Sharp Career
TECHNOLOGIES

Batch: June, 2024 (Data Analytic)

Project B

- AARYA SUHAS PAWAR
  IIT (BHU) VARANASI
- VINAY AGGARWAL
  GROUP 2

1. INTRODUCTION

What is Customer Churn?

Model's purpose is to bring churned customers to light. In a targeted approach industry try to identify which customers are likely to churn. Churn Customer refers to the number of existing customers who may leave the service provider over a given period.

Terms related to telecom company -

There are promotional costs known as acquisition costs and retention costs in a telco company. The acquisition cost is the price a company pays to gain new consumers. Retention costs, on the other hand, are the costs of keeping existing clients. It is very difficult to predict which customers would churn and which customers will be maintained due to human limitations. As a result, the allocation of money may be incorrect, resulting in a higher amount of cash being issued. according to some reports, the acquisition cost is 5 times that of the retention cost.

The customer churn is the direct loss in terms of revenue to the company. Appropriate steps can be taken, and a better service can be provided to such customers.

2. DESCRIPTION OF THE DATASET

This dataset has 7043 customer entries with the following columns 'customerID', 'gender','SeniorCitizen', 'Partner', 'Dependents','tenure', 'PhoneService', 'MultipleLines', 'InternetService','OnlineSecurity', 'OnlineBackup', 'DeviceProtection','TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges','TotalCharges', 'Churn'.

Content

Each row represents a customer, each column contains customer's attributes described on the column Metadata.

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Month to month contracts, absence of online security and tech support seem to be positively correlated with churn. While, tenure, two year contracts seem to be negatively correlated with churn.

Interestingly, services such as Online security, streaming TV, online backup, tech support, etc. without internet connection seem to be negatively related to churn.

Data Cleaning:

- **Handling Missing Values**: Filling in missing data or removing rows/columns with missing values.

Data Transformation:

- **Normalization/Standardization**: Scaling numerical features to a standard range.
- **Encoding Categorical Variables**: Converting categorical data into numerical format (e.g., one-hot encoding, label encoding).

Data Splitting:

- Dividing the data into training, validation, and test sets.

3. EXPLANATORY ANALYSIS

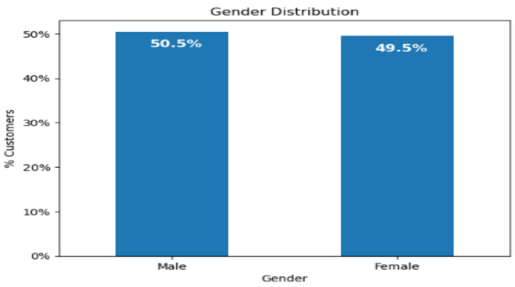
- Distribution of individual variables and then slice and dice our data for any interesting trends.

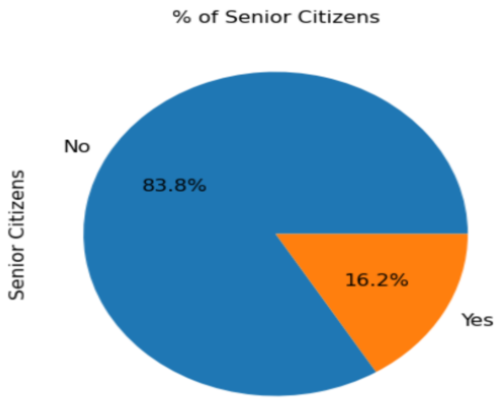
Demographics - Let us first understand the gender, age range, partner and dependent status of the customers

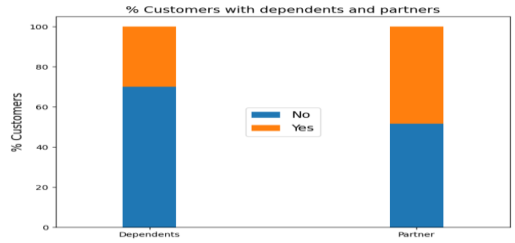
A.) Gender Distribution - About half of the customers in our data set are male while the other half are female.

B.) % Senior Citizens - There are only 16% of the customers who are senior citizens. Thus most of our customers in the data are younger people.

C.) Partner and dependent status - About 50% of the customers have a partner, while only 30% of the total customers have dependents.

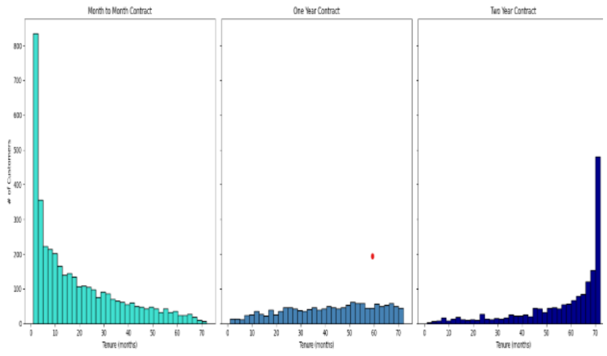






- **Customer Account Information**: Let us now look at the tenure, contract¶

1. **Tenure:** After looking at the below histogram we can see that a lot of customers have been with the telecom company for just a month, while quite a many are there for about 72 months. This could be potentially because different customers have different contracts. Thus based on the contract they are into it could be more/less easier for the customers to stay/leave the telecom company.
2. **Contracts:** To understand the above graph, lets first look at the # of customers by different contracts.

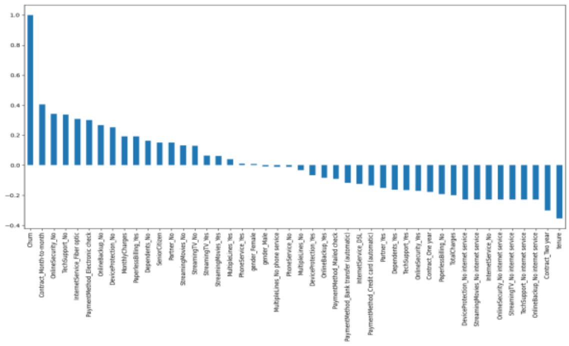


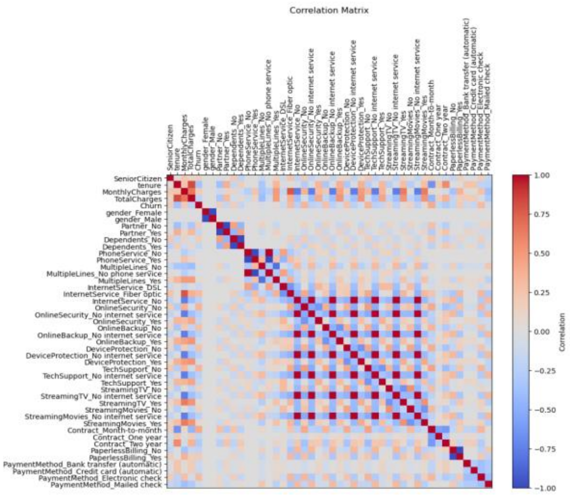
3. **Partner and dependent status** - About 50% of the customers have a partner, while only 30% of the total customers have dependents.

CORRELATION MATRIX

Correlation Analysis: Conducted correlation analysis to identify relationships between individual attributes.

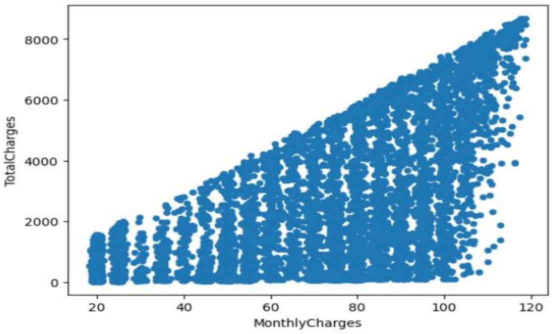
Correlation coefficients provide insights guiding the selection of potentially influential features.





- Now let's take a quick look at the relation between monthly and total charges¶

The total charges increases as the monthly bill for a customer increases.



- Finally, let's take a look at out predictor variable (Churn) and understand its interaction with other important variables as was found out in the correlation plot.¶

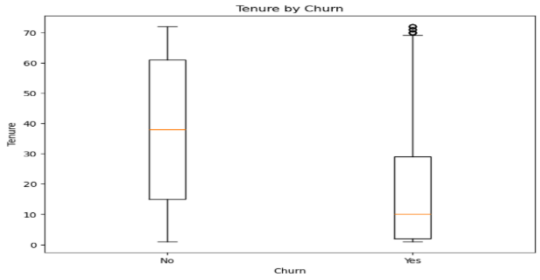
1. Lets first look at the churn rate in our data

   In our data, 74% of the customers do not churn. Clearly the data is skewed as we would expect a large majority of the customers to not churn. This is important to keep in mind for our modelling as skewness could lead to a lot of false negatives. We will see in the modelling section on how to avoid skewness in the data.

2. Churn rate by tenure, seniority, contract type, monthly charges and total charges to see how it varies by these variables.

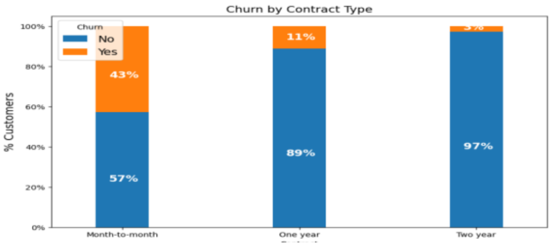
i. Churn vs Tenure**: As we can see form the below plot, the customers who do not churn, they tend to stay for a longer tenure with the telecom company.**

ii.

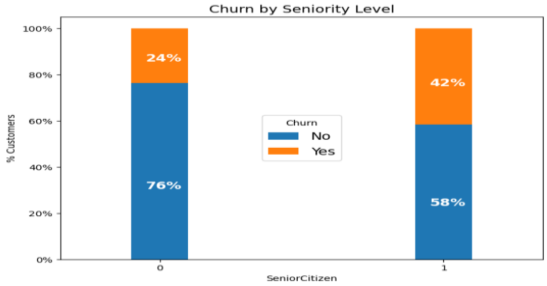


n

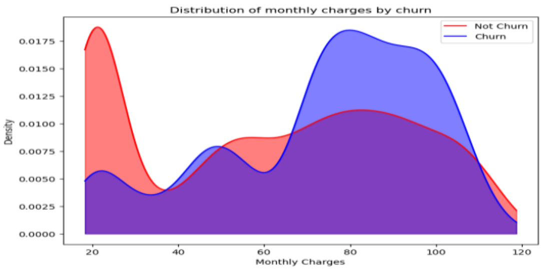
by Contract Type**: Similar to what we saw in the correlation plot, the customers who have a month to month contract have a very high churn rate.**



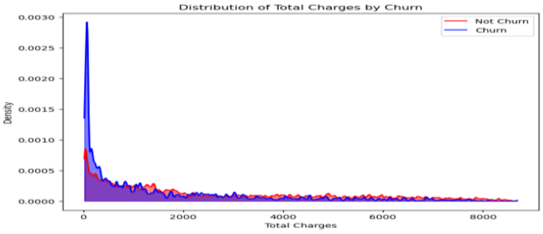
iii. Churn by Seniority**: Senior Citizens have almost double the churn rate than younger population.**



iv. Churn by Monthly Charges**: Higher % of customers churn when the monthly charges are high.**



v. Churn by Total Charges**: It seems that there is higher churn when the total charges are lower.**



Data Splitting: Training-Validation-Testing Split: Split the dataset into training, validation, and testing sets for model training, tuning, and evaluation respectively. This ensures unbiased model evaluation and helps in generalization to unseen data.

SOME PREDICTIVE MODELS AND THEIR COMPARISONS

1. Logistic Regression

Logistic regression is a statistical method used for binary classification problems. It is used to model the probability of a certain class or event, such as pass/fail, win/lose, healthy/sick, etc. Logistic regression is a type of regression analysis where the dependent variable is binary or dichotomous, that is, it has only two possible outcomes.

It is important to scale the variables in logistic regression so that all of them are within a range of 0 to 1. This helped me improve the accuracy from 79.7% to 80.7%. Further, you will notice below that the importance of variables is also aligned with what we are seeing in Random Forest algorithm and the EDA we conducted above.

Observations

We can see that some variables have a negative relation to our predicted variable (Churn), while some have positive relation. Negative relation means that likeliness of churn decreases with that variable. Let us summarize some of the interesting features below:

- As we saw in our EDA, having a 2 month contract reduces chances of churn. 2 month contract along with tenure have the most negative relation with Churn as predicted by logistic regressions
- Having DSL internet service also reduces the probability of Churn
- Lastly, total charges, monthly contracts, fibre optic internet services and seniority can lead to higher churn rates. This is interesting because although fibre optic services are faster, customers are likely to churn because of it. I think we need to explore more to better understand why this is happening.

2. Random Forest

A Random Forest is a powerful ensemble learning algorithm used for both classification and regression tasks in machine learning. It operates by constructing multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Observations:

- From random forest algorithm, monthly contract, tenure and total charges are the most important predictor variables to predict churn.
- The results from random forest are very similar to that of the logistic regression and in line to what we had expected from our EDA

CONCLUSION

Logistic Regression got us an accuracy of 80.7%

Using Random Forest the accuracy came out to be 80.88%

With ADA Boost accuracy came out to be 81.2%

Final Accuracy of 81.4% was obtained

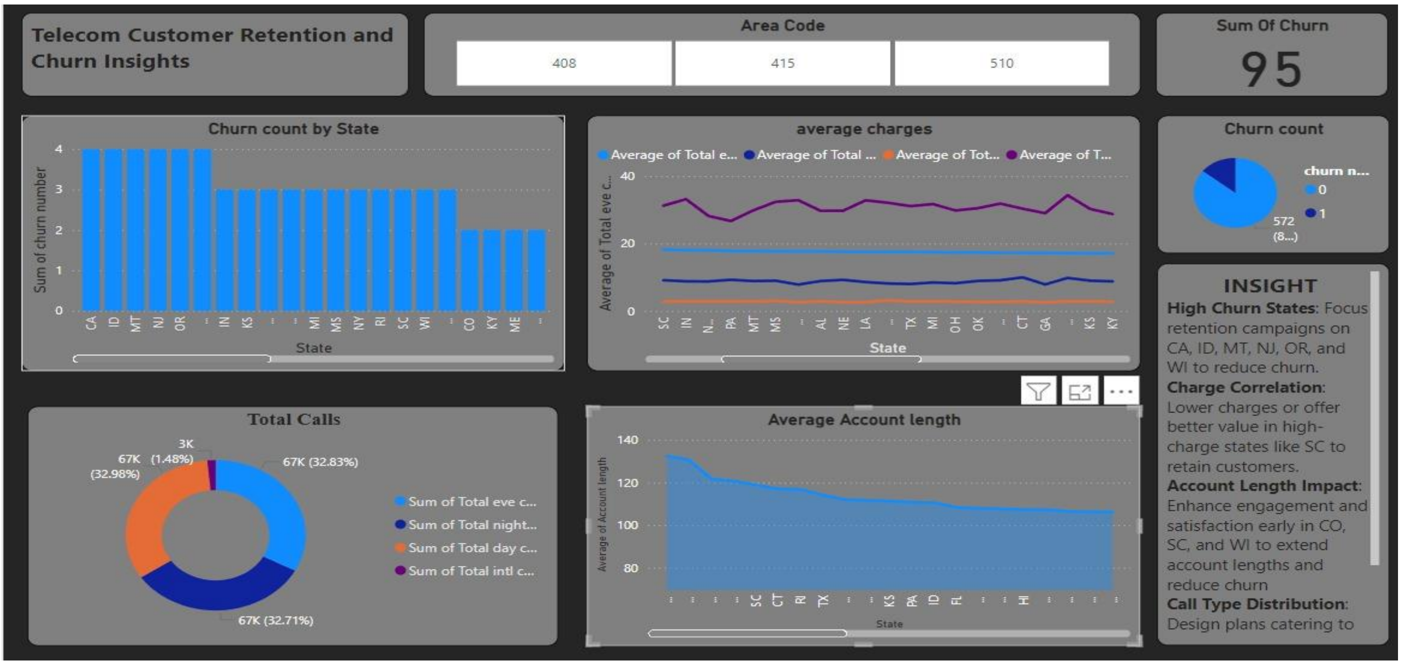
POWER BI DASHBOARD

Power BI dashboard provides a comprehensive overview of telecom customer retention and churn patterns, focusing on key metrics such as churn count by state, average charges, total calls, and average account length. The total number of churned customers is displayed prominently, with a sum of 95.

One of the key insights is the "Churn Count by State," which highlights states with the highest churn, indicating a need for targeted retention strategies in states like CA, ID, MT, NJ, OR, and WI. The "Average Charges" section shows the average charges for different call types (evening, day, night, international) across states, helping to identify effective pricing strategies. The "Total Calls" pie chart displays the distribution of call types, providing an understanding of customer usage patterns. The "Average Account Length" graph visualizes the average length of customer accounts, with shorter lengths in certain states suggesting areas for improvement in customer satisfaction and engagement.

The dashboard offers actionable insights and recommendations, such as focusing retention campaigns on high-churn states and adjusting charges to offer better value in these areas. Enhancing early engagement and satisfaction in states with shorter average account lengths can help extend customer lifespans. Additionally, understanding call type distribution can aid in designing plans that better cater to customer preferences.

Filters for area codes (408, 415, 510) allow users to drill down into specific regions for detailed analysis. Interactive elements, like clickable charts, enable dynamic data exploration, providing deeper insights. The dashboard's clear and intuitive layout, combined with effective visual elements, ensures ease of navigation and a quick grasp of key metrics.



INSIGHT

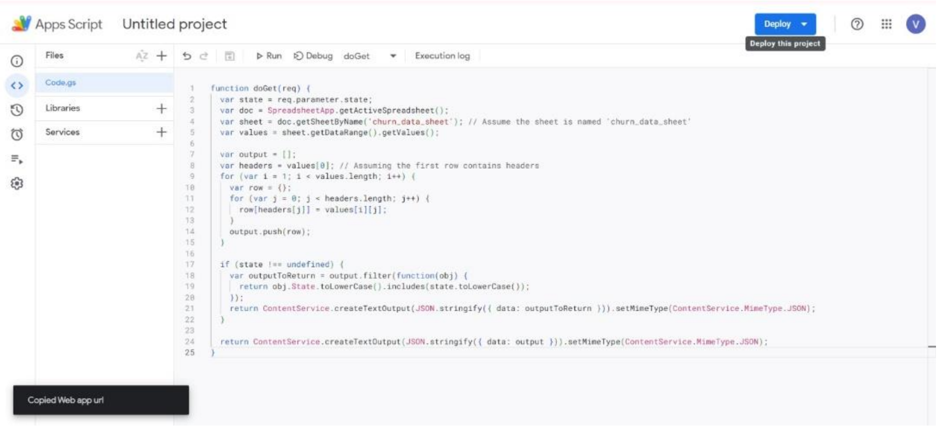
High Churn States: Focus retention campaigns on CA, ID, MT, NJ, OR, and WI to reduce churn.
Charge Correlation: Lower charges or offer better value in high-charge states like SC to retain customers.
Account Length Impact: Enhance engagement and satisfaction early in CO, SC, and WI to extend account lengths and reduce churn
Call Type Distribution: Design plans catering to

API

The API created using Google Apps Script retrieves and filters telecom churn data from a Google Sheet named 'churn_data_sheet'. It accepts a state parameter to filter the data for a specific state or returns all data if no parameter is provided.

The script accesses the active Google Sheet, retrieves data, and processes it, assuming the first row contains headers. It creates an array of objects for each row of data. If a state parameter is specified, the script filters the data to include only rows matching the state. The processed data is then returned as a JSON response.

Key fields in the JSON output include State, Account length, Area code, International plan, Voice mail plan, Number vmail messages, Total day minutes, Total day calls, Total day charge, Total eve minutes, Total eve calls, Total eve charge, Total night minutes, Total night calls, Total night charge, Total intl minutes, Total intl calls, Total intl charge, Customer service calls, and Churn.

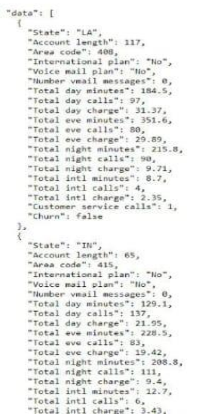


```
function doGet(req) {
  var state = req.parameter.state;
  var doc = SpreadsheetApp.getActiveSpreadsheet();
  var sheet = doc.getSheetByName('churn_data_sheet'); // Assume the sheet is named 'churn_data_sheet'
  var values = sheet.getDataRange().getValues();

  var output = [];
  var headers = values[0]; // Assuming the first row contains headers
  for (var i = 1; i < values.length; i++) {
    var row = {};
    for (var j = 0; j < headers.length; j++) {
      row[headers[j]] = values[i][j];
    }
    output.push(row);
  }

  if (state !== undefined) {
    var outputToReturn = output.filter(function(obj) {
      return obj.State.toLowerCase().includes(state.toLowerCase());
    });
    return ContentService.createTextOutput(JSON.stringify({ data: outputToReturn })).setMimeType(ContentService.MimeType.JSON);
  }

  return ContentService.createTextOutput(JSON.stringify({ data: output })).setMimeType(ContentService.MimeType.JSON);
}
```



```
"data": [
  {
    "State": "LA",
    "Account length": 117,
    "Area code": 408,
    "International plan": "No",
    "Voice mail plan": "No",
    "Number vmail messages": 0,
    "Total day minutes": 184.5,
    "Total day calls": 97,
    "Total day charge": 31.37,
    "Total eve minutes": 351.6,
    "Total eve calls": 80,
    "Total eve charge": 29.89,
    "Total night minutes": 215.8,
    "Total night calls": 90,
    "Total night charge": 9.71,
    "Total intl minutes": 8.7,
    "Total intl calls": 4,
    "Total intl charge": 2.35,
    "Customer service calls": 1,
    "Churn": false
  },
  {
    "State": "IN",
    "Account length": 65,
    "Area code": 415,
    "International plan": "No",
    "Voice mail plan": "No",
    "Number vmail messages": 0,
    "Total day minutes": 129.1,
    "Total day calls": 137,
    "Total day charge": 21.95,
    "Total eve minutes": 228.5,
    "Total eve calls": 83,
    "Total eve charge": 19.42,
    "Total night minutes": 208.8,
    "Total night calls": 111,
    "Total night charge": 9.4,
    "Total intl minutes": 12.7,
    "Total intl calls": 6,
    "Total intl charge": 3.43,
```

THANK YOU!!