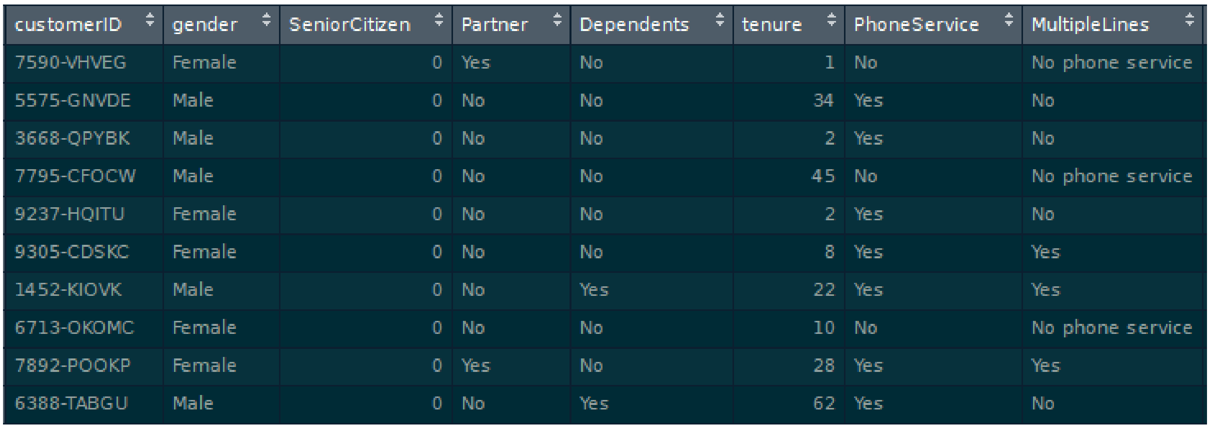
Report on Project “customer_churn”
-Banani Kashyap (DS &ML Batch)

Problem Statement –

You are the Data Scientist at a telecom company “Neo” whose customers are churning out to its competitors. You have to analyse the data of your company and find insights and stop your customers from churning out to other telecom companies.

Customer_churn Dataset:

The details regarding this ‘customer_churn’ dataset are present in the data dictionary



customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines
7590-VHVEG	Female	0	Yes	No	1	No	No phone service
5575-GNVDE	Male	0	No	No	34	Yes	No
3668-QPYBK	Male	0	No	No	2	Yes	No
7795-CFOCW	Male	0	No	No	45	No	No phone service
9237-HQITU	Female	0	No	No	2	Yes	No
9305-CDSKC	Female	0	No	No	8	Yes	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes
6713-OKOMC	Female	0	No	No	10	No	No phone service
7892-POOKP	Female	0	Yes	No	28	Yes	Yes
6388-TABGU	Male	0	No	Yes	62	Yes	No

Lab Environment: Anaconda

Domain – Telecom

Tasks to be done:

A) Data Manipulation:

a. Extract the 5th column & store it in ‘customer_5’
b. Extract the 15th column & store it in ‘customer_15’
c. Extract all the male senior citizens whose Payment Method is Electronic check & store the result in ‘senior_male_electronic’
d. Extract all those customers whose tenure is greater than 70 months or their Monthly charges is more than 100$ & store the result in ‘customer_total_tenure’
e. Extract all the customers whose Contract is of two years, payment method is Mailed check & the value of Churn is ‘Yes’ & store the result in ‘two_mail_yes’
f. Extract 333 random records from the customer_churn dataframe & store the result in ‘customer_333’
g. Get the count of different levels from the ‘Churn’ column

B) Data Visualization:

a. Build a bar-plot for the ’InternetService’ column:
   i. Set x-axis label to ‘Categories of Internet Service’
   ii. Set y-axis label to ‘Count of Categories’
   iii. Set the title of plot to be ‘Distribution of Internet Service’
   iv. Set the color of the bars to be ‘orange’

b. Build a histogram for the ‘tenure’ column:
   i. Set the number of bins to be 30
   ii. Set the color of the bins to be ‘green’
   iii. Assign the title ‘Distribution of tenure’
c. Build a scatter-plot between ‘MonthlyCharges’ & ‘tenure’. Map ‘MonthlyCharges’ to the y-axis & ‘tenure’ to the ‘x-axis’:
   i. Assign the points a color of ‘brown’
   ii. Set the x-axis label to ‘Tenure of customer’
   iii. Set the y-axis label to ‘Monthly Charges of customer’
   iv. Set the title to ‘Tenure vs Monthly Charges’
d. Build a box-plot between ‘tenure’ & ‘Contract’. Map ‘tenure’ on the y-axis & ‘Contract’ on the x-axis.

C) Linear Regression:

a. Build a simple linear model where dependent variable is ‘MonthlyCharges’ and independent variable is ‘tenure’
   i. Divide the dataset into train and test sets in 70:30 ratio.
   ii. Build the model on train set and predict the values on test set
   iii. After predicting the values, find the root mean square error
   iv. Find out the error in prediction & store the result in ‘error’

v. Find the root mean square error

D) Logistic Regression:

a. Build a simple logistic regression model where dependent variable is ‘Churn’ & independent variable is ‘MonthlyCharges’

i. Divide the dataset in 65:35 ratio

ii. Build the model on train set and predict the values on test set

iii. Build the confusion matrix and get the accuracy score

b. Build a multiple logistic regression model where dependent variable is ‘Churn’ & independent variables are ‘tenure’ & ‘MonthlyCharges’

i. Divide the dataset in 80:20 ratio

ii. Build the model on train set and predict the values on test set

iii. Build the confusion matrix and get the accuracy score

E) Decision Tree:

a. Build a decision tree model where dependent variable is ‘Churn’ & independent variable is ‘tenure’

i. Divide the dataset in 80:20 ratio

ii. Build the model on train set and predict the values on test set

iii. Build the confusion matrix and calculate the accuracy

F) Random Forest:

a. Build a Random Forest model where dependent variable is ‘Churn’ & independent variables are ‘tenure’ and ‘MonthlyCharges’

i. Divide the dataset in 70:30 ratio

ii. Build the model on train set and predict the values on test set

iii. Build the confusion matrix and calculate the accuracy

Introduction:

This report presents an analysis of customer churn using a dataset containing information about customers' demographics, services subscribed, and churn status. The analysis aims to understand factors influencing churn and to develop predictive models for identifying potential churners.

Dataset Description

The dataset consists of several features including gender, senior citizenship, partner status, tenure, internet service, payment method, monthly charges, total charges, and churn status. It contains both categorical and numerical variables.

Modules Used

1. NumPy

NumPy was imported to perform numerical computations efficiently, particularly for array operations and mathematical functions.

2. Pandas

Pandas was utilized for data manipulation tasks, including reading the dataset from a CSV file, extracting columns, filtering records, and creating dataframes.

3. Matplotlib and Seaborn

Matplotlib and Seaborn were employed for data visualization purposes. Matplotlib was used for creating basic plots such as bar plots, histograms, and scatter plots, while Seaborn was utilized for enhancing the visual aesthetics of the plots.

4. Scikit-Learn

Scikit-Learn was a crucial module for building predictive models. Various submodules from Scikit-Learn were used for tasks such as splitting the dataset into train and test sets, training machine learning models (e.g., Linear Regression, Logistic Regression, Decision Tree, Random Forest), making predictions, and evaluating model performance using metrics like accuracy score and root mean square error.

Data Manipulation

Extracted Columns:

- The 5th column 'Dependents' and 15th column 'StreamingMovies' were extracted and stored.
- Senior male citizens using electronic check as payment method were identified and stored.
- Customers with tenure greater than 70 months or monthly charges more than $100 were extracted and stored.
- Customers with a two-year contract, payment method as mailed check, and churn status 'Yes' were filtered and stored.
- 333 random records were sampled from the dataset.

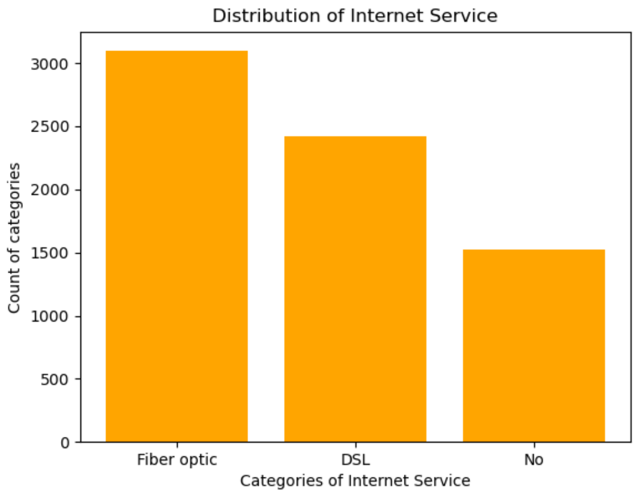
Churn Rate:

The churn rate was calculated, showing 5174 customers not churned and 1869 customers churned.

Data Visualization:

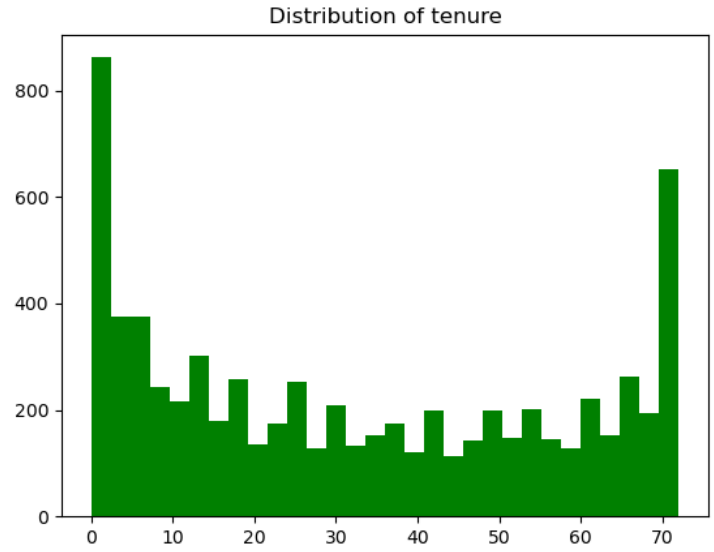
Distribution of Internet Service:

A bar plot was created to visualize the distribution of internet service categories, where DSL and Fiber optic were the primary categories.



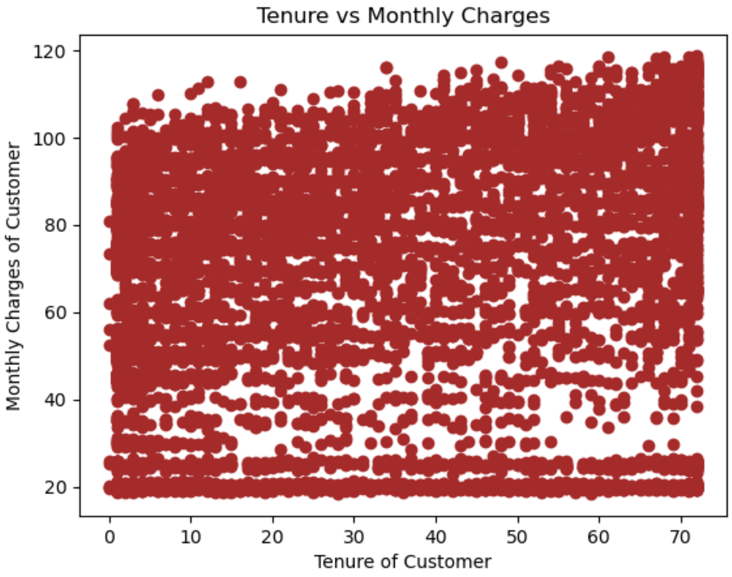
Distribution of Tenure:

A histogram was plotted to visualize the distribution of tenure, showing a skewed distribution with the majority of customers having shorter tenure.



Tenure vs Monthly Charges:

A scatter plot was generated to explore the relationship between tenure and monthly charges, showing a trend of higher charges for longer tenures.



Linear Regression

- A simple linear regression model was built with monthly charges as the dependent variable and tenure as the independent variable.
- The root mean square error (RMSE) was calculated, yielding an error of approximately 29.39.

Logistic Regression

Simple Logistic Regression:

- A logistic regression model was developed with churn as the dependent variable and monthly charges as the independent variable.
- The accuracy score was calculated to be approximately 73.60%.

Multiple Logistic Regression:

- A multiple logistic regression model was built with tenure and monthly charges as independent variables.
- The confusion matrix and classification report were generated, showing an accuracy of approximately 77.00%.

```
▪ [[934 107]
▪  [212 156]]
▪               precision    recall  f1-score   support
▪
▪           No       0.82      0.90      0.85      1041
▪          Yes       0.59      0.42      0.49       368
▪
▪     accuracy                           0.77      1409
▪    macro avg       0.70      0.66      0.67      1409
▪ weighted avg       0.76      0.77      0.76      1409
```

Decision Tree

- A decision tree model was constructed with tenure as the independent variable and churn as the dependent variable.
- The confusion matrix was computed, resulting in an accuracy of approximately 75.01%.

```
○ [[935  94]
○  [258 122]]
○ 0.7501774308019872
```

Random Forest

- A random forest model was developed with tenure and monthly charges as independent variables.
- The accuracy score of the model was approximately 65.36%

Conclusion

The analysis provides insights into factors influencing customer churn and the performance of predictive models in identifying potential churners. Strategies to improve customer retention can be devised based on the findings of this analysis.

This report summarizes the analysis conducted on the customer churn dataset, highlighting the key findings and results obtained from various data manipulation techniques, visualizations, and predictive modeling approaches.