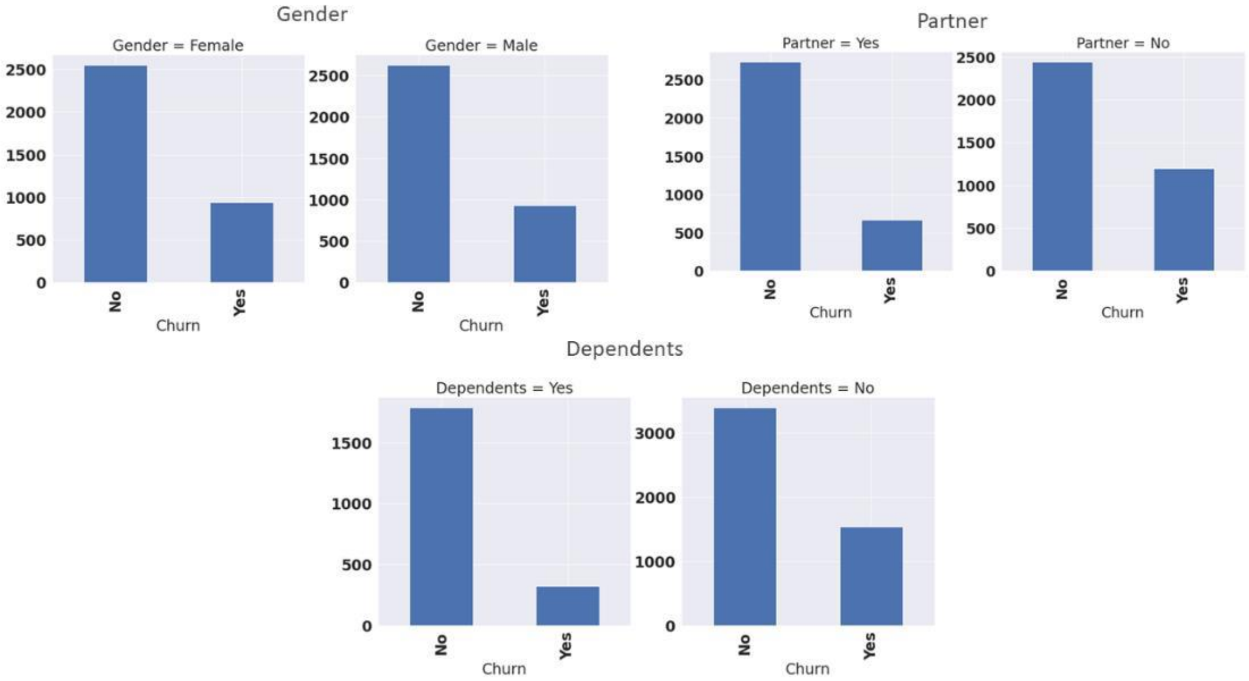


Figure 9.1: Distribution of Demographic Attributes with respect to Churn

4.3.2 Distribution of variables with respect to Churn

In Figure 4.8, the distribution of features with respect to churn can be observed with the help of a stacked histogram. Directing focus on the customers who have churned will help us identify the likely churn patterns. Some of the observations made from the visualizations can be confirmed in the feature selection techniques. The number of males and females churning is equal. Most people that churn do not have dependents. Customers with dependents are less likely to churn as they have settled into the ecosystem and do no change it unless there is something significant. Customers who use tech support are less likely to churn as they are willing to work with the telecom operator to fix the issues they may face. Electronic check is the most used payment method by the customers that churn – this signifies the method that has the maximum friction compared to an automatic deduction. Customers that do not use online security are also more likely to churn. A combination of these behaviour patterns can be observed from the charts that have been plotted in Figure 4.8, where the distribution of churned and not churned customer gives us insights into customer behaviour.

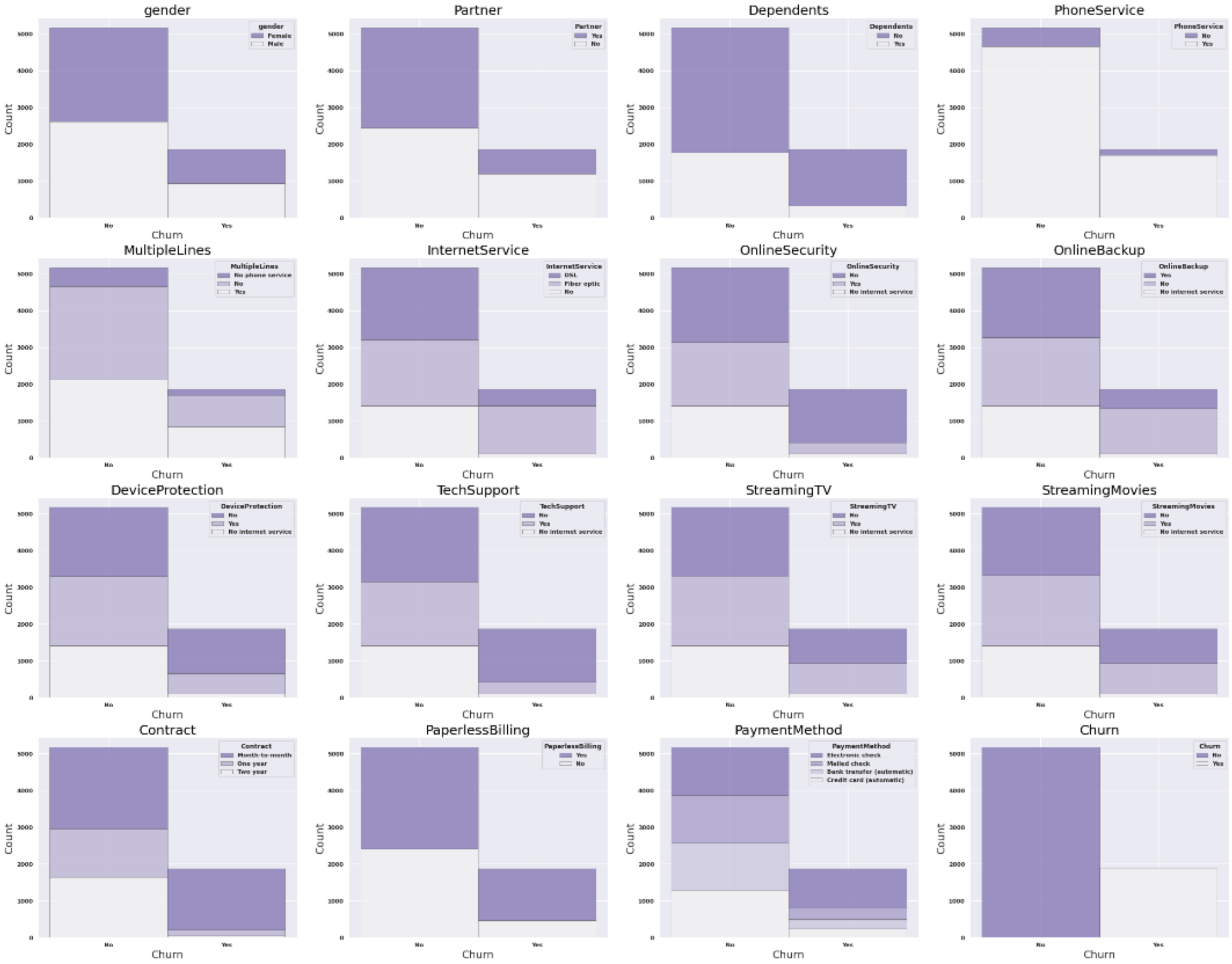


Figure 9.2: Distribution of all features with respect to Churn

Clients on a monthly contract are more likely to churn as it is easier to move out when there is no long-term commitment with the telecom operator. The observations from the chart will be leveraged in the model building phase to drive better insights for the business using interpretable machine learning.

4.4 Methods

In this section, the methods and standards that will be leveraged in this study. The conventions followed through the study will be highlighted in the form of the data split, the encoding used, and the feature engineering employed to predict the customers at a high risk of churn.

4.4.1 Data Split

The dataset will be split at a train-test ratio of 80% train data and 20% test data using the sklearn model selection library. The split will be done in a stratified manner by the train-test package leveraged in python. The main objective of the stratified train-test split is to keep the same proportion of train and test class samples as the original data.

4.4.2 Feature Engineering

Feature engineering is the process of creating new features by transforming existing features into a new feature space. Feature engineering does have the potential to improve model performance (Khurana et al., 2017). However, in our use-case where there are two numerical attributes, monthly charges and total charges, feature engineering will not make sense here as generating a new feature will bring about high multicollinearity in the data. Box-Cox transformation was also applied on the dataset for specific columns, such as monthly charges.

4.4.3 Implementation

All of the analysis and implementation has been done on Google Colab. The virtual machine's configuration is two CPU cores of the Haswell CPU family at 2.30GHz with RAM of 16 GB and disk space of 25 GB. All of the packages that have been leveraged are open source python packages. For instance, for importing the data and working with data frames, NumPy and pandas have been used. For the visualization, packages like matplotlib, seaborn, pandas-profiling and sweetviz have been used. Machine learning models have been implemented leveraging packages such as sklearn, Xgboost and catboost. The code was developed on the Colab platform using the native inbuild CPU and compute power on the Edge Browser. The data was sourced from Kaggle and pulled in.

4.5 Analysis

In this section, the baselines and implementation of research models will be decided. It will include pre-processing, feature selection, class balancing, ensemble models, cross-validation and model interpretability.

4.5.1 Models

Multiple models will be leveraged to predict customer churn from the data. Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, K-Nearest Neighbour, Gradient Boosting Classifier, Stochastic Gradient Descent, Light Gradient Boosted Machine were leveraged to predict churn and individual model performance was analyzed.

4.5.2 Feature selection

Feature selection techniques help us understand the features that are of higher importance for the prediction of churn. The feature selection techniques leveraged are Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier and Light GBM. The features that were showcased as necessary is shown in Figure 4.9 and Figure 4.10.

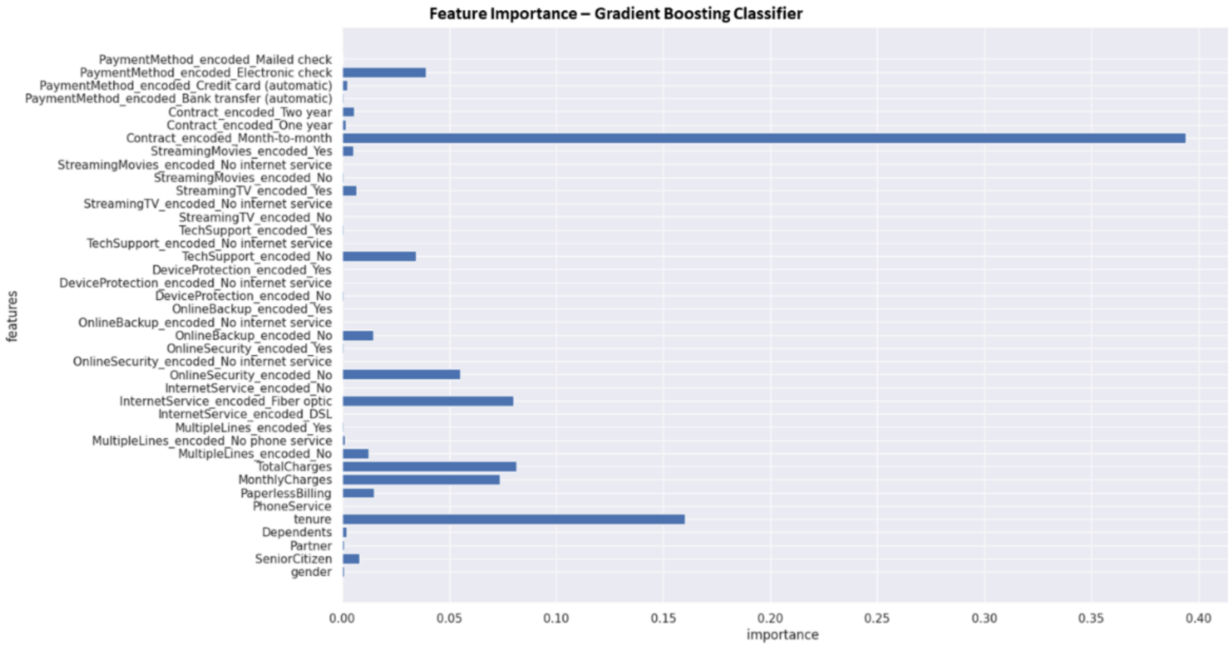


Figure 9.3: Feature Selection using Gradient Boosting Classifier

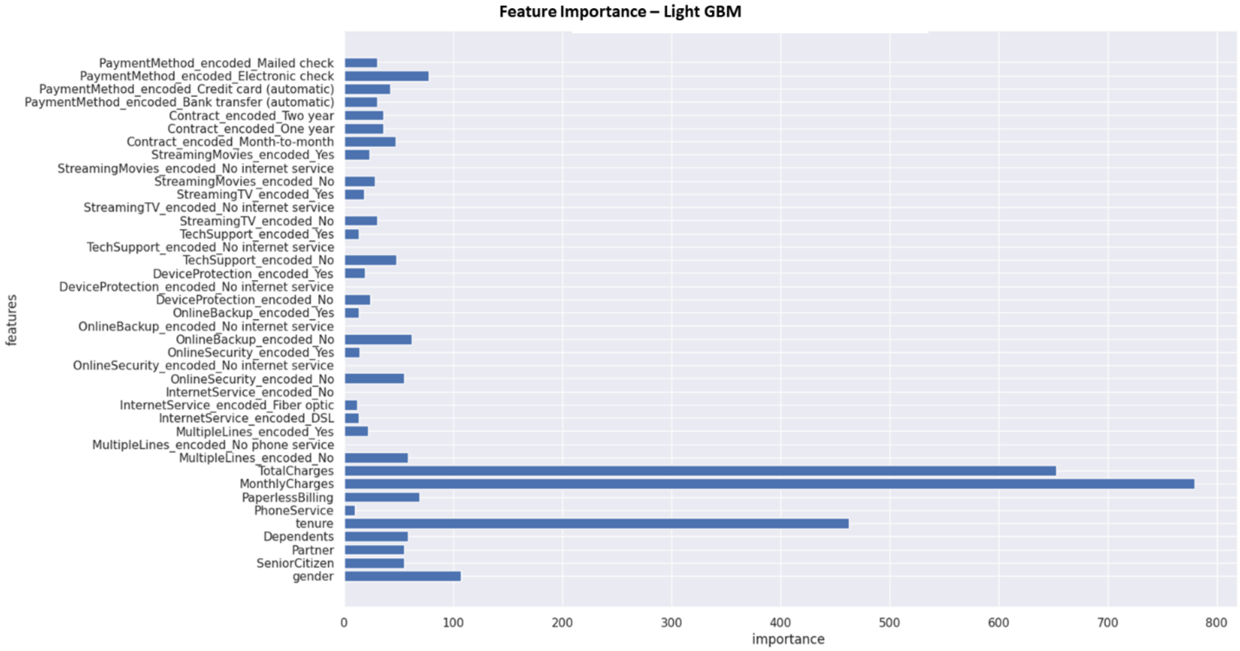


Figure 9.4: Feature Selection using Gradient Boosting Classifier and Light GBM

From the charts and graphs shown above in Figure 4.15 and 4.16, it is noted that the crucial features are the month to month contracts, the tenure of the customer, the total charges and the monthly charges. The attributes have the highest variance compared to the categorical values in the other attributes that signify customer behaviour.

CHAPTER 5: RESULTS AND DISCUSSIONS

In this chapter, the results and the discussion, the interpretation of the results are made in detail. In this chapter, the work done to keep up with industry best practices will be showcased and possible results when the focus is put on following best practices.

5.1 Interpretation of Visualisations

In Section, where exploratory data analysis was done on the telecom dataset, multiple visualizations were plotted to get a more profound intuition of the dataset and the relation of the attributes with the target variable. The most vital indicators of customer churn are total charges and monthly charges. The customers that have the highest monthly charges generally have internet service, as noticed from the heatmap, as they have a high correlation of -0.8 for no internet service as seen in The observations made from the chart will be leveraged in the model building phase to drive better insights for the business using interpretable machine learning.

When a customer does have internet service, the monthly charges are higher, and hence, the chance of churn is lower. The customer is less likely to churn because they are deep in the ecosystem of the telecom operator, and hence, there is high friction to move to another telecom operator.

Another notable attribute that is an indicator of churn is when the mode of the contract is month to month. Customers who do not have a one-year or two-year contract have a higher tendency to churn. The distribution of monthly charges based on churn shows that the customers who do not have dependents are more likely to churn. Customers who have partners or dependents tend to be more stable in their choices and do not have time to look out for offerings from other telecom operators. The interpretations are primarily from feature importance graphics. A high positive correlation between monthly charges and total charges is observed with an equal distribution of male and female customers in Figure 4.2 Scatter plot of Monthly Charges versus Total Charges. This correlation indicates that as monthly charges increase, the customer is more invested in the subscription they have opted for, thus indicating a significantly higher customer lifetime value. Customers who opt-in for more services such as streaming movies and online security are less likely to churn. One of the primary interpretations is that the more services the customers use from the telecom operator, the more likely they will be loyal.

5.2 Model Results

In the following section, the various model results will be analyzed. Along with the individual model results, the ensemble model results have also been analyzed. After exploratory data analysis was performed on the dataset, missing value analysis and outlier analysis was performed. One-hot encoding was performed on the categorical attributes present in the dataset. Box cox transformation was done on the skewed variables, and the best results were taken into consideration. There is a split of 80% train, 10% validation and 10% test for the models as stated in Table 2.

Model	Train (%)	Validation (%)	Test (%)	AUC
Gaussian Naïve Bayes	70.38	70.03	67.52	0.73
Bernoulli Naïve Bayes	72.93	72.87	69.36	0.74
Logistic Regression	80.8	81.39	78.3	0.71
Random Forest	99.75	79.83	75.04	0.67
Support Vector Machine	81.82	81.25	77.45	0.68
Decision Tree	99.75	75.85	70.21	0.65
K Nearest Neighbour	83.51	74.86	70.21	0.65
Gradient Boosting	83.3	79.83	77.16	0.69
Stochastic Gradient Descent	76.98	79.83	73.62	0.72
Light Gradient Boosting Machine	88.34	79.12	76.31	0.68

Table 2: Model Results of Individual and Ensemble Models

It is observed that the model with the highest accuracy is logistic regression, with an accuracy of 78.3% with an AUC score of 0.71. The model with the highest AUC score of 0.74 is Bernoulli, which has an accuracy of 69.36% on the test data. The results of every model used, both the individual model and the ensemble models such as Gradient Boosting, Stochastic, and Light GBM, have been used for this study. The visual representation of the scores can be seen in Figure 5.1, where it is noticed that the decision tree and random forest classifier may be overfitting the data. Overall, all of the individual and ensemble models have an accuracy score above 67%, and this is an indication that models are getting trained satisfactorily for a preliminary run on the IBM Telecom Churn dataset.

CHAPTER 6:
CONCLUSIONS AND RECOMMENDATIONS

In this study, the classification of telecom customers that will churn has been done with the help of machine learning models. Multiple papers related to customer churn in the telecom industry was analyzed to perform a preliminary analysis to ensure that best practices were implemented. Some studies focused on data processing, and there were research papers that focused on finding the best model that would give us the best results. While there were papers that focused on bringing about the best results, the focus of this study has been to bring about the best possible results along with the focus on model interpretability.

6.1 Discussion and Conclusion

In this study, the performance of individual and ensemble models was carried out to classify churned customers. A baseline was set using the logistic regression and decision tree classifier, where the test accuracy was noted. The preliminary analysis of the data was done by looking at the fundamental statistics of the data. Then, the distribution of the variables was analyzed, followed by missing value analysis and outlier analysis. Univariate and bivariate analysis with respect to the target variable churn was done. The distribution of the variables with respect to churn was analyzed to deep dive into the latent relationships within the dataset. The was followed by analysing the Pearson’s correlation coefficient by plotting heatmaps for categorical variables and the numerical attributes.

In any analysis, before more advanced techniques are implemented on the data, it is essential to understand the data in depth. The probability distribution of the numerical variables is analyzed using a non-parametric kernel density estimation. For variables that are skewed, the box-cox transformation was applied to normalize the distribution.

The various feature selections methods that can be employed were also discussed to understand the crucial features that will help understand if a customer is likely to churn or not. This approach will help in model results is as optimized as possible and interpretation of real-world applications. This study aimed to build a model that can be deployed in real-world scenarios.

It was observed that ensemble models tend to have better performance as compared to individual models. Decision Tree with AdaBoost, Decision Tree with Bagging, CatBoost, Linear Support Vector Classification, Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbour, Naïve Bayes, Decision Tree and SVM with radial basis function kernel were implemented. The highest AUC scores were by the decision tree with AdaBoost and decision tree with bagging with scores of 0.84 for both.

6.2 Future Recommendations

There are various areas of research that one can take going ahead. The model has now been performed on a static dataset. Implementing a similar pipeline at an enterprise level at a fixed cadence can help track customer behaviour and reinforce the model. Natural Language Processing can be leveraged on feedback gathered from focus groups as to why customers are churning from ratings, reviews, social media and calls from customer service agents. Some patterns can be analyzed, such as geography, demographic information, and other factors analyzed further. This model can be improved to calculate the percentage of revenue saved by the company based on the evaluation metrics. Based on this information, the lift in sales can be analyzed.

REFERENCES

1. Agrawal, S., (2018) Customer Churn Prediction Modelling Based on Behavioural patterns Analysis using Deep Learning. *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, pp.1–6.
2. 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.
3. Ahmed, A. and Linen, D.M., (2017) A review and analysis of churn prediction methods for customer retention in telecom industries. In: *2017 4th International Conference on Advanced Computing and Communication Systems, ICACCS 2017*. Institute of Electrical and Electronics Engineers Inc.
4. Ahmed, A.A. and Maheswari, D., (2017) A Review And Analysis Of Churn Prediction Methods For Customer Retention In Telecom Industries. *2017 International Conference on Advanced Computing and Communication Systems*.
5. Ambildhuke, G.M., Rekha, G. and Tyagi, A.K., (2021) Performance Analysis of Undersampling Approaches for Solving Customer Churn Prediction. [online] Springer, Singapore, pp.341–347. Available at: https://link.springer.com/chapter/10.1007/978-981-15-9689-6_37 [Accessed 25 May 2022].
6. Andrews, R., (2019) Churn Prediction in Telecom Sector Using Machine Learning. *International Journal of Information Systems and Computer Sciences*, 82, pp.132–134.
7. Anon (2021) *Cognos Analytics - IBM Business Analytics Community*. [online] Available at: https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113 [Accessed 24 May 2022].

8. Anon (2021) *Digital transformation for 2020 and beyond eight telco considerations*. [online] Available at: https://www.ey.com/en_in/tmt/digital-transformation-for-2020-and-beyond-eight-telco-considera [Accessed 25 May 2022].
9. Anon (2021) *Why is the telecom industry struggling with product success?* [online] Available at:https://internationalfinance.com/why-telecom-industry-struggling-product-success/ [Accessed 25 May 2022].

10. Castanedo, F., Valverde, G., Zaratiegui, J. and Vazquez, A., (2014) Using Deep Learning
11. to Predict Customer Churn in a Mobile Telecommunication Network Federico. pp.1–8.
12. 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 24 May. 2022].
13. Fonseca Coelho, A., (n.d.) *Churn Prediction in Telecom Sector: A completed data engineering Framework*.
14. Hadden, J., Tiwari, A., Roy, R. and Ruta, D., (2006) Churn Prediction: Does Technology Matter. *International Journal of Intelligent Technology*, 1, pp.104–110.
15. Halibas, A.S., Cherian Matthew, A., Pillai, I.G., Harold Reazol, J., Delvo, E.G. and Bonachita Reazol, L., (2019) Determining the intervening effects of exploratory data analysis and feature engineering in telecoms customer churn modelling. *2019 4th MEC International Conference on Big Data and Smart City, ICBDSC 2019*.
16. Hargreaves, C.A., (2019) A Machine Learning Algorithm for Churn Reduction & Revenue Maximization: An Application in the Telecommunication Industry. *International Journal of Future Computer and Communication*, 84, pp.109–113.
17. Havrylovych, M. and Nataliia Kuznietsova, ©, (2019) *Survival analysis methods for churn prevention in telecommunications industry*.
18. Jain, H., Yadav, G. and Manoov, R., (2021) Churn Prediction and Retention in Banking, Telecom and IT Sectors Using Machine Learning Techniques. [online] Springer, Singapore, pp.137–156. Available at: https://link.springer.com/chapter/10.1007/978-981-15-5243-4_12 [Accessed 21 Apr. 2022].

19. Kaggle, (2018) *Telco Customer Churn*. *Kaggle.com*. Available at: https://www.kaggle.com/blastchar/telco-customer-churn [Accessed 9 Apr .2022].

20. Karimi, N., Dash, A., Rautaray, S.S. and Pandey, M., (2021) A Proposed Model for Customer Churn Prediction and Factor Identification Behind Customer Churn in Telecom Industry. [online] Springer, Singapore, pp.359–369. Available at: https://link.springer.com/chapter/10.1007/978-981-15-7511-2_34 [Accessed 10 May. 2022].
21. Khurana, U., Nargesian, F., Samulowitz, H., Khalil, E.B. and Turaga, D., (2017) Learning Feature Engineering for Classification. [online] Available at: https://www.researchgate.net/publication/318829821 [Accessed 19 May 2022].
22. Kriti, (2019) *Customer churn: A study of factors affecting customer churn using machine learning*. [online] Available at: https://lib.dr.iastate.edu/creativecomponents [Accessed 14 May 2022].
23. Labhsetwar, S.R., (n.d.) Predictive Analysis Of Customer Churn in Telecom Industry using Supervised Learning.
24. Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P., (2021) Customer churn prediction system: a machine learning approach. *Computing*.
25. Mahdi, A., Alzubaidi, N. and Al-Shamery, E.S., (2020) Projection pursuit Random Forest using discriminant feature analysis model for churners prediction in telecom industry discriminant random forest Linear discriminant analysis oblique tree Project pursuit index Support vector machines. *International Journal of Electrical and Computer Engineering (IJECE)*, 102, pp.1406–1421.
26. Momin, S., Bohra, T. and Raut, P., (2020) *Prediction of Customer Churn Using Machine Learning*. *EAI/Springer Innovations in Communication and Computing*.

M_Sc_data_analytics_Project

June 13, 2022

1 Prediction of Customer Attrition in the Telecom Industry (using Machine Learning)

This notebook is based on the IBM Watson Telecom Dataset. We will go through all of the steps:

- Data Selection
- Data Preprocessing
- Data Transformation
- Data Visualization
- Class Balancing
- Model Building
- Model Evaluation
- Model Review

Aim: The purpose of this notebook is to analyze, visualize and model (classification) [IBM Telecom Dataset]

1 Loading the Data

2 Quick EDA

1.1 1 Loading the Data

[84]:

```
# Installing the required packages

!pip uninstall -y pandas-profiling &> /dev/null                # Package for␣
 ↪pandas profiling - visualization
!pip install pandas-profiling[notebook,html] &> /dev/null      # Uninstalling␣
 ↪and reinstalling it due to a bug in Google Colab
!pip install sweetviz &> /dev/null                             # Package for␣
 ↪some sweet visualizations
!pip install -U dataprep &> /dev/null                          # Package for␣
 ↪instant data preparation
!pip install --upgrade plotly &> /dev/null
!pip install pyyaml==5.4.1
!pip3 install --upgrade pip
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: pyyaml==5.4.1 in /usr/local/lib/python3.7/dist-
packages (5.4.1)
WARNING: Running pip as the 'root' user can result in broken permissions
and conflicting behaviour with the system package manager. It is recommended to
use a virtual environment instead: https://pip.pypa.io/warnings/venv
Looking in indexes: https://pypi.org/simple, https://us-
python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pip in /usr/local/lib/python3.7/dist-packages
(22.1.2)
WARNING: Running pip as the 'root' user can result in broken permissions
and conflicting behaviour with the system package manager. It is recommended to
use a virtual environment instead: https://pip.pypa.io/warnings/venv
```

[94]:

```python
from sklearn import svm, tree, linear_model, neighbors, naive_bayes, ensemble,␣
 ↪discriminant_analysis, gaussian_process
from sklearn.ensemble import AdaBoostClassifier, RandomForestClassifier,␣
 ↪GradientBoostingClassifier
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor

from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
import warnings
import re
import os

# Removing the minimum display columns to 500
pd.set_option('display.max_columns', 1000)
pd.set_option('display.max_rows', 1000)

# Hide Warnings
warnings.filterwarnings('ignore') #adjust display options

%matplotlib inline
```

[95]:

```python
# Let us import the required packages
```

```
from pandas_profiling import ProfileReport          # Pandas Profile to␣
 ↪visualize the data
from dataprep.eda import plot_correlation           # Importing a package to␣
 ↪visualize data correlation
from dataprep.eda import plot_missing               # Importing packge to plot␣
 ↪the missing values of the dataset
from IPython.display import display                 # Displaying widgets
import matplotlib.ticker as ticker                  # User-defined function for␣
 ↪formatting graphs
import plotly.graph_objects as go                   # To use graph objects␣
 ↪within visualizations

from plotly.offline import iplot                    # Used for interactive plots
import matplotlib.pyplot as plt                     # Importing matplotlib for␣
 ↪visualization
from dataprep.eda import plot                       # Importing package to plot/
 ↪ visualize features of the dataset
import ipywidgets as widgets                        # Creating widgets
import plotly.express as px                         # Importing plotly express␣
 ↪for visualizations
import matplotlib.cm as cm                          # Colormaps, colormaps␣
 ↪handling utilities

import pandas_profiling                             # Automatic EDA
import sweetviz as sv                               # Importing sweetviz for␣
 ↪some sweet visualizations
import seaborn as sns                               # Importing seaborn for␣
 ↪visualization
import pandas as pd                                 # Importing pandas
import numpy as np                                  # Importing numpy
%matplotlib inline
import warnings                                     # Importing package to␣
 ↪toggle warnings
import IPython                                      # Importing ipython for␣
 ↪displaying html files in the notebook

# Removing the minimum display columns to 500
pd.set_option('display.max_columns', 500)
pd.set_option('display.max_rows', 500)

# Hide Warnings
warnings.filterwarnings('ignore')
```

[96]:

```
# Reading the file onto Google Colab from GitHub using pandas library
telecomDf = pd.read_csv("/content/drive/MyDrive/
 ↪WA_Fn-UseC_-Telco-Customer-Churn.csv")     # Reading file CSV
```

[97]:

```
df = telecomDf.copy(deep = True)
# Viewing the head of the data
df.head()
```

[97]:

```
   customerID  gender  SeniorCitizen Partner Dependents  tenure PhoneService  \
0  7590-VHVEG  Female              0     Yes         No       1           No   
1  5575-GNVDE    Male              0      No         No      34          Yes   
2  3668-QPYBK    Male              0      No         No       2          Yes   
3  7795-CFOCW    Male              0      No         No      45           No   
4  9237-HQITU  Female              0      No         No       2          Yes   

      MultipleLines InternetService OnlineSecurity OnlineBackup  \
0  No phone service             DSL             No          Yes   
1                No             DSL            Yes           No   
2                No             DSL            Yes          Yes   
3  No phone service             DSL            Yes           No   
4                No     Fiber optic             No           No   

  DeviceProtection TechSupport StreamingTV StreamingMovies        Contract  \
0               No          No          No              No  Month-to-month   
1              Yes          No          No              No        One year   
2               No          No          No              No  Month-to-month   
3              Yes         Yes          No              No        One year   
4               No          No          No              No  Month-to-month   

  PaperlessBilling              PaymentMethod MonthlyCharges TotalCharges  \
0              Yes           Electronic check          29.85        29.85   
1               No               Mailed check          56.95       1889.5   
2              Yes               Mailed check          53.85       108.15   
3               No  Bank transfer (automatic)          42.30      1840.75   
4              Yes           Electronic check          70.70       151.65   

  Churn  
0    No  
1    No  
2   Yes  
3    No  
4   Yes  
```

[98]:

```
df.shape
```

[98]:

```
(7043, 21)
```

[99]:

```
# Let's view a summary of the dataset now
print(df.info(verbose=True))
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
```

```
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
None
```

Variable Descriptions - gender –> Whether the customer is a male or a female - SeniorCitizen –> Whether the customer is a senior citizen or not (1, 0) - Partner –> Whether the customer has a partner or not (Yes, No) - Dependents –> Whether the customer has dependents or not (Yes, No) - tenure –> Number of months the customer has stayed with the company - PhoneService –> Whether the customer has a phone service or not (Yes, No) - MultipleLines –> Whether the customer has multiple lines or not (Yes, No, No phone service) - InternetService –> Customer's internet service provider (DSL, Fiber optic, No) - OnlineSecurity –> Whether the customer has online security or not (Yes, No, No internet service) - OnlineBackup –> Whether the customer has online backup or not (Yes, No, No internet service) - DeviceProtection –> Whether the customer has device protection or not (Yes, No, No internet service) - TechSupport –> Whether the customer has tech support or not (Yes, No, No internet service) - StreamingTV –> Whether the customer has streaming TV or not (Yes, No, No internet service) - StreamingMovies –> Whether the customer has streaming movies or not (Yes, No, No internet service) - Contract –> The contract term of the customer (Month-to-month, One year, Two year) - PaperlessBilling –> Whether the customer has paperless billing or not (Yes, No) - PaymentMethod –> The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)) - MonthlyCharges –> The amount charged to the customer monthly - TotalCharges –> The total amount charged to the customer - Churn –> Whether the customer churned or not (Yes or No)

1.2 2 Quick EDA

Now that we have some basic data understanding of the data we are dealing with, it's time to try and understand things a little more in details. We will proceed to analyze and explore the data. There is a common term in the Data Science community that we use to describe this: Exploratory Data Analysis (EDA)

Exploratory Data Analysis is used to get a feel of the data. We use it to understand the attributes, gaps and behaviour that the data has. A traditional way of doing it can be looking at the data column by column, row by row. As more developers contribute to the open-source python ecosystem, we have great packages that can help us analyze the data with minimal effort.

1.3 ### 2.1 Pandas Profiling

[100]:

```python
# Making a copy of the dataset as df
df = telecomDf.copy(deep = True)
```

[101]:

```python
# Generating the profile report and feeding it into a variable
Profile = ProfileReport(df, title = 'Telecom Data: Pandas Profiling Report', html = {'style':{'full_width':True}})
```

[102]:

```python
# Showcasing the Pandas Profiling Report for the Zomato Dataset
Profile
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

<IPython.core.display.HTML object>
```

[102]:

1.3.1 2.3 Dataprep

Dataprep is a python package to collect data, perform eda, clean and standardize data.

[103]:

```python
# Plotting the correlation of the dataframe
plot_correlation(df)
```

```
[103]: <dataprep.eda.container.Container at 0x7f4397e69e90>
```

[104]:

```python
# Analyzing the missing values from the dataset and visualizing it
plot_missing(df)
```

```
[104]: <dataprep.eda.container.Container at 0x7f439847b050>
```

- Tenure & Monthly Charges are correlated with churn

[ ]:

```

```

1.4 3 Pre-processing

[105]:

```
df.info()
```

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

[106]:

```
df.head().T
```

[106]:

```
                                 0           1           2  \
customerID              7590-VHVEG  5575-GNVDE  3668-QPYBK
gender                      Female        Male        Male
SeniorCitizen                    0           0           0
Partner                        Yes          No          No
Dependents                      No          No          No
tenure                           1          34           2
PhoneService                    No         Yes         Yes
MultipleLines     No phone service          No          No
```

```
InternetService                DSL           DSL             DSL
OnlineSecurity                  No           Yes             Yes
OnlineBackup                   Yes            No             Yes
DeviceProtection                No           Yes              No
TechSupport                     No            No              No
StreamingTV                     No            No              No
StreamingMovies                 No            No              No
Contract            Month-to-month      One year  Month-to-month
PaperlessBilling               Yes            No             Yes
PaymentMethod     Electronic check  Mailed check    Mailed check
MonthlyCharges               29.85         56.95           53.85
TotalCharges                 29.85        1889.5          108.15
Churn                           No            No             Yes

                                          3                 4
customerID                       7795-CFOCW        9237-HQITU
gender                                 Male            Female
SeniorCitizen                             0                 0
Partner                                  No                No
Dependents                               No                No
tenure                                   45                 2
PhoneService                             No               Yes
MultipleLines              No phone service                No
InternetService                         DSL       Fiber optic
OnlineSecurity                          Yes                No
OnlineBackup                             No                No
DeviceProtection                        Yes                No
TechSupport                             Yes                No
StreamingTV                              No                No
StreamingMovies                          No                No
Contract                           One year    Month-to-month
PaperlessBilling                         No               Yes
PaymentMethod     Bank transfer (automatic)  Electronic check
MonthlyCharges                         42.3              70.7
TotalCharges                        1840.75            151.65
Churn                                    No               Yes
```

Now that we have seen what the data looks like, let's make a justified action plan on how we care going to process the data before we go ahead with modelling:

- Ensure correct data type (TotalCharges)
- One-Hot encoding
- Normalization & Standardization
- Removal of Customer ID column (high cardinality)
- Correlation Analysis
- Multi Collinearity Analysis

One-hot encoding

[107]:

```
df = telecomDf.copy(deep = True)
```

[108]:

```
df.pop('customerID')
```

[108]:

```
0       7590-VHVEG
1       5575-GNVDE
2       3668-QPYBK
3       7795-CFOCW
4       9237-HQITU
           …
7038    6840-RESVB
7039    2234-XADUH
7040    4801-JZAZL
7041    8361-LTMKD
7042    3186-AJIEK
Name: customerID, Length: 7043, dtype: object
```

[109]:

```
# Adjusting columns that are categorical to numeric
listOfCategoricalColumnsToConvert = ["TotalCharges", "MonthlyCharges"]
df[listOfCategoricalColumnsToConvert] = df[listOfCategoricalColumnsToConvert].
 ↪apply(pd.to_numeric, errors='coerce')
```

[110]:

```
cols = df.columns
num_cols = df._get_numeric_data().columns
num_cols
```

[110]:

```
Index(['SeniorCitizen', 'tenure', 'MonthlyCharges', 'TotalCharges'],
dtype='object')
```

[111]:

```
listOfCategoricalColumns = list(set(cols) - set(num_cols))
listOfCategoricalColumns
```

[111]:

```
['StreamingTV',
 'TechSupport',
 'Dependents',
 'PhoneService',
 'Churn',
 'OnlineBackup',
 'Partner',
 'PaymentMethod',
 'OnlineSecurity',
 'Contract',
 'gender',
 'StreamingMovies',
 'PaperlessBilling',
 'InternetService',
 'MultipleLines',
 'DeviceProtection']
```

[112]:

```
pd.get_dummies(data=df, columns= listOfCategoricalColumns)
```

[112]:

```
      SeniorCitizen  tenure  MonthlyCharges  TotalCharges  StreamingTV_No  \
0                 0       1           29.85         29.85               1
1                 0      34           56.95       1889.50               1
2                 0       2           53.85        108.15               1
3                 0      45           42.30       1840.75               1
4                 0       2           70.70        151.65               1
…             …     …             …           …             …
7038              0      24           84.80       1990.50               0
7039              0      72          103.20       7362.90               0
7040              0      11           29.60        346.45               1
7041              1       4           74.40        306.60               1
7042              0      66          105.65       6844.50               0

      StreamingTV_No internet service  StreamingTV_Yes  TechSupport_No  \
0                                   0                0               1
1                                   0                0               1
2                                   0                0               1
3                                   0                0               0
4                                   0                0               1
…                               …              …             …
7038                                0                1               0
7039                                0                1               1
7040                                0                0               1
7041                                0                0               1
7042                                0                1               0

      TechSupport_No internet service  TechSupport_Yes  Dependents_No  \
0                                   0                0              1
1                                   0                0              1
2                                   0                0              1
3                                   0                1              1
4                                   0                0              1
…                               …              …            …
7038                                0                1              0
7039                                0                0              0
7040                                0                0              0
7041                                0                0              1
7042                                0                1              1

      Dependents_Yes  PhoneService_No  PhoneService_Yes  Churn_No  Churn_Yes  \
0                  0                1                 0         1          0
1                  0                0                 1         1          0
2                  0                0                 1         0          1
3                  0                1                 0         1          0
4                  0                0                 1         0          1
```

```
…              …              …               …       …        …
7038               1                0                 1         1          0
7039               1                0                 1         1          0
7040               1                1                 0         1          0
7041               0                0                 1         0          1
7042               0                0                 1         1          0

      OnlineBackup_No  OnlineBackup_No internet service  OnlineBackup_Yes  \
0                   0                                 0                 1
1                   1                                 0                 0
2                   0                                 0                 1
3                   1                                 0                 0
4                   1                                 0                 0
…               …                               …               …
7038                1                                 0                 0
7039                0                                 0                 1
7040                1                                 0                 0
7041                1                                 0                 0
7042                1                                 0                 0

      Partner_No  Partner_Yes  PaymentMethod_Bank transfer (automatic)  \
0              0            1                                        0
1              1            0                                        0
2              1            0                                        0
3              1            0                                        1
4              1            0                                        0
…          …          …                                      …
7038           0            1                                        0
7039           0            1                                        0
7040           0            1                                        0
7041           0            1                                        0
7042           1            0                                        1

      PaymentMethod_Credit card (automatic)  PaymentMethod_Electronic check  \
0                                         0                               1
1                                         0                               0
2                                         0                               0
3                                         0                               0
4                                         0                               1
…                                     …                             …
7038                                      0                               0
7039                                      1                               0
7040                                      0                               1
7041                                      0                               0
7042                                      0                               0

      PaymentMethod_Mailed check  OnlineSecurity_No  \
0                              0                  1
1                              1                  0
2                              1                  0
3                              0                  0
4                              0                  1
…                          …                …
7038                           1                  0
7039                           0                  1
7040                           0                  0
7041                           1                  1
7042                           0                  0

      OnlineSecurity_No internet service  OnlineSecurity_Yes  \
0                                      0                   0
1                                      0                   1
2                                      0                   1
3                                      0                   1
4                                      0                   0
…                                  …                 …
7038                                   0                   1
7039                                   0                   0
7040                                   0                   1
7041                                   0                   0
7042                                   0                   1

      Contract_Month-to-month  Contract_One year  Contract_Two year  \
0                           1                  0                  0
1                           0                  1                  0
2                           1                  0                  0
3                           0                  1                  0
4                           1                  0                  0
…                       …                …                …
7038                        0                  1                  0
7039                        0                  1                  0
7040                        1                  0                  0
7041                        1                  0                  0
7042                        0                  0                  1

      gender_Female  gender_Male  StreamingMovies_No  \
0                 1            0                   1
1                 0            1                   1
2                 0            1                   1
3                 0            1                   1
4                 1            0                   1
…             …          …                 …
7038              0            1                   0
7039              1            0                   0
```

```
7040              1            0                   1
7041              0            1                   1
7042              0            1                   0

      StreamingMovies_No internet service  StreamingMovies_Yes  \
0                                       0                    0
1                                       0                    0
2                                       0                    0
3                                       0                    0
4                                       0                    0
…                                   …                  …
7038                                    0                    1
7039                                    0                    1
7040                                    0                    0
7041                                    0                    0
7042                                    0                    1

      PaperlessBilling_No  PaperlessBilling_Yes  InternetService_DSL  \
0                       0                     1                    1
1                       1                     0                    1
2                       0                     1                    1
3                       1                     0                    1
4                       0                     1                    0
…                   …                   …                  …
7038                    0                     1                    1
7039                    0                     1                    0
7040                    0                     1                    1
7041                    0                     1                    0
7042                    0                     1                    0

      InternetService_Fiber optic  InternetService_No  MultipleLines_No  \
0                               0                   0                 0
1                               0                   0                 1
2                               0                   0                 1
3                               0                   0                 0
4                               1                   0                 1
…                           …                 …               …
7038                            0                   0                 0
7039                            1                   0                 0
7040                            0                   0                 0
7041                            1                   0                 0
7042                            1                   0                 1

      MultipleLines_No phone service  MultipleLines_Yes  DeviceProtection_No  \
0                                  1                  0                    1
1                                  0                  0                    0
2                                  0                  0                    1
```

```
3                                  1                  0                    0
4                                  0                  0                    1
…                              …                …                  …
7038                               0                  1                    0
7039                               0                  1                    0
7040                               1                  0                    1
7041                               0                  1                    1
7042                               0                  0                    0

      DeviceProtection_No internet service  DeviceProtection_Yes
0                                        0                     0
1                                        0                     1
2                                        0                     0
3                                        0                     1
4                                        0                     0
…                                    …                   …
7038                                     0                     1
7039                                     0                     1
7040                                     0                     0
7041                                     0                     0
7042                                     0                     1

[7043 rows x 47 columns]
```

[113]:
```
dfDummy = pd.get_dummies(df)
dfDummy.head()
```

[113]:
```
   SeniorCitizen  tenure  MonthlyCharges  TotalCharges  gender_Female  \
0              0       1           29.85         29.85              1
1              0      34           56.95       1889.50              0
2              0       2           53.85        108.15              0
3              0      45           42.30       1840.75              0
4              0       2           70.70        151.65              1

   gender_Male  Partner_No  Partner_Yes  Dependents_No  Dependents_Yes  \
0            0           0            1              1               0
1            1           1            0              1               0
2            1           1            0              1               0
3            1           1            0              1               0
4            0           1            0              1               0

   PhoneService_No  PhoneService_Yes  MultipleLines_No  \
0                1                 0                 0
1                0                 1                 1
2                0                 1                 1
3                1                 0                 0
4                0                 1                 1
```

```
   MultipleLines_No phone service  MultipleLines_Yes  InternetService_DSL  \
0                               1                  0                    1
1                               0                  0                    1
2                               0                  0                    1
3                               1                  0                    1
4                               0                  0                    0

   InternetService_Fiber optic  InternetService_No  OnlineSecurity_No  \
0                            0                   0                  1
1                            0                   0                  0
2                            0                   0                  0
3                            0                   0                  0
4                            1                   0                  1

   OnlineSecurity_No internet service  OnlineSecurity_Yes  OnlineBackup_No  \
0                                   0                   0                0
1                                   0                   1                1
2                                   0                   1                0
3                                   0                   1                1
4                                   0                   0                1

   OnlineBackup_No internet service  OnlineBackup_Yes  DeviceProtection_No  \
0                                 0                 1                    1
1                                 0                 0                    0
2                                 0                 1                    1
3                                 0                 0                    0
4                                 0                 0                    1

   DeviceProtection_No internet service  DeviceProtection_Yes  TechSupport_No  \
0                                     0                     0               1
1                                     0                     1               1
2                                     0                     0               1
3                                     0                     1               0
4                                     0                     0               1

   TechSupport_No internet service  TechSupport_Yes  StreamingTV_No  \
0                                0                0               1
1                                0                0               1
2                                0                0               1
3                                0                1               1
4                                0                0               1

   StreamingTV_No internet service  StreamingTV_Yes  StreamingMovies_No  \
0                                0                0                   1
1                                0                0                   1
2                                0                0                   1
```

```
3                                0                0                   1
4                                0                0                   1

   StreamingMovies_No internet service  StreamingMovies_Yes  \
0                                    0                    0
1                                    0                    0
2                                    0                    0
3                                    0                    0
4                                    0                    0

   Contract_Month-to-month  Contract_One year  Contract_Two year  \
0                        1                  0                  0
1                        0                  1                  0
2                        1                  0                  0
3                        0                  1                  0
4                        1                  0                  0

   PaperlessBilling_No  PaperlessBilling_Yes  \
0                    0                     1
1                    1                     0
2                    0                     1
3                    1                     0
4                    0                     1

   PaymentMethod_Bank transfer (automatic)  \
0                                        0
1                                        0
2                                        0
3                                        1
4                                        0

   PaymentMethod_Credit card (automatic)  PaymentMethod_Electronic check  \
0                                      0                               1
1                                      0                               0
2                                      0                               0
3                                      0                               0
4                                      0                               1

   PaymentMethod_Mailed check  Churn_No  Churn_Yes
0                           0         1          0
1                           1         1          0
2                           1         0          1
3                           0         1          0
4                           0         0          1
```

[114]:

```
# df = pd.concat([df, pd.get_dummies(df)], axis=1)
dfDummy.pop("Churn_No")
```

[114]:
```
0       1
1       1
2       0
3       1
4       0
       ..
7038    1
7039    1
7040    1
7041    0
7042    1
Name: Churn_No, Length: 7043, dtype: uint8
```

[115]:
```
dfDummy = dfDummy.rename(columns={'Churn_No': 'Churn'})
```

[116]:
```
dfDummy.head()
```

[116]:
```
   SeniorCitizen  tenure  MonthlyCharges  TotalCharges  gender_Female  \
0              0       1           29.85         29.85              1
1              0      34           56.95       1889.50              0
2              0       2           53.85        108.15              0
3              0      45           42.30       1840.75              0
4              0       2           70.70        151.65              1

   gender_Male  Partner_No  Partner_Yes  Dependents_No  Dependents_Yes  \
0            0           0            1              1               0
1            1           1            0              1               0
2            1           1            0              1               0
3            1           1            0              1               0
4            0           1            0              1               0

   PhoneService_No  PhoneService_Yes  MultipleLines_No  \
0                1                 0                 0
1                0                 1                 1
2                0                 1                 1
3                1                 0                 0
4                0                 1                 1

   MultipleLines_No phone service  MultipleLines_Yes  InternetService_DSL  \
0                               1                  0                    1
1                               0                  0                    1
2                               0                  0                    1
3                               1                  0                    1
4                               0                  0                    0

   InternetService_Fiber optic  InternetService_No  OnlineSecurity_No  \
0                            0                   0                  1
```

```
1                            0                   0                  0
2                            0                   0                  0
3                            0                   0                  0
4                            1                   0                  1

   OnlineSecurity_No internet service  OnlineSecurity_Yes  OnlineBackup_No  \
0                                   0                   0                0
1                                   0                   1                1
2                                   0                   1                0
3                                   0                   1                1
4                                   0                   0                1

   OnlineBackup_No internet service  OnlineBackup_Yes  DeviceProtection_No  \
0                                 0                 1                    1
1                                 0                 0                    0
2                                 0                 1                    1
3                                 0                 0                    0
4                                 0                 0                    1

   DeviceProtection_No internet service  DeviceProtection_Yes  TechSupport_No  \
0                                     0                     0               1
1                                     0                     1               1
2                                     0                     0               1
3                                     0                     1               0
4                                     0                     0               1

   TechSupport_No internet service  TechSupport_Yes  StreamingTV_No  \
0                                0                0               1
1                                0                0               1
2                                0                0               1
3                                0                1               1
4                                0                0               1

   StreamingTV_No internet service  StreamingTV_Yes  StreamingMovies_No  \
0                                0                0                   1
1                                0                0                   1
2                                0                0                   1
3                                0                0                   1
4                                0                0                   1

   StreamingMovies_No internet service  StreamingMovies_Yes  \
0                                    0                    0
1                                    0                    0
2                                    0                    0
3                                    0                    0
4                                    0                    0
```

```
   Contract_Month-to-month  Contract_One year  Contract_Two year  \
0                        1                  0                  0
1                        0                  1                  0
2                        1                  0                  0
3                        0                  1                  0
4                        1                  0                  0

   PaperlessBilling_No  PaperlessBilling_Yes  \
0                    0                     1
1                    1                     0
2                    0                     1
3                    1                     0
4                    0                     1

   PaymentMethod_Bank transfer (automatic)  \
0                                        0
1                                        0
2                                        0
3                                        1
4                                        0

   PaymentMethod_Credit card (automatic)  PaymentMethod_Electronic check  \
0                                      0                               1
1                                      0                               0
2                                      0                               0
3                                      0                               0
4                                      0                               1

   PaymentMethod_Mailed check  Churn_Yes
0                           0          0
1                           1          0
2                           1          1
3                           0          0
4                           0          1
```

1.5 3 Modelling

```
[117]: dfModel = dfDummy.copy(deep = True)
```

```
[118]: dfModel.shape
```

```
[118]: (7043, 46)
```

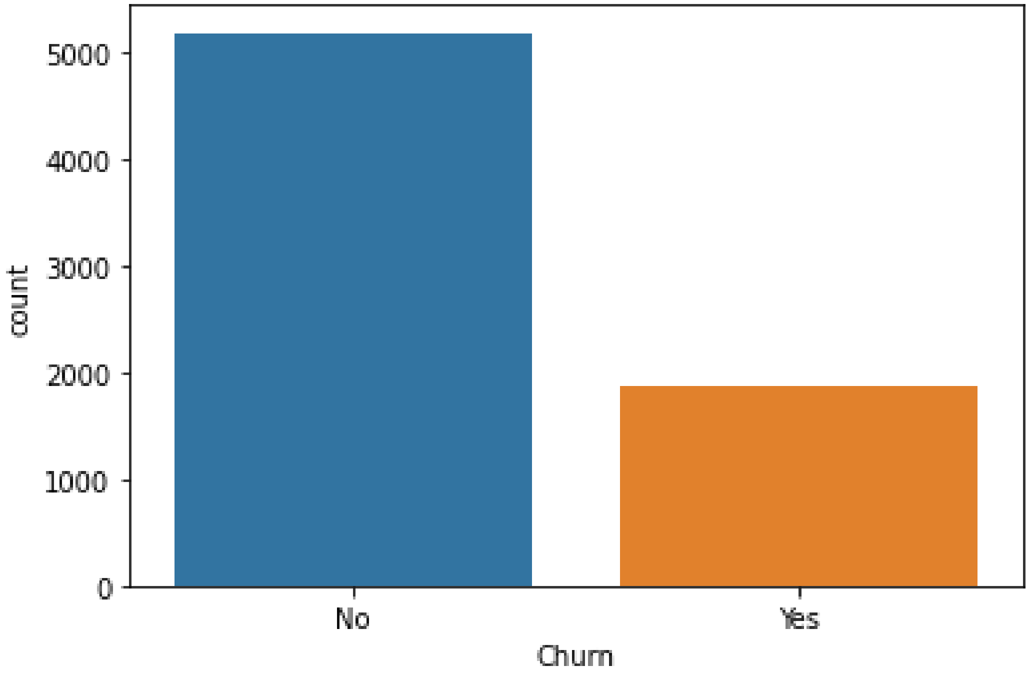
```
[119]: # Printing all the columns with atleast one null value
       dfModel.columns[dfModel.isna().any()].tolist()
```

```
[119]: ['TotalCharges']
```

```
[120]: dfModel['TotalCharges'].fillna(dfModel['TotalCharges'].mode()[0], inplace=True)
```

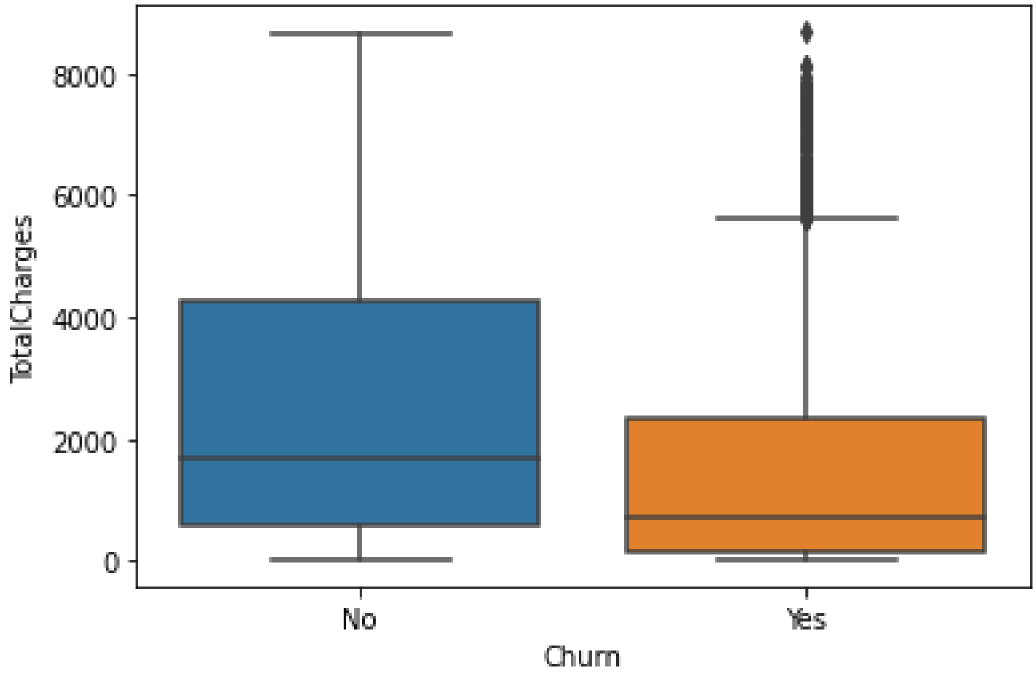
```
[121]: sns.countplot(data = df, x ='Churn') #BaggingBalancedClassifier
```

```
[121]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4398342e10>
```



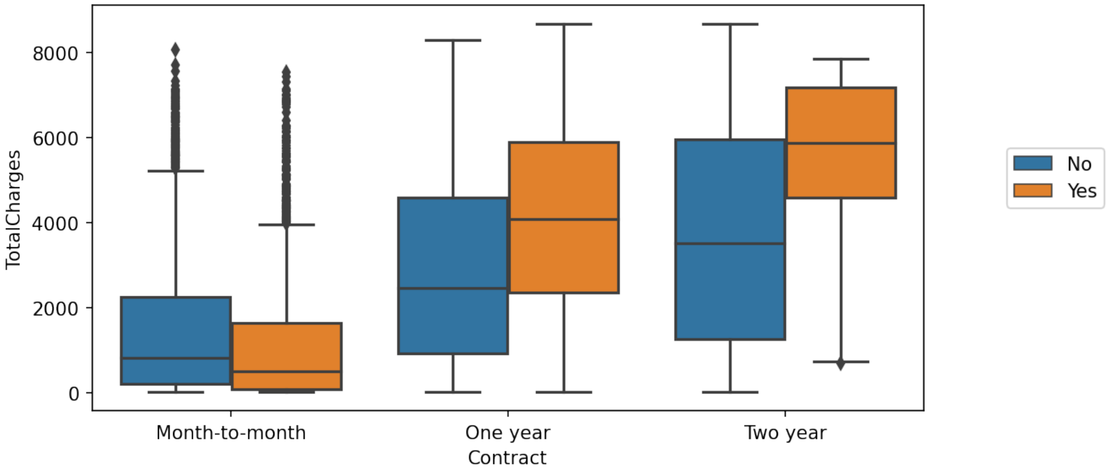
```
[122]: sns.boxplot(data = df, x = 'Churn', y= 'TotalCharges')
```

```
[122]: <matplotlib.axes._subplots.AxesSubplot at 0x7f439e1d2150>
```



```
[123]: plt.figure(figsize = (8,4), dpi = 150)
       sns.boxplot(data = df, y = 'TotalCharges', x = 'Contract', hue = 'Churn')
       plt.legend(loc = (1.1, 0.5))
```

```
[123]: <matplotlib.legend.Legend at 0x7f439e6760d0>
```



```
[124]: df.columns
```

[124]:
```
Index(['gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure',
       'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity',
       'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV',
       'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod',
       'MonthlyCharges', 'TotalCharges', 'Churn'],
      dtype='object')
```

[125]:
```
corr_df = pd.get_dummies(df[['gender', 'SeniorCitizen', 'Partner',␣
 ↪'Dependents','PhoneService', 'MultipleLines',
 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',␣
 ↪'InternetService',
   'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling',␣
 ↪'PaymentMethod', 'Churn']]).corr()
```

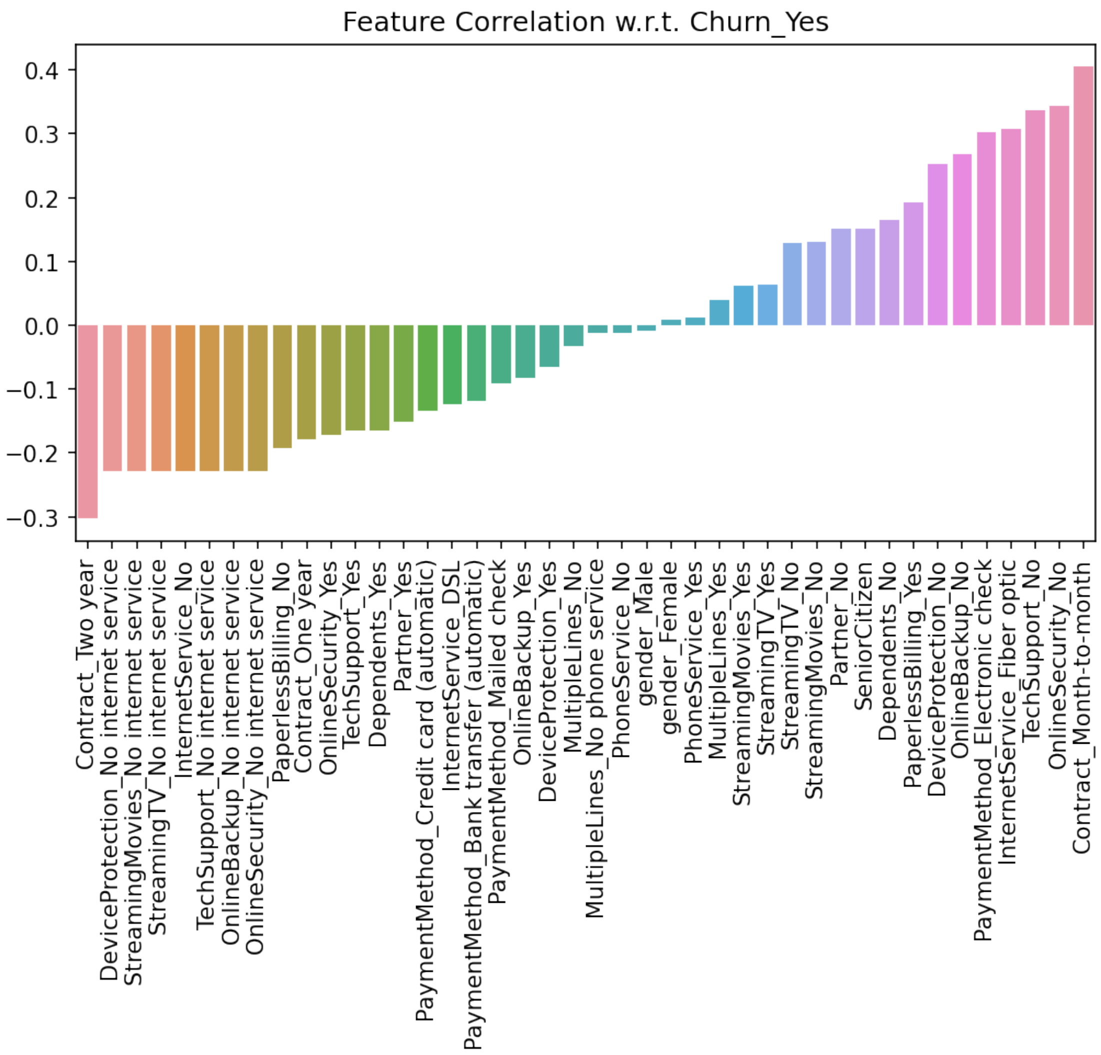
[126]:
```
corr_df['Churn_Yes'].sort_values().iloc[1:-1]
```

[126]:
```
Contract_Two year                          -0.302253
DeviceProtection_No internet service       -0.227890
StreamingMovies_No internet service        -0.227890
StreamingTV_No internet service            -0.227890
InternetService_No                         -0.227890
TechSupport_No internet service            -0.227890
OnlineBackup_No internet service           -0.227890
OnlineSecurity_No internet service         -0.227890
PaperlessBilling_No                        -0.191825
Contract_One year                          -0.177820
OnlineSecurity_Yes                         -0.171226
TechSupport_Yes                            -0.164674
Dependents_Yes                             -0.164221
Partner_Yes                                -0.150448
PaymentMethod_Credit card (automatic)      -0.134302
InternetService_DSL                        -0.124214
PaymentMethod_Bank transfer (automatic)    -0.117937
PaymentMethod_Mailed check                 -0.091683
OnlineBackup_Yes                           -0.082255
DeviceProtection_Yes                       -0.066160
MultipleLines_No                           -0.032569
MultipleLines_No phone service             -0.011942
PhoneService_No                            -0.011942
gender_Male                                -0.008612
gender_Female                               0.008612
PhoneService_Yes                            0.011942
MultipleLines_Yes                           0.040102
StreamingMovies_Yes                         0.061382
StreamingTV_Yes                             0.063228
StreamingTV_No                              0.128916
StreamingMovies_No                          0.130845
```

```
Partner_No                                0.150448
SeniorCitizen                             0.150889
Dependents_No                             0.164221
PaperlessBilling_Yes                      0.191825
DeviceProtection_No                       0.252481
OnlineBackup_No                           0.268005
PaymentMethod_Electronic check            0.301919
InternetService_Fiber optic               0.308020
TechSupport_No                            0.337281
OnlineSecurity_No                         0.342637
Contract_Month-to-month                   0.405103
Name: Churn_Yes, dtype: float64
```

[127]:

```
plt.figure(figsize = (8,4), dpi =150)
sns.barplot(x = corr_df['Churn_Yes'].sort_values().iloc[1:-1].index, y =␣
 ↪corr_df['Churn_Yes'].sort_values().iloc[1:-1].values)
plt.title("Feature Correlation w.r.t. Churn_Yes")
plt.xticks(rotation = 90);
```

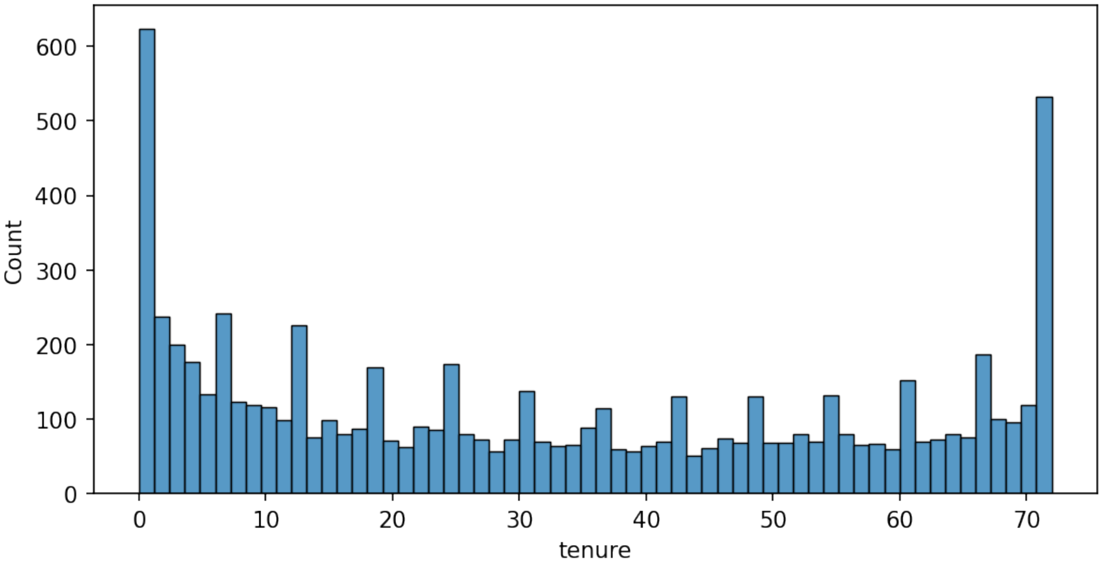


```
[128]: df['Contract'].unique()
```

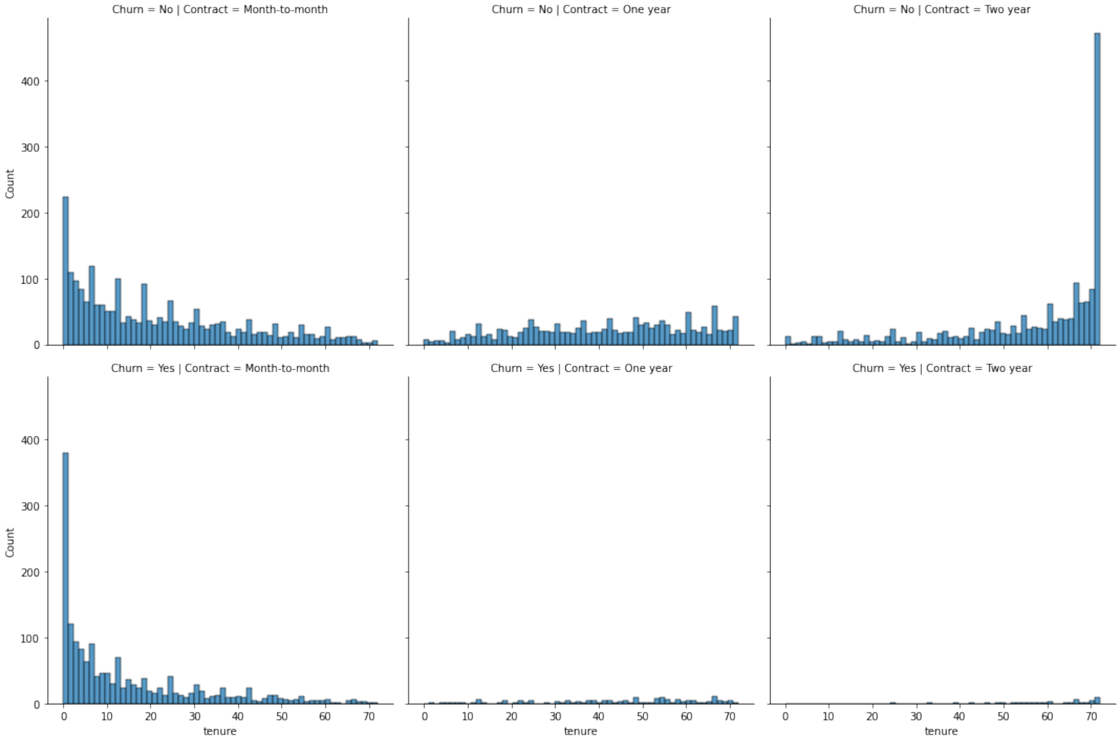
```
[128]: array(['Month-to-month', 'One year', 'Two year'], dtype=object)
```

```
[129]: plt.figure(figsize = (8,4), dpi =150)
       sns.histplot(data =df, x= 'tenure', bins = 60)
```

```
[129]: <matplotlib.axes._subplots.AxesSubplot at 0x7f43909b7a10>
```

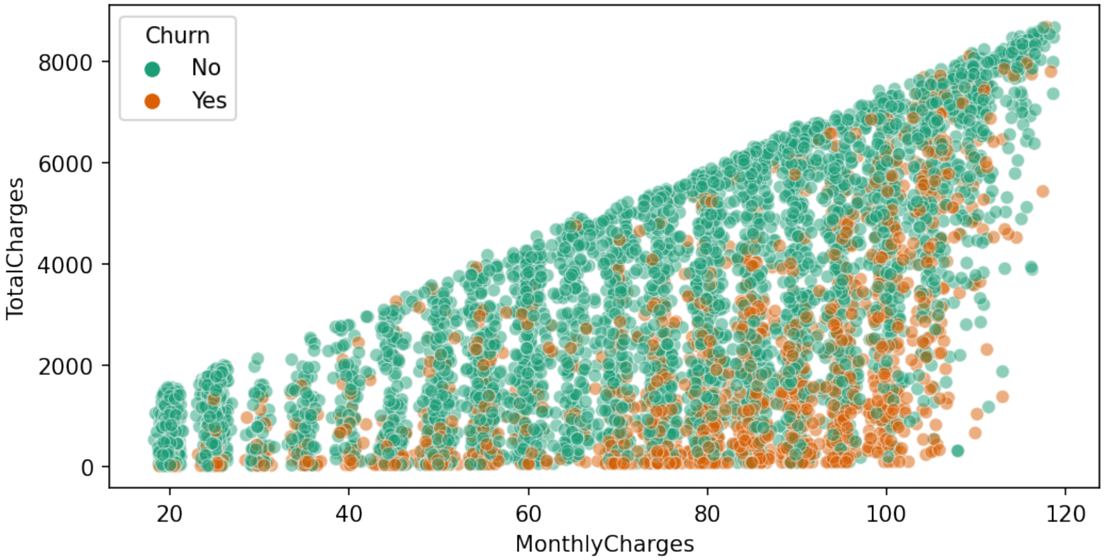


[130]:
```
sns.displot(data = df, x= 'tenure', bins = 60, col = 'Contract', row = 'Churn');
```



[131]:
```
plt.figure(figsize = (8,4), dpi =150)
sns.scatterplot(data =df, x = 'MonthlyCharges', y= 'TotalCharges', hue =␣
↪'Churn', alpha = 0.5, palette = 'Dark2')
```

[131]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4394695fd0>
```



[132]:
```
no_churn = df.groupby(['Churn', 'tenure']).count().transpose()['No']
yes_churn = df.groupby(['Churn', 'tenure']).count().transpose()['Yes']
```

[133]:
```
churn_rate =  100*yes_churn/(no_churn+yes_churn)
```

[134]:
```
def cohert(tenure):
    if tenure <13:
        return '0-12 months'
    elif tenure < 25:
        return '12-24 months'
    elif tenure < 49:
        return '24-48 months'

    else:
        return 'over 48 months'
```

[135]:
```
df['Tenure Cohort'] = df['tenure'].apply(cohert)
```

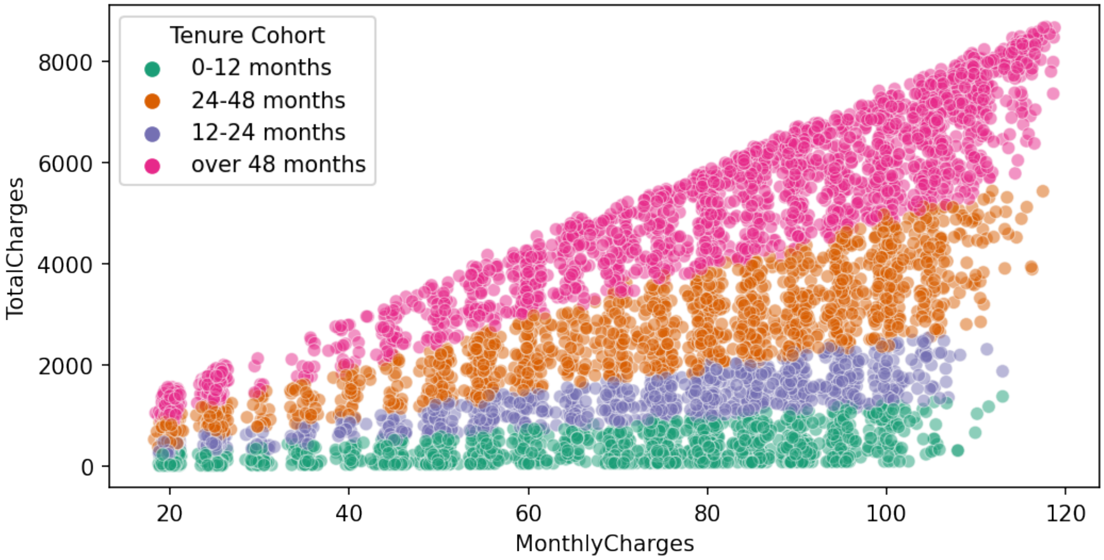
[136]:
```
df.head(10)[['tenure', 'Tenure Cohort']]
```

[136]:
```
   tenure   Tenure Cohort
0       1     0-12 months
```

```
1       34     24-48 months
2        2      0-12 months
3       45     24-48 months
4        2      0-12 months
5        8      0-12 months
6       22     12-24 months
7       10      0-12 months
8       28     24-48 months
9       62   over 48 months
```

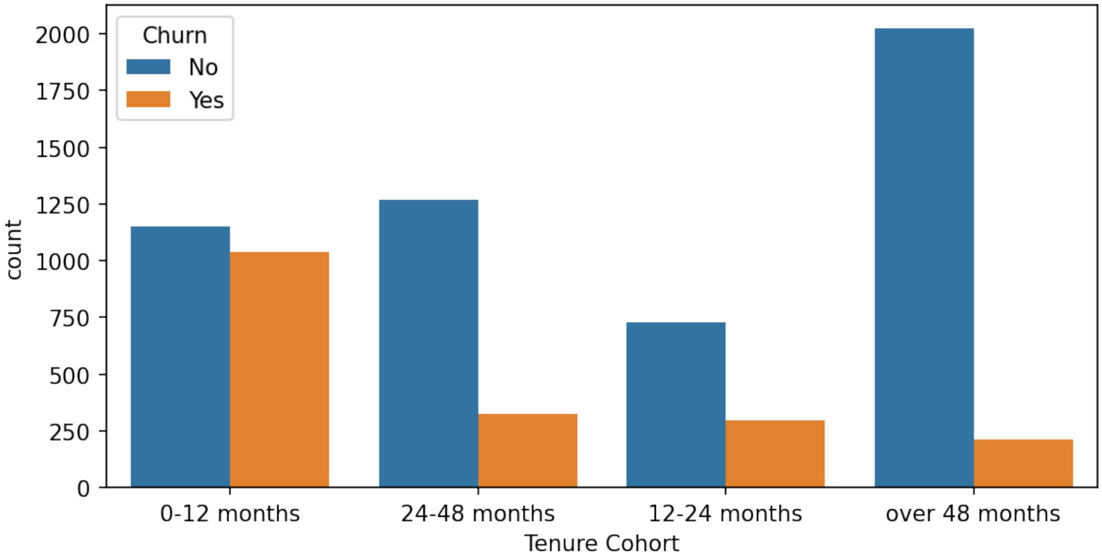
[137]:
```
plt.figure(figsize = (8,4), dpi =150)
sns.scatterplot(data =df, x = 'MonthlyCharges', y= 'TotalCharges', hue =␣
 ↪'Tenure Cohort', alpha = 0.5, palette = 'Dark2')
```

[137]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f439e1a6c50>
```



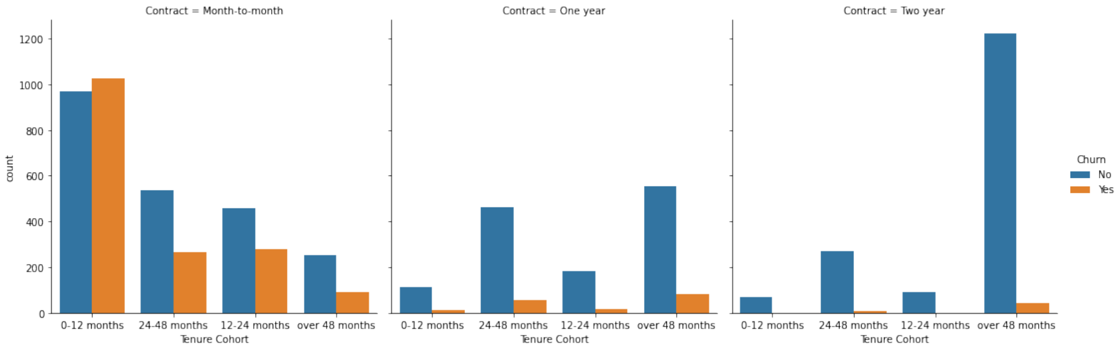
[138]:
```
plt.figure(figsize = (8,4), dpi =150)
sns.countplot(data = df, x = 'Tenure Cohort', hue = 'Churn')
```

[138]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f439f929b10>
```



[139]:
```
sns.catplot(data = df, x ='Tenure Cohort', hue = 'Churn', col ='Contract', kind␣
↪= 'count')
```

[139]:
```
<seaborn.axisgrid.FacetGrid at 0x7f439e57e610>
```



[140]:
```
X = dfModel.iloc[:, :-1]
y = dfModel.iloc[:, -1]
```

[141]:
```
X.shape
```

[141]:
```
(7043, 45)
```

[142]:
```
y.shape
```

[142]:
```
(7043,)
```

[143]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,␣
 ↪random_state=42)
```

[144]:

```
def modelConfusionMatrix(y_test, y_pred):
  '''
  Function to print the confusion matrix
  Inputs:
  @y_test (dataframe) - Test Dataframe
  @y_pred (dataframe) - Predicted Dataframe

  Output:
  @cnf_matrix (2D Array) - Confusion Matrix
  '''
  cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
  return cnf_matrix

def modelConfusionMatrixVisual(cnf_matrix):
  '''
  Function to print confusion matrix
  '''
  class_names=[0,1] # name  of classes
  fig, ax = plt.subplots()
  tick_marks = np.arange(len(class_names))
  plt.xticks(tick_marks, class_names)
  plt.yticks(tick_marks, class_names)
  # create heatmap
  sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
  ax.xaxis.set_label_position("top")
  plt.tight_layout()
  plt.title('Confusion matrix', y=1.1)
  plt.ylabel('Actual label')
  plt.xlabel('Predicted label')

def modelAccuracyMatrix(y_test, y_pred, modelName):
  '''
  Function to print the model accuracy
  '''
  accuracy = round((metrics.accuracy_score(y_test, y_pred).round(4))*100, 2)
  print("Accuracy (" + modelName +"): " + str(accuracy) + "%")

  precision = round((metrics.precision_score(y_test, y_pred).round(4))*100, 2)
  print("Precision (" + modelName +"): " + str(precision) + "%")

  recall = round((metrics.recall_score(y_test, y_pred).round(4))*100, 2)
  print("Recall (" + modelName +"): " + str(recall) + "%")

def modelRocCurve(X_test, y_test):
  '''
  Function to print the AUC-ROC Curve
  '''
  y_pred_proba = lm.predict_proba(X_test)[::,1]
  fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
  auc = metrics.roc_auc_score(y_test, y_pred_proba)
  plt.plot(fpr,tpr,label="data 1, auc="+str(auc))
  plt.legend(loc=4)
  plt.show()
```

1.5.1 3.1 Logistic Regression

[145]:
```
# import the metrics class #modelConfusionMatrix
from sklearn import metrics

# import required modules
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

[146]:
```
# Importing RFE and LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

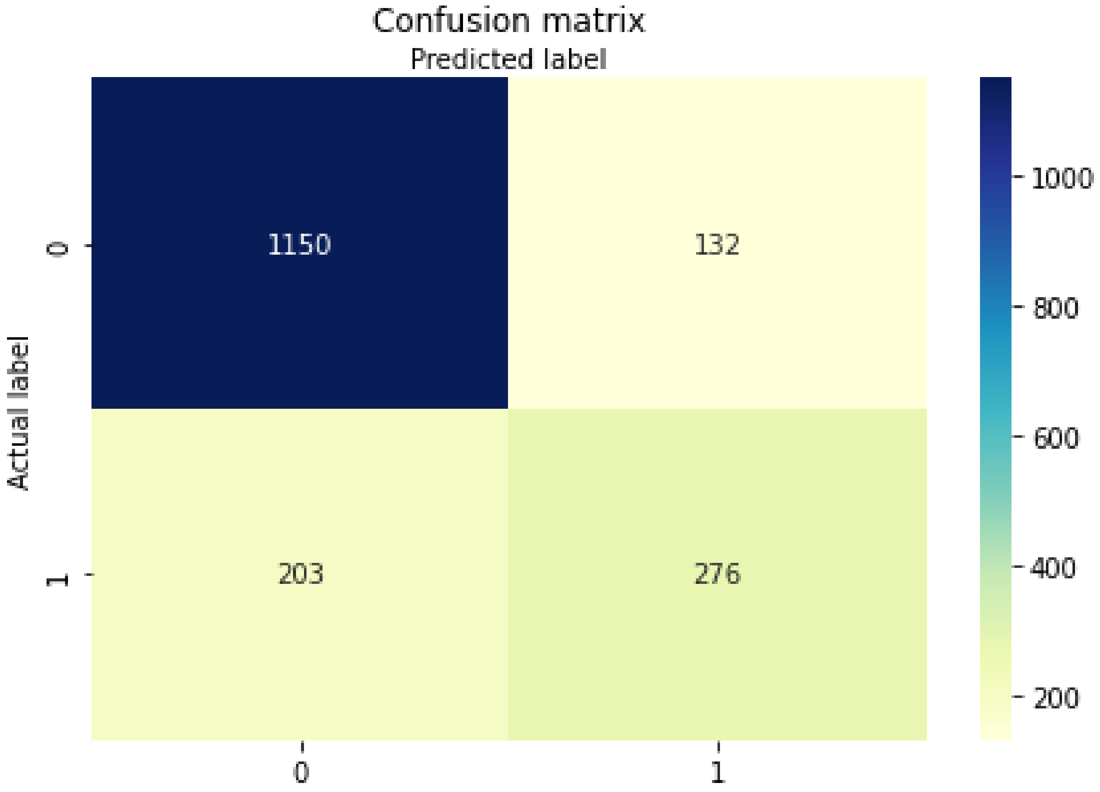
[147]:
```
# Running RFE with the output number of the variable equal to 20

# Making a linear regression model object
lm = LogisticRegression()

# Fitting the model on the training dataset
lm.fit(X_train, y_train)

# Outputting the top 20 features
# rfe = RFE(lm, 20)
# rfe = rfe.fit(X_train, y_train)
```

[147]:
```
LogisticRegression()
```

[148]:
```
# Making predictions
y_pred = lm.predict(X_test)
```

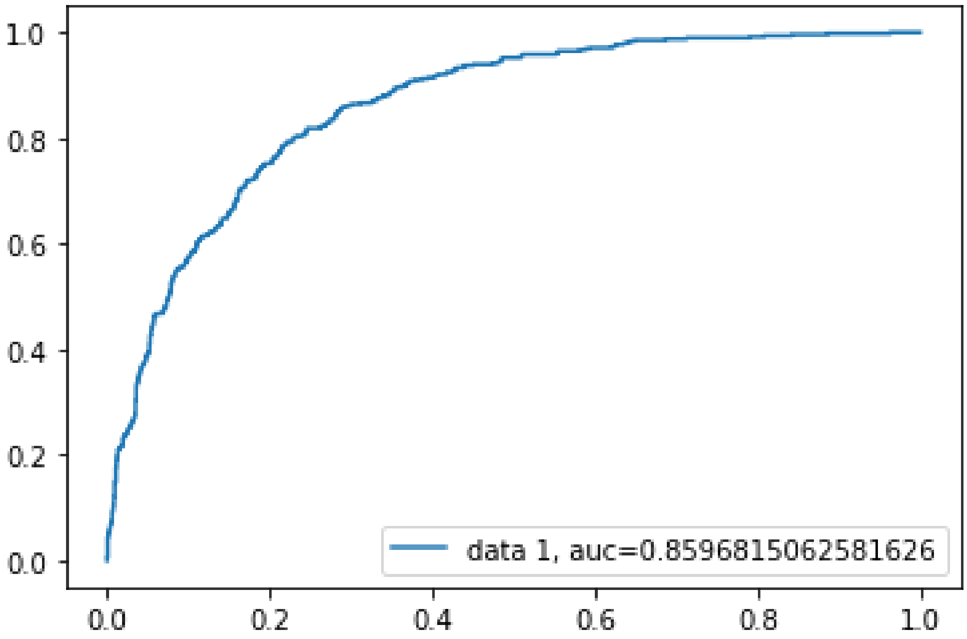
[149]:
```
# Assigning the model name to variable modelName
modelName = "Logistic Regression"

# printing the model name
print(modelName)
print("\n")

# get the confusion matrix
cnf_matrix = modelConfusionMatrix(y_test, y_pred)
print(cnf_matrix)
print("\n")

# visualize the confusion matrix
modelConfusionMatrixVisual(cnf_matrix)
modelAccuracyMatrix(y_test, y_pred, modelName)
print("\n")
```

```
Logistic Regression


[[1150  132]
 [ 203  276]]


Accuracy (Logistic Regression): 80.98%
Precision (Logistic Regression): 67.65%
Recall (Logistic Regression): 57.62%
```



[150]:

```
# Area under ROC Curve
modelRocCurve(X_test, y_test)
```



1.5.2 3.2 K-Nearest Neighbour

[151]:

```
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors=5)
classifier.fit(X_train, y_train)
```

[151]:

```
KNeighborsClassifier()
```

[152]:

```
y_pred = classifier.predict(X_test)
```

[154]:

```
from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

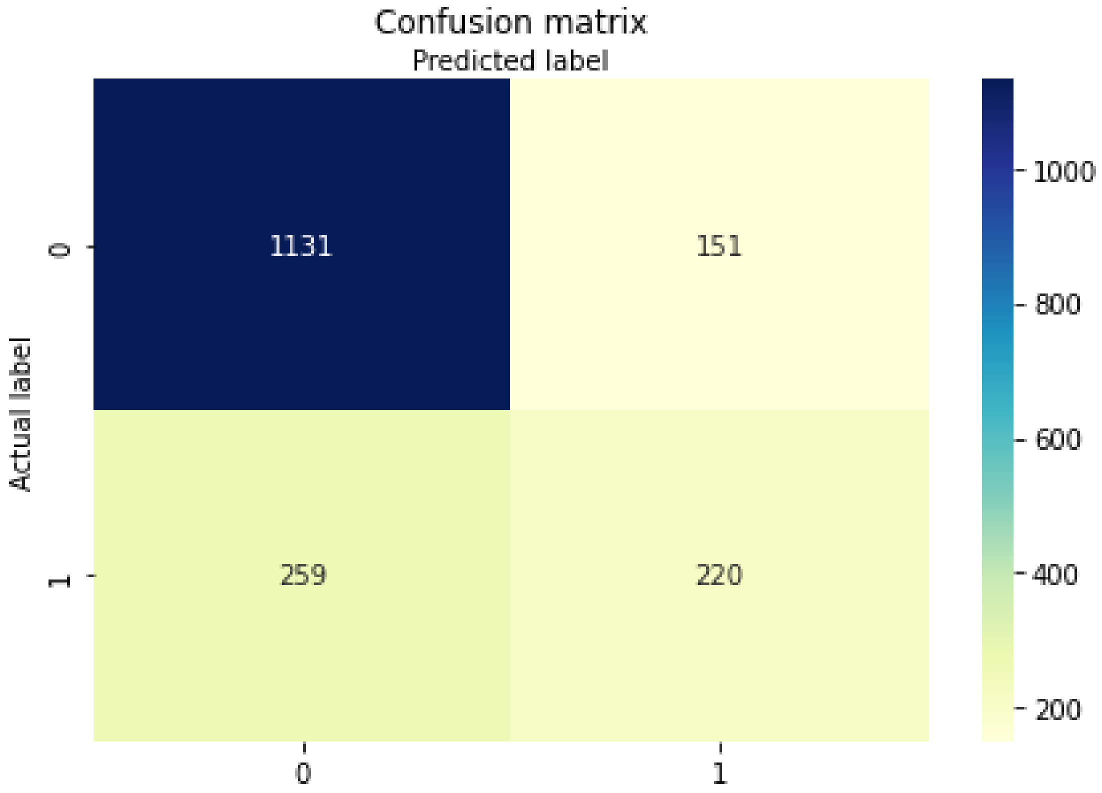
```
[[1131  151]
 [ 259  220]]
              precision    recall  f1-score   support

           0       0.81      0.88      0.85      1282
           1       0.59      0.46      0.52       479
```

```
    accuracy                           0.77      1761
   macro avg       0.70      0.67      0.68      1761
weighted avg       0.75      0.77      0.76      1761
```

[155]:

```
# Assigning the model name to variable modelName
modelName = "K-Nearest Neighbour"

# printing the model name
print(modelName)
print("\n")

# get the confusion matrix
cnf_matrix = modelConfusionMatrix(y_test, y_pred)
print(cnf_matrix)
print("\n")

# visualize the confusion matrix
modelConfusionMatrixVisual(cnf_matrix)
modelAccuracyMatrix(y_test, y_pred, modelName)
print("\n")
```

```
K-Nearest Neighbour


[[1131  151]
 [ 259  220]]


Accuracy (K-Nearest Neighbour): 76.72%
Precision (K-Nearest Neighbour): 59.3%
Recall (K-Nearest Neighbour): 45.93%
```



1.5.3 3.3 Random Forest

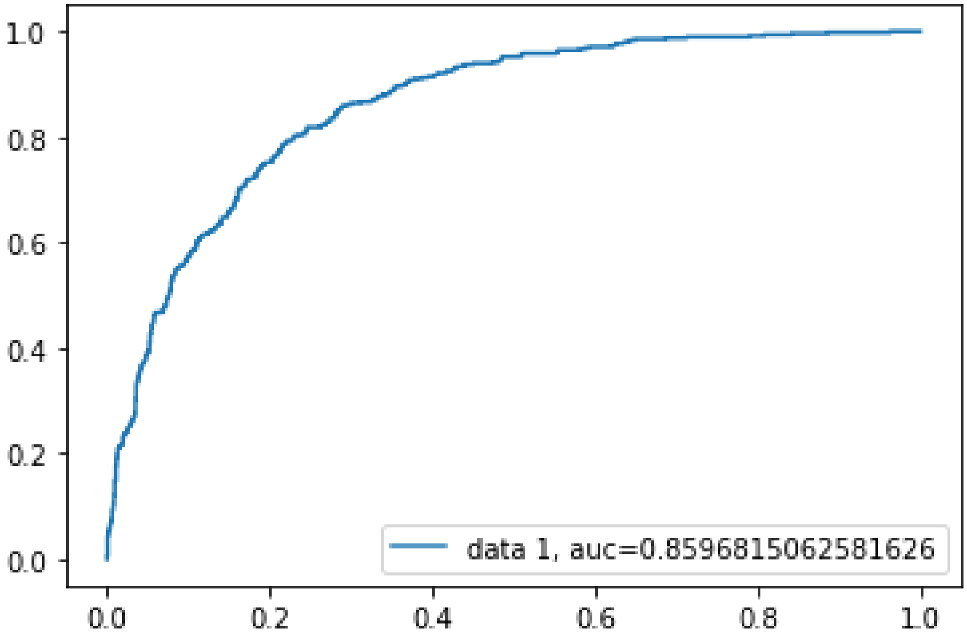
[156]:
```
from sklearn.ensemble import RandomForestRegressor

regressor = RandomForestRegressor(n_estimators=20, random_state=0)
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)
```

[157]:
```
from sklearn import metrics

# Assigning the model name to variable modelName
modelName = "Random Forest"

accuracy = round((metrics.mean_absolute_error(y_test, y_pred).round(4))*100, 2)
print("Mean Absolute Error: (" + modelName +"): " + str(accuracy) + "%")

accuracy = round((metrics.mean_squared_error(y_test, y_pred).round(4))*100, 2)
print("Mean Squared Error: (" + modelName +"): " + str(accuracy) + "%")

accuracy = round((metrics.mean_squared_error(y_test, y_pred).round(4))*100, 2)
print("Root Mean Squared Error: (" + modelName +"): " + str(accuracy) + "%")
```

```
Mean Absolute Error: (Random Forest): 27.3%
Mean Squared Error: (Random Forest): 14.94%
Root Mean Squared Error: (Random Forest): 14.94%
```

[158]:
```
# printing the model name
print(modelName)
print("\n")

# Area under ROC Curve
modelRocCurve(X_test, y_test)
```

```
Random Forest
```



1.5.4 3.4 Decision Trees

[159]:
```
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier()
classifier.fit(X_train, y_train)
```

[159]:
```
DecisionTreeClassifier()
```

[160]:
```
y_pred = classifier.predict(X_test)
```

[69]:

```
from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
[[1058  224]
 [ 239  240]]
              precision    recall  f1-score   support

           0       0.82      0.83      0.82      1282
           1       0.52      0.50      0.51       479

    accuracy                           0.74      1761
   macro avg       0.67      0.66      0.66      1761
weighted avg       0.73      0.74      0.74      1761
```

[161]:

```
from sklearn import metrics

# Assigning the model name to variable modelName
modelName = "Decision Trees"

accuracy = round((metrics.mean_absolute_error(y_test, y_pred).round(4))*100, 2)
print("Mean Absolute Error: (" + modelName +"): " + str(accuracy) + "%")

accuracy = round((metrics.mean_squared_error(y_test, y_pred).round(4))*100, 2)
print("Mean Squared Error: (" + modelName +"): " + str(accuracy) + "%")

accuracy = round((metrics.mean_squared_error(y_test, y_pred).round(4))*100, 2)
print("Root Mean Squared Error: (" + modelName +"): " + str(accuracy) + "%")
```

```
Mean Absolute Error: (Decision Trees): 3487.9%
Mean Squared Error: (Decision Trees): 26.24%
Root Mean Squared Error: (Decision Trees): 26.24%
```

1.5.5 3.5 XGBoost

[162]:

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.ensemble import GradientBoostingClassifier
```

[163]:

```
lr_list = [0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1]

for learning_rate in lr_list:
    gb_clf = GradientBoostingClassifier(n_estimators=20,␣
 ↪learning_rate=learning_rate, max_features=2, max_depth=2, random_state=0)
    gb_clf.fit(X_train, y_train)
```

[164]:

```
print("Learning rate: ", learning_rate)
print("Accuracy score (training): {0:.3f}".format(gb_clf.score(X_train,␣
 ↪y_train)))
print("Accuracy score (validation): {0:.3f}".format(gb_clf.score(X_test,␣
 ↪y_test)))
```

```
Learning rate:  1
Accuracy score (training): 0.797
Accuracy score (validation): 0.798
```

[165]:

```
gb_clf2 = GradientBoostingClassifier(n_estimators=20, learning_rate=1,␣
 ↪max_features=2, max_depth=2, random_state=0)
gb_clf2.fit(X_train, y_train)
predictions = gb_clf2.predict(X_test)

print("Confusion Matrix:")
print(confusion_matrix(y_test, predictions))

print("Classification Report")
print(classification_report(y_test, predictions))
```

```
Confusion Matrix:
[[1161  121]
 [ 235  244]]
Classification Report
              precision    recall  f1-score   support

           0       0.83      0.91      0.87      1282
           1       0.67      0.51      0.58       479

    accuracy                           0.80      1761
   macro avg       0.75      0.71      0.72      1761
weighted avg       0.79      0.80      0.79      1761
```

[167]:

```
!jupyter nbconvert --to PDF "M.Sc_data_analytics_Project.ipynb"
```

```
[NbConvertApp] WARNING | pattern 'M.Sc_data_analytics_Project.ipynb' matched no
files
This application is used to convert notebook files (*.ipynb)
        to various other formats.

        WARNING: THE COMMANDLINE INTERFACE MAY CHANGE IN FUTURE RELEASES.

Options
=======
The options below are convenience aliases to configurable class-options,
as listed in the "Equivalent to" description-line of the aliases.
To see all configurable class-options for some <cmd>, use:
    <cmd> --help-all

--debug
    set log level to logging.DEBUG (maximize logging output)
    Equivalent to: [--Application.log_level=10]
--show-config
    Show the application's configuration (human-readable format)
    Equivalent to: [--Application.show_config=True]
--show-config-json
    Show the application's configuration (json format)
    Equivalent to: [--Application.show_config_json=True]
--generate-config
    generate default config file
    Equivalent to: [--JupyterApp.generate_config=True]
-y
    Answer yes to any questions instead of prompting.
    Equivalent to: [--JupyterApp.answer_yes=True]
--execute
    Execute the notebook prior to export.
    Equivalent to: [--ExecutePreprocessor.enabled=True]
--allow-errors
    Continue notebook execution even if one of the cells throws an error and
include the error message in the cell output (the default behaviour is to abort
conversion). This flag is only relevant if '--execute' was specified, too.
    Equivalent to: [--ExecutePreprocessor.allow_errors=True]
--stdin
    read a single notebook file from stdin. Write the resulting notebook with
default basename 'notebook.*'
    Equivalent to: [--NbConvertApp.from_stdin=True]
--stdout
    Write notebook output to stdout instead of files.
    Equivalent to: [--NbConvertApp.writer_class=StdoutWriter]
--inplace
    Run nbconvert in place, overwriting the existing notebook (only
            relevant when converting to notebook format)
    Equivalent to: [--NbConvertApp.use_output_suffix=False
--NbConvertApp.export_format=notebook --FilesWriter.build_directory=]
--clear-output
    Clear output of current file and save in place,
            overwriting the existing notebook.
    Equivalent to: [--NbConvertApp.use_output_suffix=False
--NbConvertApp.export_format=notebook --FilesWriter.build_directory=
--ClearOutputPreprocessor.enabled=True]
--no-prompt
    Exclude input and output prompts from converted document.
    Equivalent to: [--TemplateExporter.exclude_input_prompt=True
--TemplateExporter.exclude_output_prompt=True]
--no-input
```

```
    Exclude input cells and output prompts from converted document.
            This mode is ideal for generating code-free reports.
    Equivalent to: [--TemplateExporter.exclude_output_prompt=True
--TemplateExporter.exclude_input=True
--TemplateExporter.exclude_input_prompt=True]
--allow-chromium-download
    Whether to allow downloading chromium if no suitable version is found on the
system.
    Equivalent to: [--WebPDFExporter.allow_chromium_download=True]
--disable-chromium-sandbox
    Disable chromium security sandbox when converting to PDF..
    Equivalent to: [--WebPDFExporter.disable_sandbox=True]
--show-input
    Shows code input. This flag is only useful for dejavu users.
    Equivalent to: [--TemplateExporter.exclude_input=False]
--embed-images
    Embed the images as base64 dataurls in the output. This flag is only useful
for the HTML/WebPDF/Slides exports.
    Equivalent to: [--HTMLExporter.embed_images=True]
--log-level=<Enum>
    Set the log level by value or name.
    Choices: any of [0, 10, 20, 30, 40, 50, 'DEBUG', 'INFO', 'WARN', 'ERROR',
'CRITICAL']
    Default: 30
    Equivalent to: [--Application.log_level]
--config=<Unicode>
    Full path of a config file.
    Default: ''
    Equivalent to: [--JupyterApp.config_file]
--to=<Unicode>
    The export format to be used, either one of the built-in formats
            ['asciidoc', 'custom', 'html', 'latex', 'markdown', 'notebook',
'pdf', 'python', 'rst', 'script', 'slides', 'webpdf']
            or a dotted object name that represents the import path for an
            ``Exporter`` class
    Default: ''
    Equivalent to: [--NbConvertApp.export_format]
--template=<Unicode>
    Name of the template to use
    Default: ''
    Equivalent to: [--TemplateExporter.template_name]
--template-file=<Unicode>
    Name of the template file to use
    Default: None
    Equivalent to: [--TemplateExporter.template_file]
--theme=<Unicode>
    Template specific theme(e.g. the name of a JupyterLab CSS theme distributed
    as prebuilt extension for the lab template)
```

```
  Default: 'light'
  Equivalent to: [--HTMLExporter.theme]
--writer=<DottedObjectName>
  Writer class used to write the
                    results of the conversion
  Default: 'FilesWriter'
  Equivalent to: [--NbConvertApp.writer_class]
--post=<DottedOrNone>
  PostProcessor class used to write the
                    results of the conversion
  Default: ''
  Equivalent to: [--NbConvertApp.postprocessor_class]
--output=<Unicode>
  overwrite base name use for output files.
        can only be used when converting one notebook at a time.
  Default: ''
  Equivalent to: [--NbConvertApp.output_base]
--output-dir=<Unicode>
  Directory to write output(s) to. Defaults
                 to output to the directory of each notebook.
To recover
                 previous default behaviour (outputting to the
current
                 working directory) use . as the flag value.
  Default: ''
  Equivalent to: [--FilesWriter.build_directory]
--reveal-prefix=<Unicode>
  The URL prefix for reveal.js (version 3.x).
      This defaults to the reveal CDN, but can be any url pointing to a
copy
      of reveal.js.
      For speaker notes to work, this must be a relative path to a local
      copy of reveal.js: e.g., "reveal.js".
      If a relative path is given, it must be a subdirectory of the
      current directory (from which the server is run).
      See the usage documentation
      (https://nbconvert.readthedocs.io/en/latest/usage.html#reveal-js-
html-slideshow)
      for more details.
  Default: ''
  Equivalent to: [--SlidesExporter.reveal_url_prefix]
--nbformat=<Enum>
  The nbformat version to write.
      Use this to downgrade notebooks.
  Choices: any of [1, 2, 3, 4]
  Default: 4
  Equivalent to: [--NotebookExporter.nbformat_version]
```

```
Examples
--------

    The simplest way to use nbconvert is

            > jupyter nbconvert mynotebook.ipynb --to html

            Options include ['asciidoc', 'custom', 'html', 'latex', 'markdown',
'notebook', 'pdf', 'python', 'rst', 'script', 'slides', 'webpdf'].

            > jupyter nbconvert --to latex mynotebook.ipynb

            Both HTML and LaTeX support multiple output templates. LaTeX
includes
            'base', 'article' and 'report'.  HTML includes 'basic', 'lab' and
            'classic'. You can specify the flavor of the format used.

            > jupyter nbconvert --to html --template lab mynotebook.ipynb

            You can also pipe the output to stdout, rather than a file

            > jupyter nbconvert mynotebook.ipynb --stdout

            PDF is generated via latex

            > jupyter nbconvert mynotebook.ipynb --to pdf

            You can get (and serve) a Reveal.js-powered slideshow

            > jupyter nbconvert myslides.ipynb --to slides --post serve

            Multiple notebooks can be given at the command line in a couple of
            different ways:

            > jupyter nbconvert notebook*.ipynb
            > jupyter nbconvert notebook1.ipynb notebook2.ipynb

            or you can specify the notebooks list in a config file, containing::

                c.NbConvertApp.notebooks = ["my_notebook.ipynb"]

            > jupyter nbconvert --config mycfg.py

To see all available configurables, use `--help-all`.
```

[ ]: