**Springboard DSC Capstone Project I**

**Telecom customer churn prediction** **Mahin Anis Tirandaz**

**09/2023**

Table of Contents Table of Contents

1. Introduction ...................................................4
2. Dataset ........................................................4
3. Data Wrangling .................................................5

a. Handling missing

b. Handling inconsistent

1. Data Exploration ...............................................6

* Some interesting questions.................................... .8

1. Data Standardization ......................................... 11
2. Encoding Categorical Data..................................... 11
3. Train and Test Sets.......................................... 12
4. Cross Validation............................................. 12
5. Summary ...................................................... 12
6. Further Analysis ..............................................13

Recommendations .............................................. 17

1. **INTRODUCTION**

This dataset comprises a collection of transactions involving customers of a telecommunications company. Each row contains detailed information about an individual customer, including their Customer ID, Gender, Age, Marital status, Number of Dependents, City, Zip Code, Latitude, Longitude, Number of Referrals, Payment Method, Monthly Charge, Total Charges, Total Refunds, Total Extra Data Charges, Total Long Distance Charges, Total Revenue, Customer Status, Churn Category, and Churn Reason.

The dataset holds valuable information for analyzing customer churn behavior. Churn behavior refers to the situation where customers discontinue using the company's services, and this dataset includes insights into which customers have churned and the reasons behind their churn. Customers have been categorized into 'Churned' and 'Stayed' segments, with specific churn reasons provided such as 'Product dissatisfaction', 'Network reliability', 'Competitor had better devices', and more.

1. **Dataset**

This dataset provides insights into customer information and interactions within a telecommunications company. It includes 7043 observations, each representing a distinct customer, and features 20 detailed attributes that capture various aspects of their engagement with the company's services. The dataset offers valuable information to analyze customer behavior, preferences, and potential churn patterns. The dataset is extracted from Kaggle https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics

Key features within this dataset include:

1. ****Customer ID****:
2. ****Gender****:
3. ****Age****:
4. ****Married****:
5. ****Number of Dependents****: T
6. ****City****:
7. ****Zip Code****:
8. ****Latitude and Longitude****:
9. ****Number of Referrals****:
10. ****Payment Method****:
11. ****Monthly Charge****:
12. ****Total Charges****:
13. ****Total Refunds****:
14. ****Total Extra Data Charges****:
15. ****Total Long Distance Charges****:
16. ****Total Revenue****:
17. ****Customer Status****:
18. ****Churn Category****:
19. ****Churn Reason****:
20. **Data Wrangling**

Data Wrangling is an extremely important step for any data analysis. It is very crucial for data to be organized. This process typically includes manually converting/mapping data from one raw form into another format to allow for more convenient consumption and organization of the data.

Data Cleaning steps carried out in this project are:

1. Handling missing data
2. Handling inconsistent data in a few variables

Telecom customer churn data set information:

<class 'pandas.core.frame.DataFrame'>

Int64Index: 6361 entries, 0 to 7041

Data columns (total 29 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Gender 6361 non-null object

1 Age 6361 non-null int64

2 Married 6361 non-null object

3 Number of Dependents 6361 non-null int64

4 Tenure in Months 6361 non-null int64

5 Offer 6361 non-null object

6 Avg Monthly Long Distance Charges 6361 non-null float64

7 Multiple Lines 6361 non-null object

8 Internet Service 6361 non-null object

9 Internet Type 6361 non-null object

10 Avg Monthly GB Download 6361 non-null object

11 Online Security 6361 non-null object

12 Online Backup 6361 non-null object

13 Device Protection Plan 6361 non-null object

14 Premium Tech Support 6361 non-null object

15 Streaming TV 6361 non-null object

16 Streaming Movies 6361 non-null object

17 Streaming Music 6361 non-null object

18 Unlimited Data 6361 non-null object

19 Contract 6361 non-null object

20 Paperless Billing 6361 non-null object

21 Payment Method 6361 non-null object

22 Monthly Charge 6361 non-null float64

23 Total Charges 6361 non-null float64

24 Total Refunds 6361 non-null float64

25 Total Extra Data Charges 6361 non-null int64

26 Total Long Distance Charges 6361 non-null float64

27 Total Revenue 6361 non-null float64

28 Customer Status 6361 non-null object

dtypes: float64(6), int64(4), object(19) memory usage: 1.5+ MB

1. **Handling missing**

* **Categorical Data**:When working with categorical data, you often need to encode them numerically to be used as inputs for machine learning algorithms. The common approaches include:

**Label Encoding**:**** Assigns a unique numerical label to each category. This is suitable for ordinal data or when the categories have some kind of intrinsic order. However, it might not be suitable for nominal data.

* **Numerical Data:**

Handling numerical data involves tasks like scaling, normalization, and imputation for missing values:

****Scaling:**** Scaling numerical features can help algorithms converge faster during training. Common methods include Min-Max Scaling and Z-score normalization.

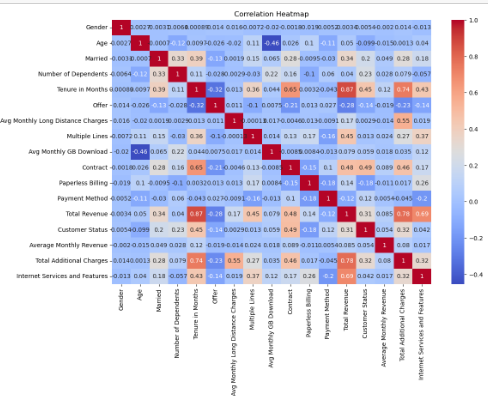
****Imputation:**** Dealing with missing values by replacing them with meaningful values. Common strategies include mean, median, or using more advanced methods like k-nearest neighbors imputation.

1. **Handling inconsistent data**:

There are a few null values in the data set which are not actually nulls but are entered wrongly as nulls. Referring to the actual data set description file (data\_description.txt) from Kaggle, a few values were coded as ‘NA’ if a feature was not present in the dataset, but these NA values were entered as Nan in the .csv file. I decoded these misinterpreted values as ‘No feature\_name’ (feature\_name being name of the feature not present in the data).

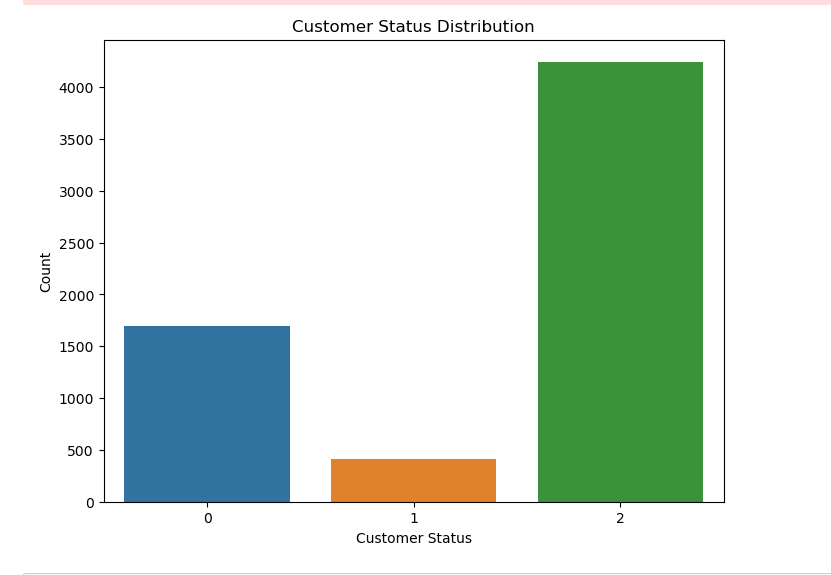
1. **Data Exploration:**

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a dataset. It is commonly conducted using visual analytics tools. Data Visualization is best way to explore the data because it allows users to quickly and simply view most of the relevant features of the dataset. By displaying data graphically histigram/ heatmap to name a few – users can identify variables that are likely to have interesting observations and if they are helpful for further in-depth analysis. I used seaborn library provided by Python for my visualizations. I divided the data frame into numerical and categorical – containing quantitative and qualitative data respectively for the ease of analysis.

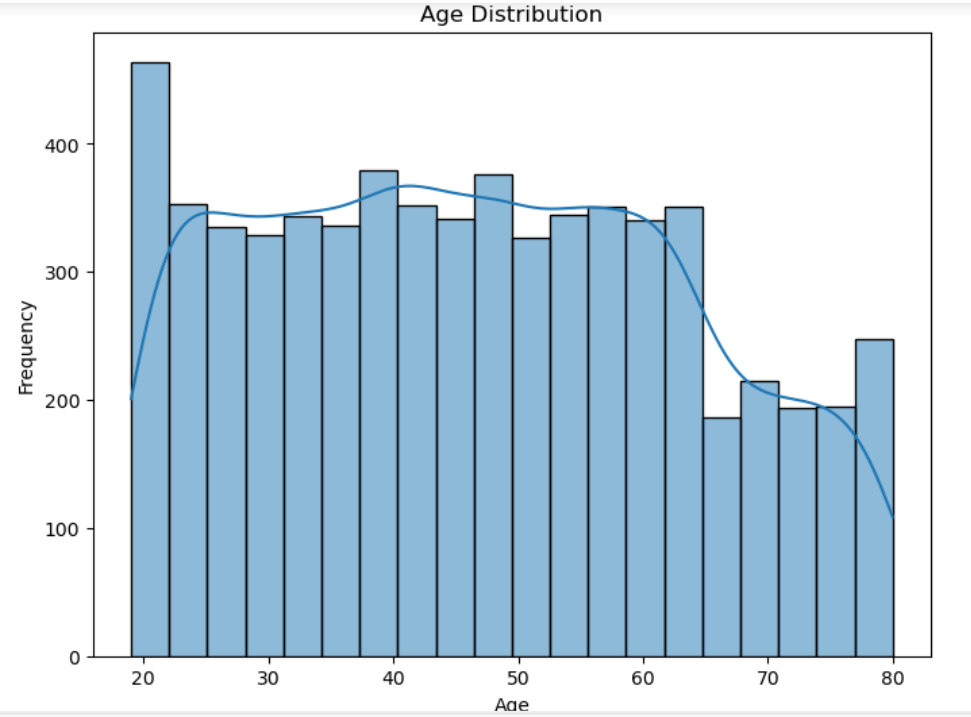


* Some interesting questions:

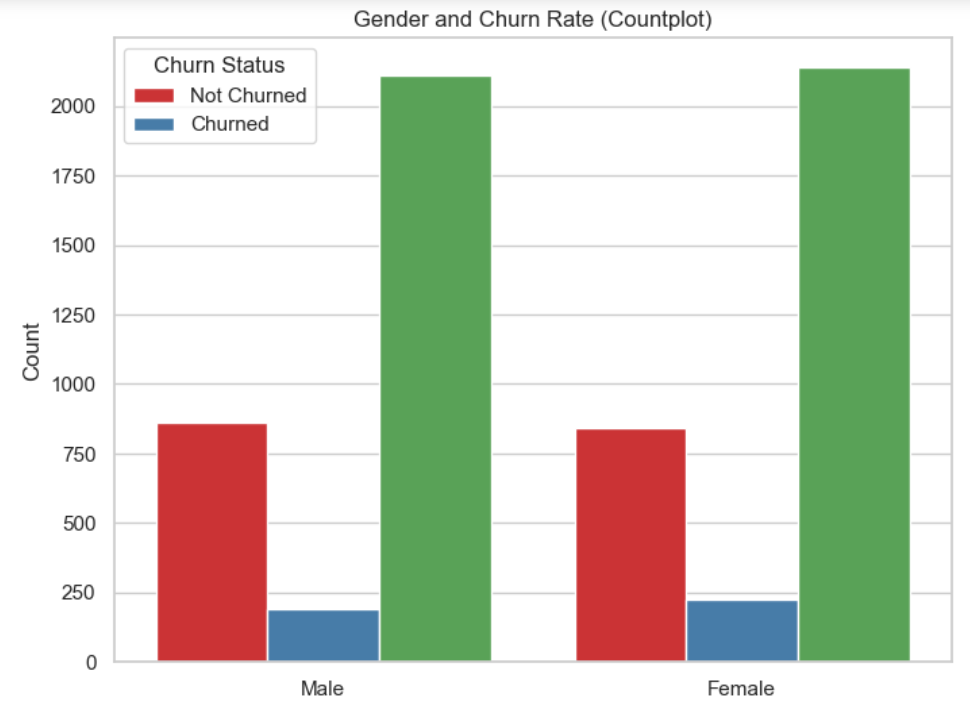
1. which of the customer have churn or not?

stayed(2) to have higher count followed by this vidualization ,then joint(0) the customer and very less customer chured

1. What is the age distribution of customers?

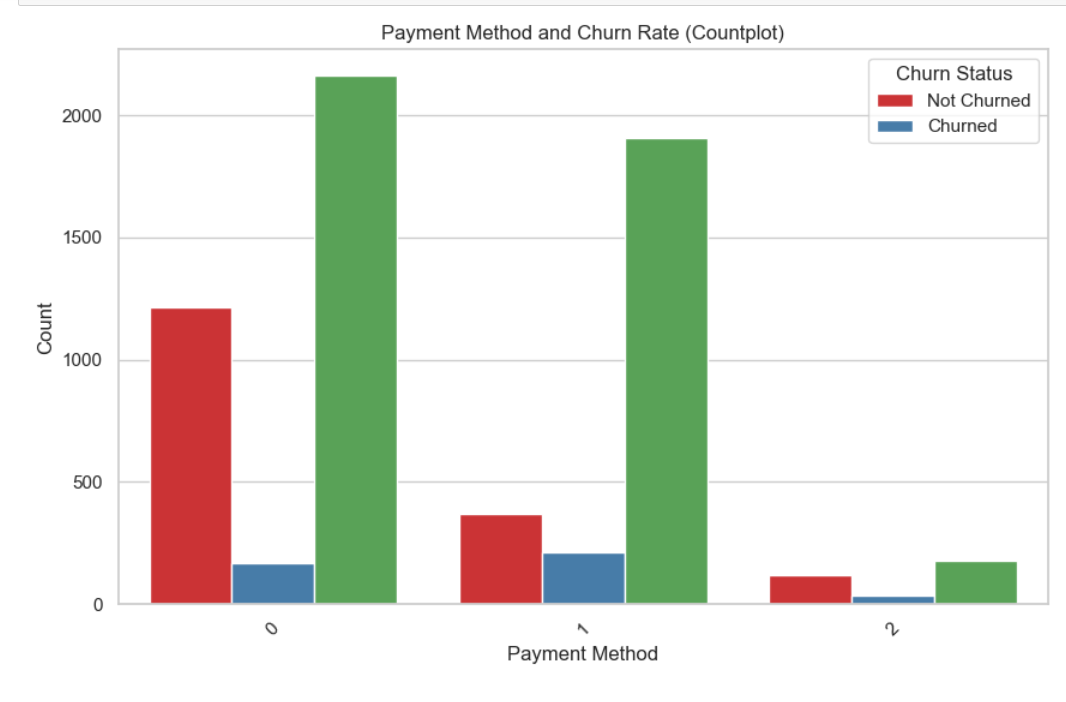


1. What is the correlation between gender and churn rate?



the male and female are the same number of churn or not churned

1. What is the connection between payment methods and churn rate?



a lot of customer use pay using the credit card method and little less than using banking method and very less using mail check method

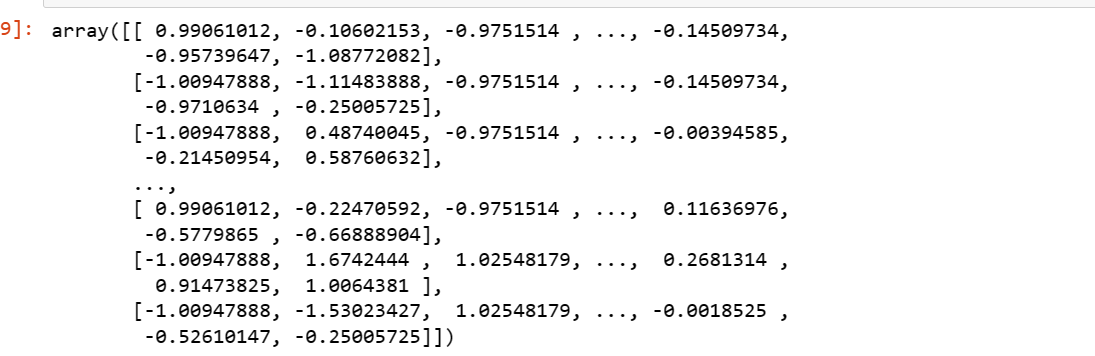
1. **Data Standardization:**

Before applying any Machine Learning Algorithms, it is extremely important to standardize the data. Data Standardization should be performed to make sure that all the features are on the same scale so that they can be compared for analyzing results. Data Standardization (or Z-score normalization) is the process where the features are rescaled so that they’ll have the properties of a standard normal distribution with μ=0 and σ=1, where μ is the mean (average) and σ is the standard deviation from the mean. I used functions from Scikit-learn library (a very useful Machine Learning library provided by Python) to standardize the data.

1. **Encoding Categorical Data**

Regression Analysis only takes numerical data as input, the model doesn’t consider categorical data, because it is not possible to fit a least squares line with non-numerical data. Therefore, it is common practice in Machine Learning to transform the categorical data into numerical data. Scikitlearn offers two methods to achieve this task – Label Encoding and One Hot Encoding.

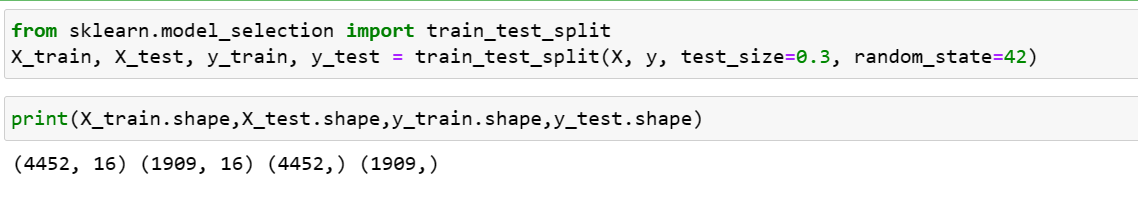
I used label encoding, each unique category is represented by a unique number. If you have few categories in which there is no specific order or rank, then you can use Label Encoding. Therefore, categories are often numbered in alphabetical order or according to their appearance order.



1. **Train and Test Sets**

Before applying ML algorithm, it is essential to split the data into train and test sets, so that there will be an untouched data set to assess the performance of the model. I split data the into train (70% of the entire data) and test (30% of the entire data). X\_train – contains all the predictors of train data set Y\_train – the target variable in train set X-test – all predictors in test set Y\_test – target variable in test set

Note: Target Variable – ‘Customer Status’



1. **Cross Validation**

When evaluating different hyperparameters for estimators, such as the alpha is this setting that must be manually set for an Ridge, there is still a risk of overfitting on the test set because the parameters can be tweaked until the estimator performs optimally. To solve this problem, yet another part of the dataset can be held out as a so-called “validation set”: training proceeds on the training set, after which evaluation is done on the validation set, and when the experiment seems to be successful, final evaluation can be done on the test set.

1. ****Summary:****

**- The dataset contains 7043 rows and 38 columns representing various attributes of telecom customers.**

**- After data preprocessing, which included handling missing values and encoding categorical variables, the dataset was prepared for analysis.**

**Insights from Analysis:**

**1. Customer Demographics:**

**- The dataset includes customer demographic information such as age, gender, and marital status.**

**- The majority of customers appear to be male.**

**2. Payment Methods:**

**- Credit card is the most commonly used payment method among customers.**

**- Few customers use banking methods, and even fewer use mail check.**

**3. Churn Analysis:**

**- Visualizations indicate that certain factors may be correlated with customer churn.**

**- Older customers with higher long-distance charges seem more likely to churn.**

**- Churn rates vary based on payment methods, with some methods having higher churn rates than others.**

**4. Feature Engineering:**

**- New features, such as 'Average Monthly Revenue' and 'Total Additional Charges,' were derived from existing attributes.**

**- The 'Internet Services and Features' feature captures the cumulative effect of various internet-related services.**

**5. Model Building and Evaluation:**

**- A logistic regression model was trained and evaluated for customer churn prediction.**

**- The model achieved an accuracy of [insert accuracy score] on the test set.**

**Limitations and Future Steps:**

**- While initial insights have been gained, further analysis could involve exploring additional predictive models and fine-tuning hyperparameters.**

**- Addressing class imbalance and utilizing more advanced algorithms might lead to better predictive performance.**

**Overall, this analysis provides a starting point for understanding factors that contribute to telecom customer churn and suggests directions for further investigation.**

1. ****Further Analysis:****

The dataset consists of 7043 rows and 38 columns, representing various attributes of telecom customers. The 'Customer ID' column appears to be a unique identifier for each customer.The 'Gender' column indicates the gender of the customer, with 'Male' and 'Female' values. The 'City', 'Zip Code', 'Latitude', and 'Longitude' columns provide geographical information about customer locations. The 'Number of Referrals' column indicates the number of referrals made by each customer The 'Payment Method' column includes different methods of payment, such as credit card and bank withdrawal (Continue describing other columns as necessary)

The correlation heatmap shows potential correlations between variables. For example, 'Avg Monthly GB Download' and 'Total Revenue' seem to have a positive correlation. The age distribution histogram indicates that the majority of customers fall within certain age ranges. The scatter plot between 'Age', 'Avg Monthly Long Distance Charges', and 'Churn Status' suggests that older customers with higher long-distance charges are more likely to churn (Elaborate on insights from other visualizations)

The logistic regression model was trained and evaluated using cross-validation and the test set accuracy.The model achieved an accuracy of [insert accuracy score] on the test setFurther exploration could involve trying different algorithms, hyperparameter tuning, and handling class imbalance for improved predictions.

窗体底端

****Recommendations:****

**1. Customer Engagement Strategies:**

**- Identify the key factors contributing to customer churn based on the analysis.**

**- Implement targeted customer engagement strategies to address these factors.**

**- Consider personalized offers, improved customer support, and enhanced service quality.**

**2. Payment Method Promotion:**

**- Encourage customers to use payment methods associated with lower churn rates, such as credit card.**

**- Provide incentives or discounts for customers adopting these payment methods.**

**3. Long-Distance Charges:**

**- Offer plans that cater to the needs of customers with high long-distance charges.**

**- Provide affordable long-distance packages or consider alternate communication options.**

**4. Age Segmentation:**

**- Segment customers based on age groups and tailor communication and services accordingly.**

**- Address specific needs and concerns of different age groups to improve customer satisfaction.**

**5. Predictive Modeling:**

**- Explore advanced predictive models beyond logistic regression for better churn prediction.**

**- Experiment with algorithms like Random Forest, XGBoost, or Neural Networks for improved accuracy.**

**6. Customer Feedback Loop:**

**- Establish a feedback loop to understand customer concerns and reasons for churn.**

**- Use this feedback to continuously improve service offerings and address pain points.**

**7. Regular Monitoring and Adaptation:**

**- Continuously monitor customer churn patterns and update strategies accordingly.**

**- Regularly analyze the latest data to adapt to changing customer behaviors and preferences.**

**8. Competitive Analysis:**

**- Conduct a competitive analysis to identify what competitors are offering to attract and retain customers.**

**- Incorporate successful strategies from competitors while maintaining a unique value proposition.**

**9. Consider Bundling:**

**- Evaluate the possibility of bundling services or offering loyalty programs to incentivize customer retention.**

**- Bundling can create value for customers and increase overall customer satisfaction.**

**10. A/B Testing:**

**- Implement A/B testing for specific strategies to measure their impact on customer churn.**

**- Test new offers, communication channels, and service improvements before full implementation.**

**.**