Link to Data Bricks notebook:

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3147479163590333/331398138458776/6138009851364668/latest.html

Link to dataset:

https://www.kaggle.com/blastchar/telco-customer-churn

CUSTOMER CHURN ANALYSIS

CSC-696 HIGH-PERFORMANCE COMPUTING

Sai Prasanthi Parasu

Shaifali Chandreshkumar Patel

INTRODUCTION

Customer churn refers to a customer's propensity to abandon a brand and cease being a paying client of a specific business. A customer churn (attrition) rate is the percentage of customers who stop using a company's products or services during a particular time. One method for calculating churn is to divide the number of customers lost during a given time interval by the number of customers acquired. Then, multiply that number by 100 percent. For example, if you gained 150 customers and lost three in the previous month, your monthly churn rate is 2%.

In this project, we analyze the churn of the telecom company in various aspects, building a Predictive model to predict the churn and evaluate the model.

DATA SET USED

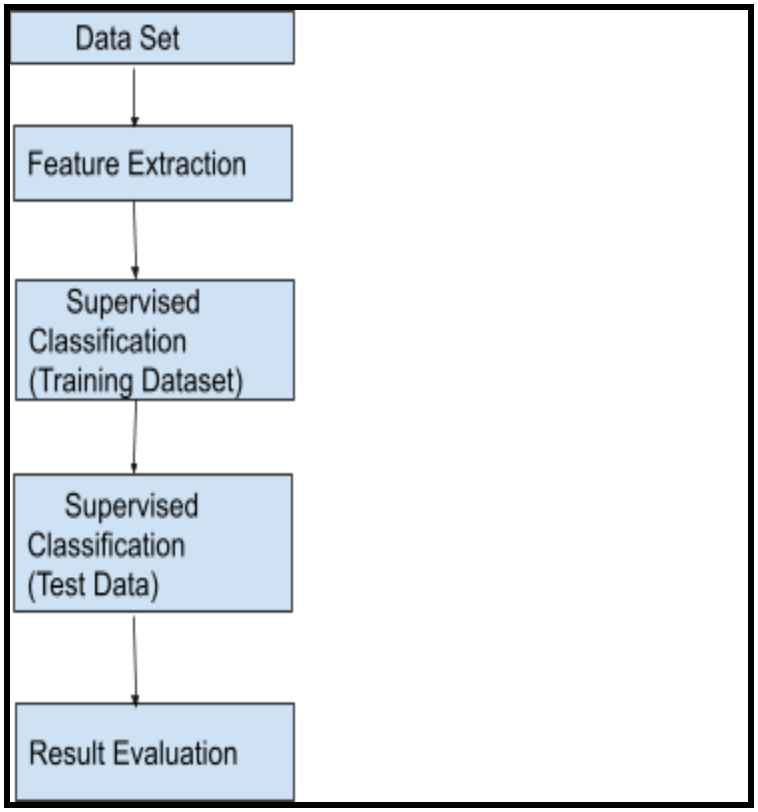
In this project, the data set selected was Telecom company. Each row represents a customer. Each column contains the customer's attributes described on the column Metadata such as Phone service, Internet service, and many more.

We have 7043 rows (customers) and 21 columns (features) in this data.

TASKS ADDRESSED

1. Analyzed the data sets in various aspects and found the relation between them.
2. Calculated the correlation of churn with the other variables.
3. Completed Exploratory data analysis on the dataset.
4. Build a Predictive model.
5. Evaluated the model using evaluation metrics.

FLOW CHART OF THE PROJECT:



TOOLS USED:

- Databricks
- Apache spark
- Pandas for Visualization.
- Matplot library for Visualizing the data
- Seaborn for Visualization.

METHODOLOGICAL APPROACH

1. Data Understanding

In this module, we studied the data and understood the data's attributes. We have 7043

rows (customers) and 21 columns (features) in this data. There are no missing and incorrect values in the data, so there is no need to clean the data.

In the dataset, the column 'TotalCharges' was of String type; since it should be of the numerical column, we changed the datatype of 'TotalCharges' to the integer type.

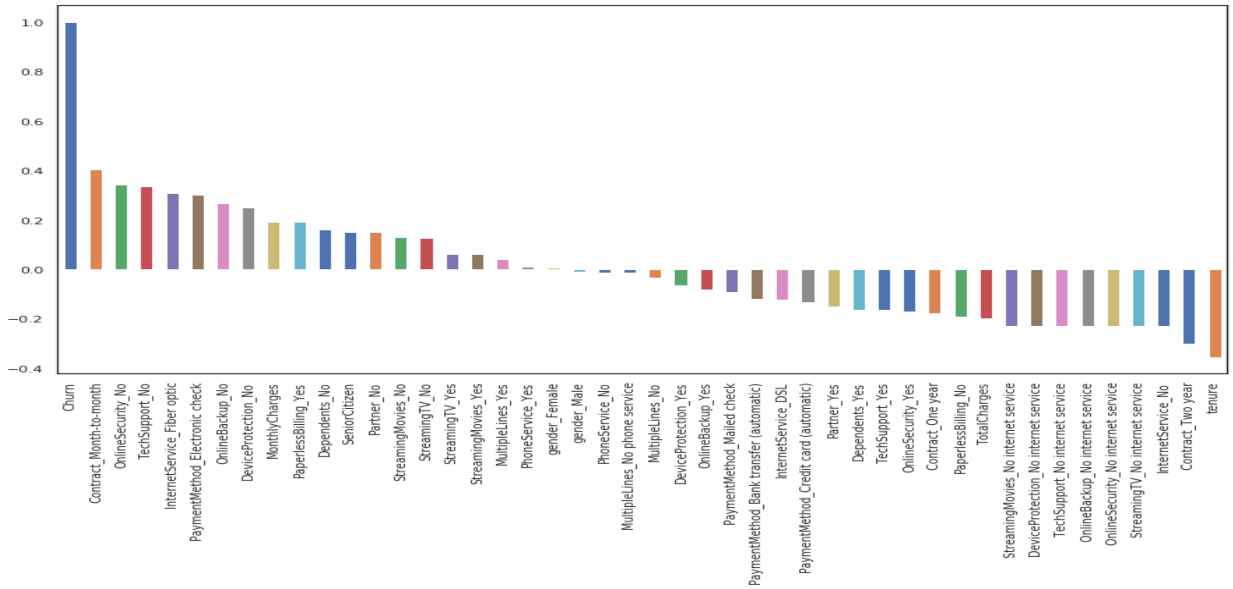
2. Data Cleaning

As we discussed earlier, since there are no missing and incorrect values in the dataset, there is no need to clean the data set.

3. Exploratory Data Analysis

1.Correlation

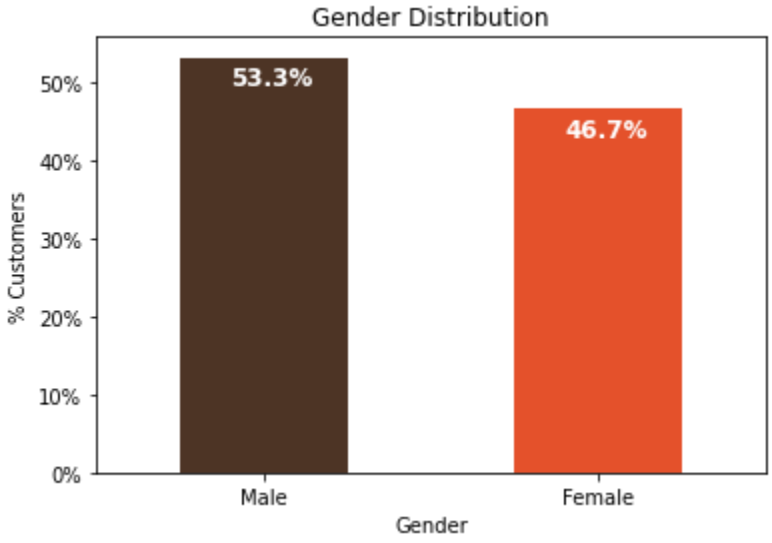
We found out the correlation of churn with the other variables.



Month-to-month contracts, absence of online security, and tech support appear positively related to churn. While tenure, two-year agreements appear to be negatively correlated with churn. Surprisingly, services such as online security, streaming TV, online backup, tech support, and so on that do not require an internet connection appear to be negatively related to churn.

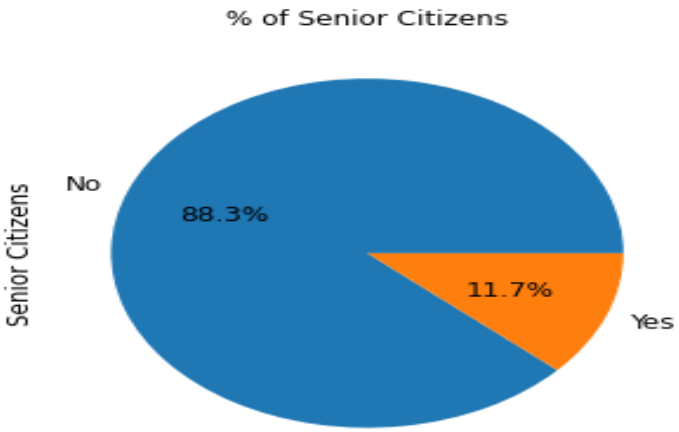
2. Gender Distribution:

In this we can see male customers are more than female customers.



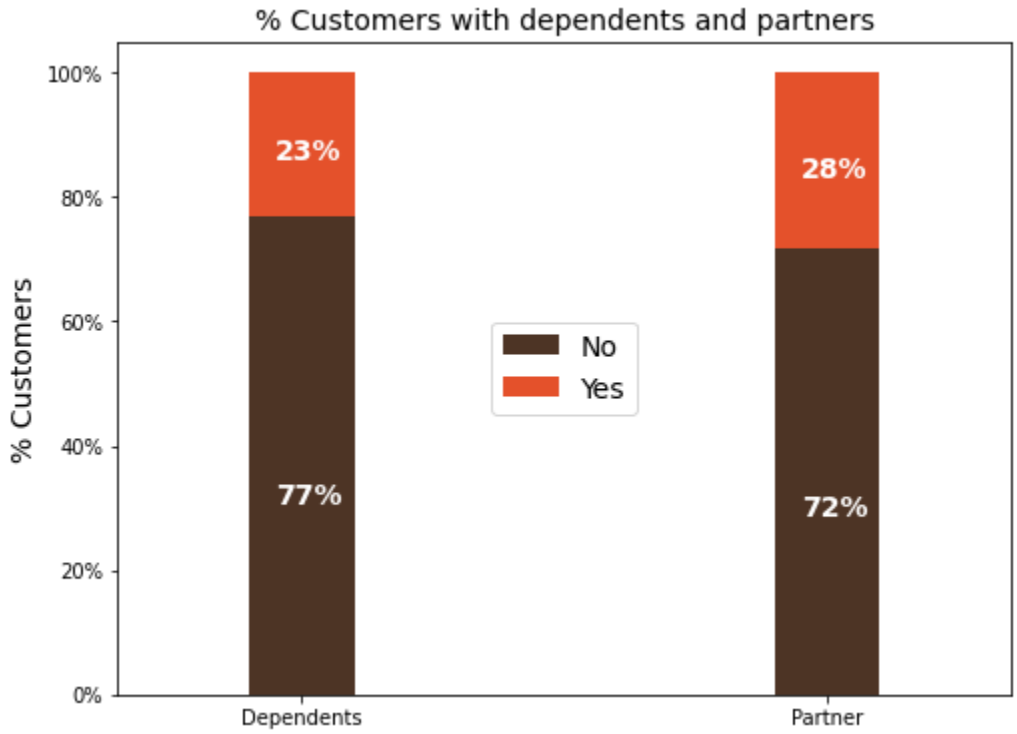
3. Senior Citizens:

In this data, the percentage of senior citizens is less. We can conclude that most senior citizens are not using any mobile services.

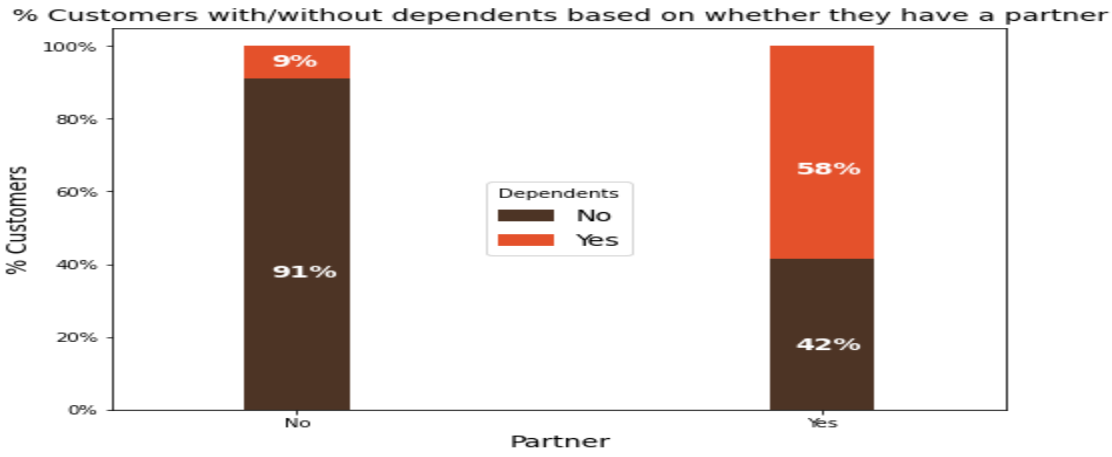


4. Dependent and Partner Status:

From the graph we can see that only less percent of customers have partners and dependents, most of them don't have partners or dependents.



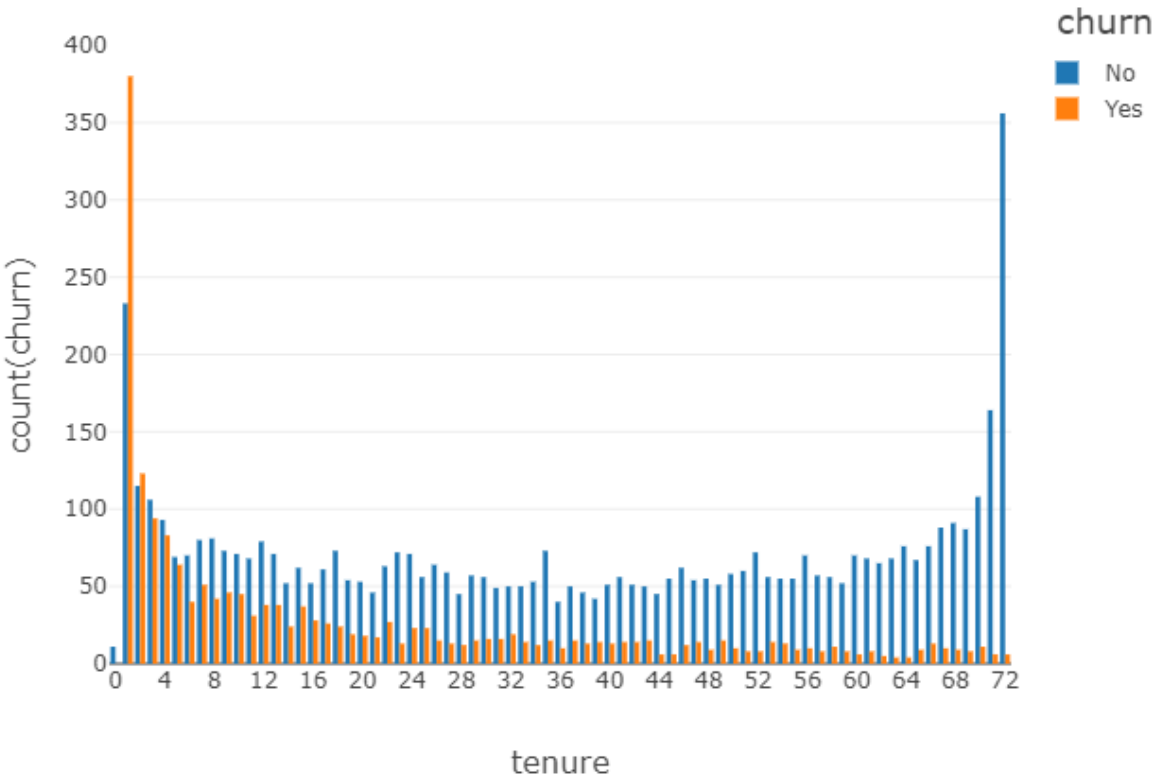
It would be interesting to look at the % of customers who have partners and dependents.



Interestingly, only about half of the partners' customers have a dependent, while the other half are independents. Additionally, as expected, most of them do not have any dependents among the customers who do not have any partners.

5. Relation between the Tenure and Churn:

From this graph, we can see that customers whose tenure is less are more likely to be churned out than people who have been using this service for many months.

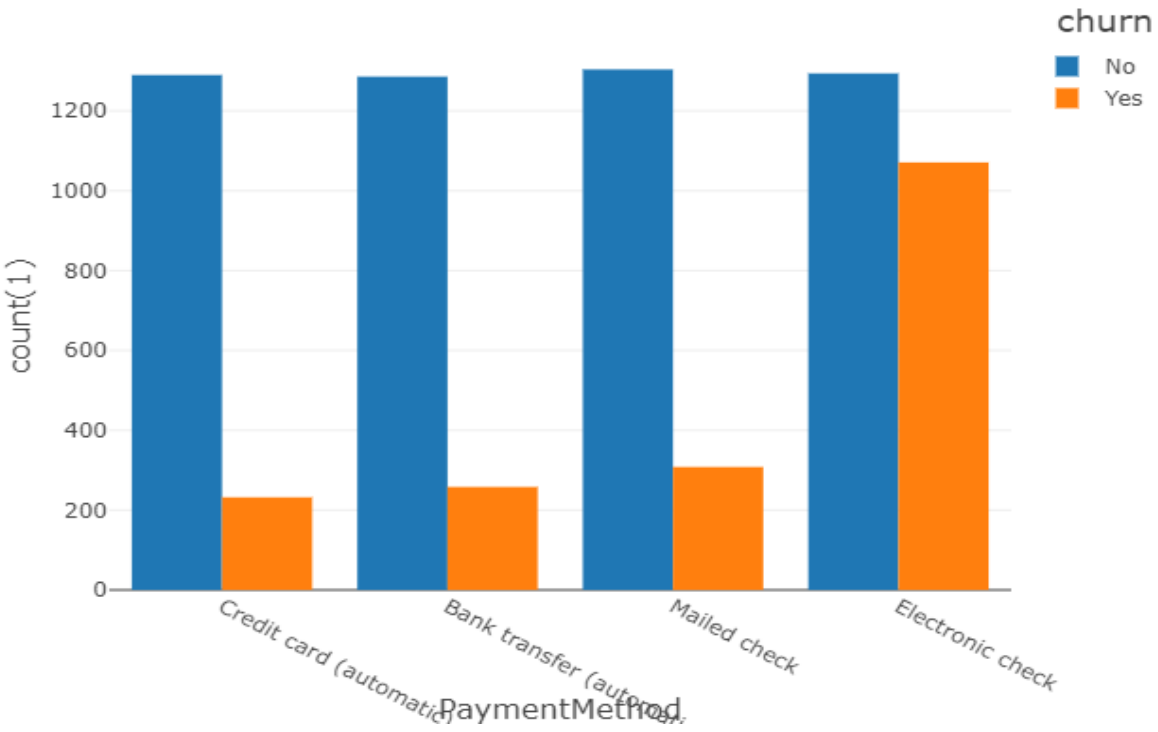


6. Internet Service:

SeniorCitizen_InternetService	DSL	Fiber optic	No
1	259	831	52
0	2162	2265	1474

From this, we can say that most senior and non-senior citizens prefer fiber optic internet service, and the next preference is DSL. Fewer customers don't use Internet service.

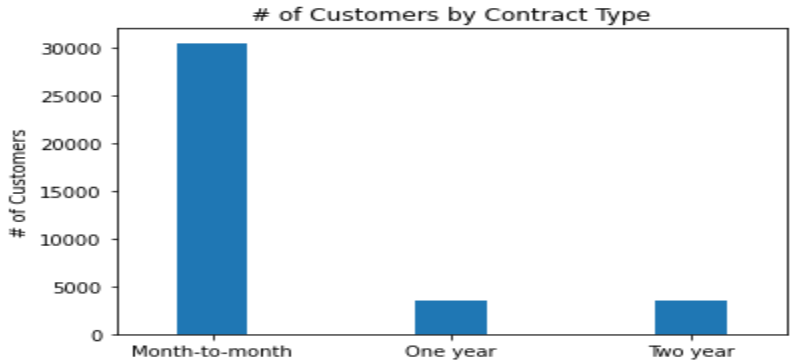
7. Payment Method:



From this, we can say that most of the customers' payments are made by Electronic check, and they are the one who is churning out the most. The remaining payment methods are less than electronic checks, and the churning rate was more petite.

8. Contract:

As we can see from this graph most of the customers are in the month to month contract. At the same time, there are an equal number of customers in the one year and 2-year contracts.

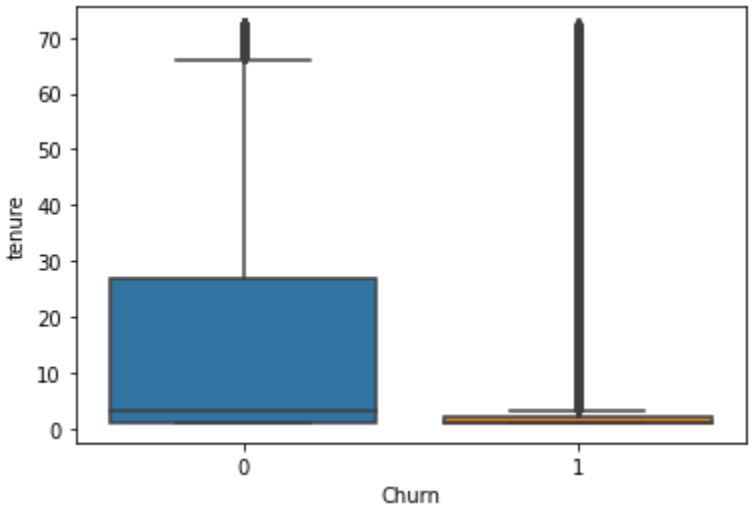


Predictor variable (Churn) and understand its interaction with other important variables as was found out in the correlation plot.

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

1. **Churn vs Tenure**

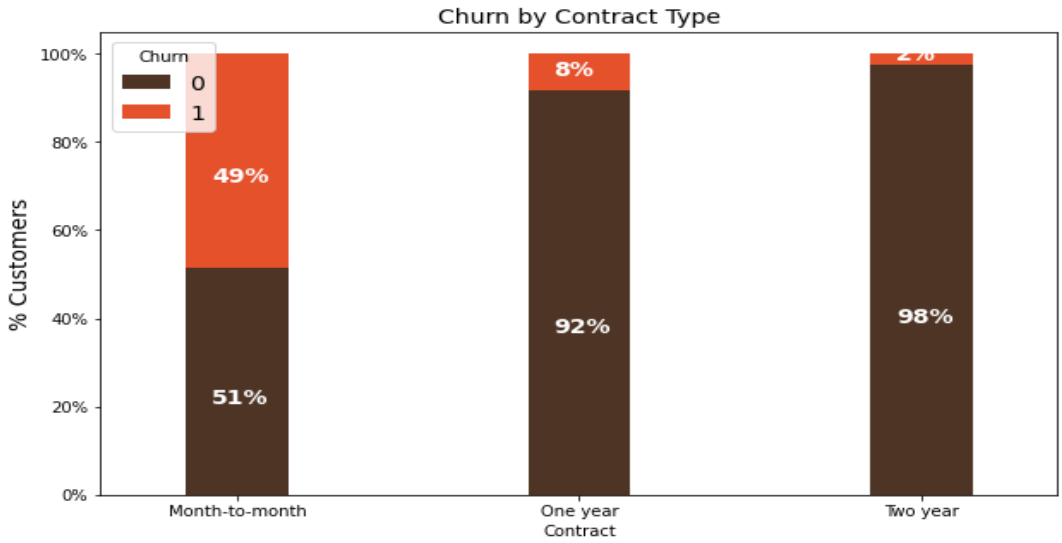
   As we can see from the below plot, the customers who do not churn tend to stay for longer tenure with the telecom company.



2. **Churn vs Contract:**

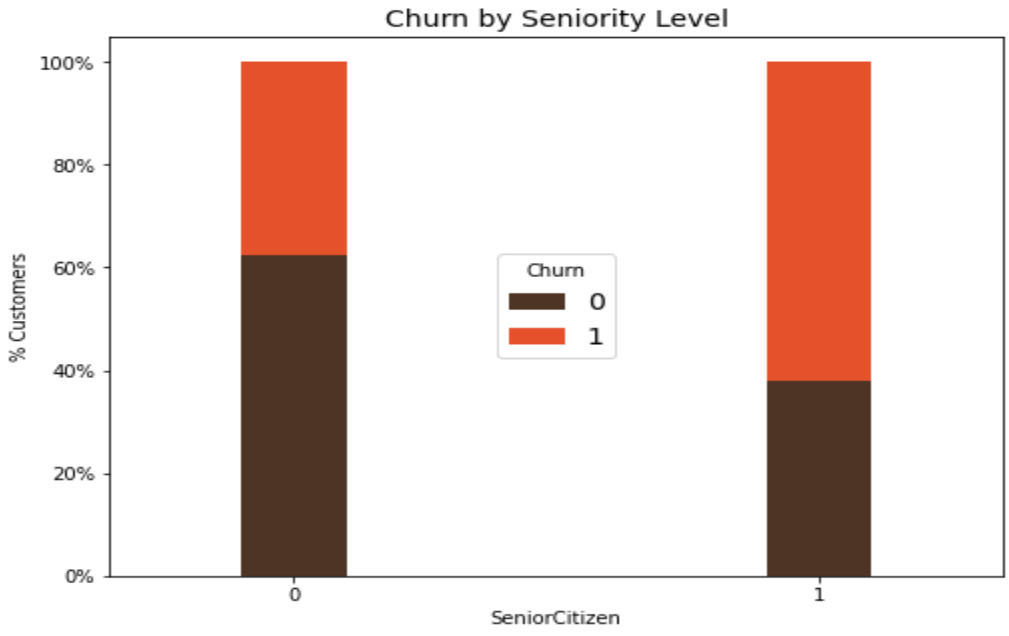
   The customers who have a monthly subscription have a high churn rate compared

to other contracts.



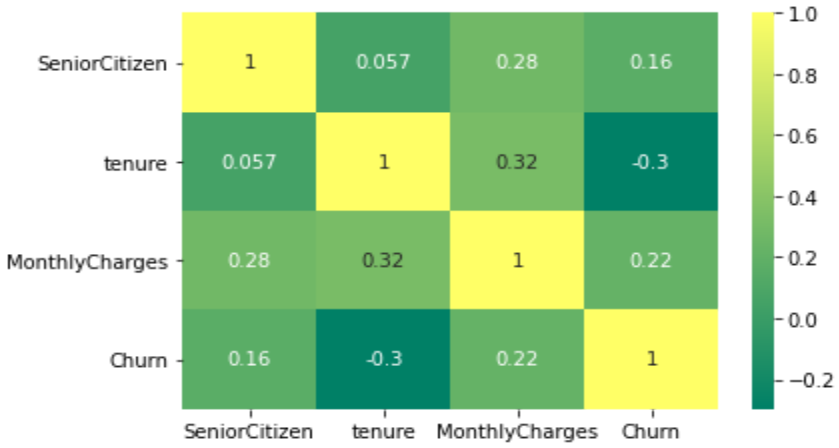
3. Churn vs Seniority:

Senior citizens have almost double the chu



rn rate compared to non-senior citizens.

4. **Correlation among different columns in the dataset.**



4. MODELING:

Data Preprocessing:

I have assigned the data frame to test_data and train_data in 70 and 30 ratios for model building. After that, I imported the pipeline and imported one hot encoder, string Indexer (converting string to numeric), vector assembler from the transformer, and converted categorical columns to numeric columns. It was then encoded the columns using OneHotEncoder and added them to stages. Using string indexer converting the churn data into boolean type.

Feature Engineering:

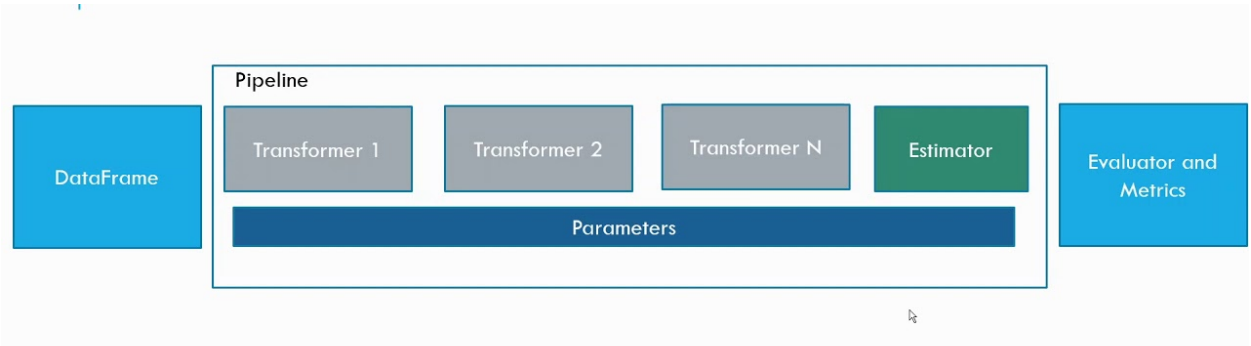
Now coming to the analysis part, the TotalCharges and MonthlyCharges have 55% correlation between them. From the above EDA, we know that customers whose tenure is more diminutive and have a high churn rate. Since this column looks interesting, we can perform feature engineering.

By using quantile discretizer. We divided the tenure into three parts, assigned it to tenure_bin, and added it to the stages. After that, we convert every column to a numeric column, add each column together, pass them into a vector assembler, assign the output to features, and append the result to stages.

By this, we were done by data pre-processing and feature engineering.

Building the model:

Pipeline:
Spark ML Pipeline Flow:



The pipeline model was trained by train_data, and then we transferred the train_data and test_data by pipeline model and assigned them to two separate data frames, testdf and traindf, respectively.

Now we have the data prepared in every aspect, did EDA, did Descriptive statistics, created pipeline, created transformers, now we can build our model.

Logistic Regression:

Logistic regression is a Machine Learning algorithm that is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

Created the LR model and trained the model with the train_data. We evaluate the model even though it is an imbalance data because we are assessing only train_data; if we calculate accuracy, precision, recall is around 75%, and Area around Roc is around 80%.

Here, we transformed the test_data, generated the predictions, and assigned them to the data frame. If we pass the prediction data frame to evaluate using binary classification Evaluation and develop roc, we got ROC around 71%, 10% less than ROC of imbalance date.

Churn Prediction Using Logistic Regression:

Like Logistic Regression, we need to train the rf model with train data and generate the predictions using test data.

After generating the predictions and evaluating them, we got,

Now we have predictions dataframe, so we can generate the predictions of churn and evaluate them. After generating the predictions we can evaluate them. First we will generate false positive, true positive, false negative, true negative. Then calculate the accuracy, precision and recall.

accuracy = (tp+tn)/count

precision = tp/(tp+fp)

recall = tp/(tp+fn)

After evaluating them we got,

```
Accuracy: 0.7444404910158334
Precision: 0.7354675652906487
Recall: 0.5756043956043956
```

Even though accuracy and precision was good, recall was not upto mark.

Random Forest:

Random forest is a supervised machine learning algorithm used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. It performs better results for classification problems.

Churn Prediction Using Random Forest.

Like Logistic Regression, we need to train the rf model with train data and generate the predictions using test data.
And then evaluating them we got

```
Accuracy: 0.7504892367906066
Precision: 0.7344799785004031
Recall: 0.6006593406593407
```

Here accuracy and precision was less compared to the Logistic regression model but recall was quite good.
After generating the predictions and evaluating them, we got,

Optimization of Random Forest Using Grid Search:

```
Accuracy: 0.811332503113325
Precision: 0.8262154176739189
Recall: 0.6760439560439561
```

After using Grid Search for optimization Accuracy, precision, and recall was increased

Grid Search:

Grid search is a tuning technique that attempts to compute the optimum values of hyperparameters. It is an exhaustive search that is performed on the specific parameter values of a model. The model is also known as an estimator.

Grid search is thus considered a very traditional hyperparameter optimization method since we are basically "brute-forcing" all possible combinations. The models are then evaluated through cross-validation. The model boasting the best accuracy is naturally considered to be the best.

Optimization of Logistic Regression Using Grid Search:

Grid Search ensures an exhaustive grid search that breeds candidates from a grid of parameter values. As we shall see later on, these values are instanced using the parameter param_grid.

We then specify the hyperparameters we seek to examine. When using param_grid, the three hyperparameters to use are regParam, elasticNetParam and maxiter. We can give each one several values to choose from.

Remember that it is possible to change these values and test them to see which values' collection gives better results. Below are my randomly chosen values.

Evaluation:

We mentioned that cross-validation is carried out to estimate the performance of a model. In k-fold cross-validation, k is the number of folds.We use cross-validation to train the model 5 times. This means that 5 would be the k value.

```
Accuracy: 0.7448852517345668
Precision: 0.7327061427780852
Recall: 0.581978021978022
```

Here we can see that accuracy and precision are better compared to other models. So we can say that this model was better.

KNOWLEDGE EXTRACTED AND FUTURE RESEARCH:

From this analysis, we can understand that customer analysis was very important for any company to know about churn rate, which service they are using the most, and the reason behind the high churn rate. We can also find the correlation between the different columns and their relationship.

In the future, this project would be helpful for many giant companies to know their customer's favorites. Here we used only two models to predict the churn rate, but many other models predict the churn. Furthermore, we can use this to know the relation between columns deeply.

CONCLUSION:

Various researchers propose various data mining methods to successfully manage the churn prediction challenge. The most common data mining techniques are based on neural networks, statistical techniques, decision trees, covering algorithms, Regression Analysis, Kmeans, etc. This project provides a detailed examination of the methods used to predict customer churn. Each of the churn prediction models discussed above has advantages and disadvantages. In order to avoid customer churn, a good prediction model is required. This can be accomplished by considering a method for future Churn prediction work to process significant inputs with higher dimensions and complex attributes. Good prediction models must be developed regularly, and a combination of the proposed methods must be used.

Contribution:

Sai Prasanthi: Selected the dataset and Coding in databricks, helped in making report
Shaifali Patel: Helped in coding part, Report.

1.