# Predicting Customer Churn

Submitted by: Somesh Choudhary

Guided by: Mr Raviraj Choudhary

# Table of contents

# Objective

- Calculating Customer Churn using various models and comparing the results
- Using Logistic Regression, Random Forest and Decision Tree, XGBoost, Support Vector Machine, Naïve Bayes Classifier.

# Introduction

- Customer churn occurs when customers or subscribers stop doing business with a company or service, also known as customer attrition.
- It is also referred as loss of clients or customers. One industry in which churn rates are particularly useful is the telecommunications industry, because most customers have multiple options from which to choose within a geographic location.
- We are going to predict customer churn using telecom dataset.

# Keywords

- **Customer Lifetime Value:**

The net present value of a future stream of contributions to profit that result from customer transactions and contacts with the company.

- **Customer Segmentation:**

It is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately

# Keywords cont…

- **Customer Value:**

It is the perception of what a product or service is worth to a customer versus the possible alternatives . Worth means whether the customer feels s/he got benefits and services over s/he paid.

- **Customer Churn:**

It occurs when customer stop doing business with a company or service. It impedes growth , so companies should have a defined method for calculating customer churn in a given period of time.

# Technology Used

- Python is used to code the model.
- Different python libraries like Pandas, Numpy, Matplotlib, SKLearn, XGBoost, Seaborn etc. are used to manipulate and represent data.

# Dataset

- The dataset we are using is IBM Sample Data Set of a wireless telecommunication company. The data consists of 7043 records. The data was used by the company to calculate the customer churn.
- The target for prediction is the ‘Churn’ column, indicating whether or not the customer cancelled their service.

# Dataset cont…

| Attribute Name | Meaning |
|---|---|
| Customer id | Customer identification number |
| Gender | Customer gender |
| Senior citizen | If customer is senior citizen or not |
| Partner | If the customer uses the service with a partner or not |
| Tenure | For how much time the customer is subscribed |
| Phone service | Whether a customer uses phone services or not |
| Multiple lines | If a customer uses more than one phone number |
| Internet service | Which internet service customer uses |

# Dataset cont…

| Attribute Name | Meaning |
|---|---|
| Dependents | If someone else is dependent on a customer |
| Online security | Online security is provided or not |
| Online backup | Online backup is done or not |
| Device Protection | Is there any device protection available? |
| Tech Support | If customer receives support or not |
| Streaming TV | If customer watches streaming TV |
| Streaming movies | If customer is streaming movies |

# Dataset cont…

| Attribute Name | Meaning |
|---|---|
| Contract | The length of the Contract |
| Paperless Billing | How the Billing is done |
| Payment Method | What is the payment method |
| Monthly Charges | What are the monthly charges |
| Total Charges | What are the total charges |
| Churn | If the customer has cancelled the service or not |

# Data Preprocessing

- The data was downloaded from IBM Sample Data Sets. Each row represents a customer, each column contains that customer's attributes.
- The raw data contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target.
- We use isnull() to find missing values in data. We found that there are 11 missing values in "TotalCharges" columns. So, let's remove all rows with missing values.

# Preprocessing cont…

- **Looking at the variables, we can see that we have some wrangling to do.**
- We will change “No internet service” to “No” for six columns, they are: “OnlineSecurity”, “OnlineBackup”, “DeviceProtection”, “TechSupport”, “streamingTV”, “streamingMovies”.
- We will change “No phone service” to “No” for column “MultipleLines”
- Since the minimum tenure is 1 month and maximum tenure is 72 months, we can group them into five tenure groups: “0–12 Month”, “12–24 Month”, “24–48 Months”, “48–60 Month”, “> 60 Month”

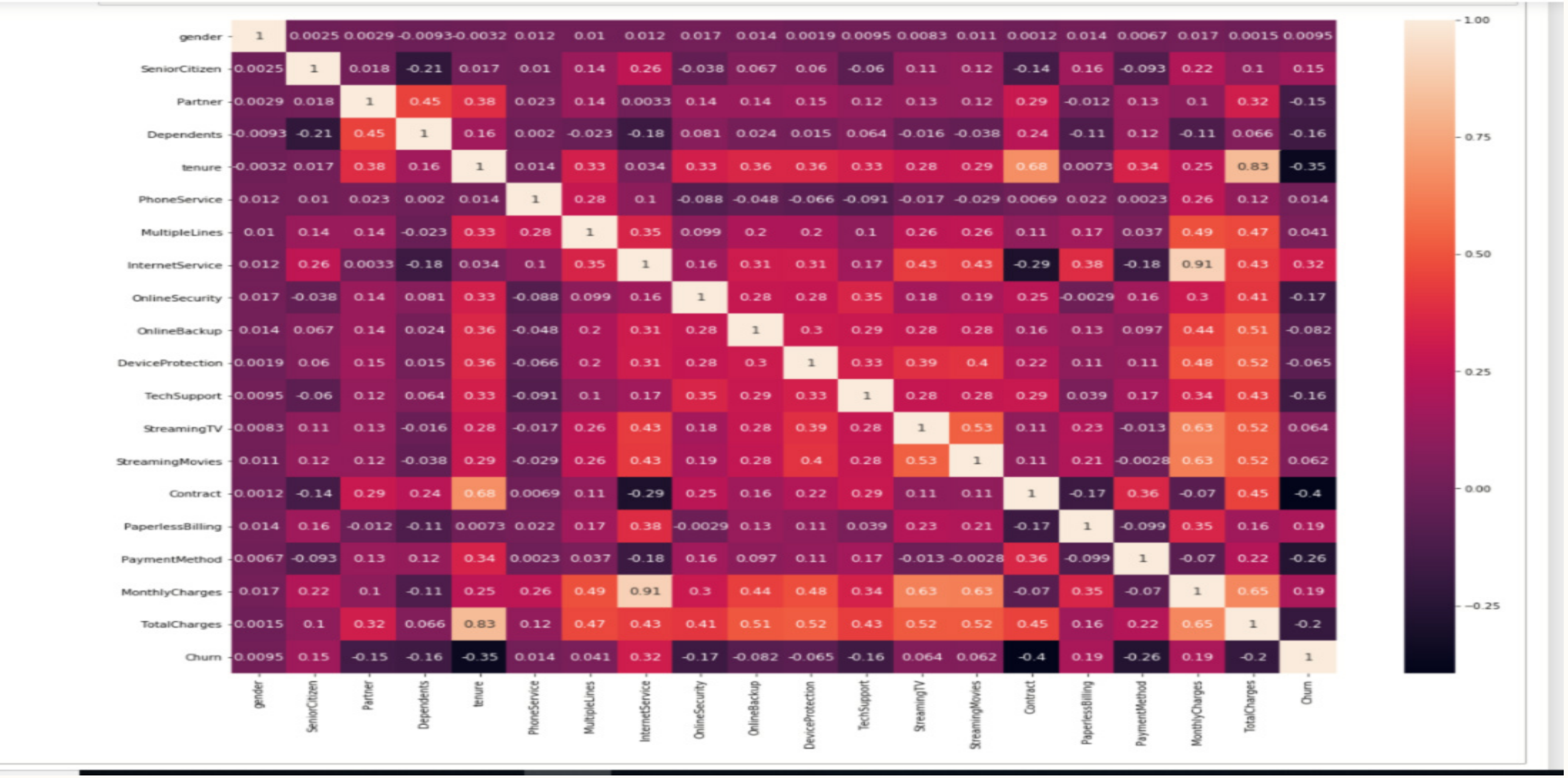
# Preprocessing cont…

- Change the values in column “SeniorCitizen” from 0 or 1 to “No” or “Yes”.
- Remove the columns we do not need for the analysis.

# Exploratory data analysis and feature selection

– **Correlation between numeric variables:** To decide which features of the data to include in our predictive churn model, we'll examine the correlation between churn and each customer feature.

# Correlation Table

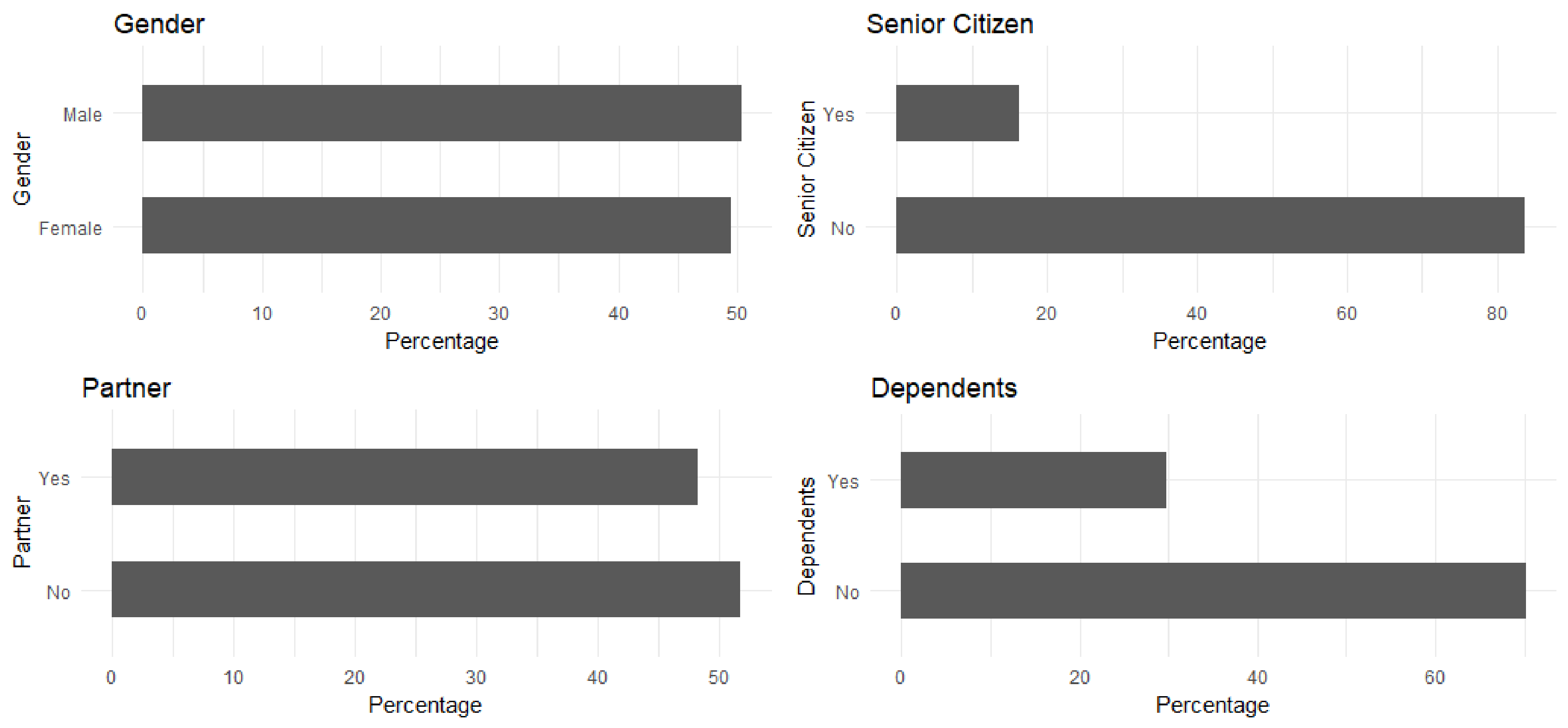


# Preprocessing cont…

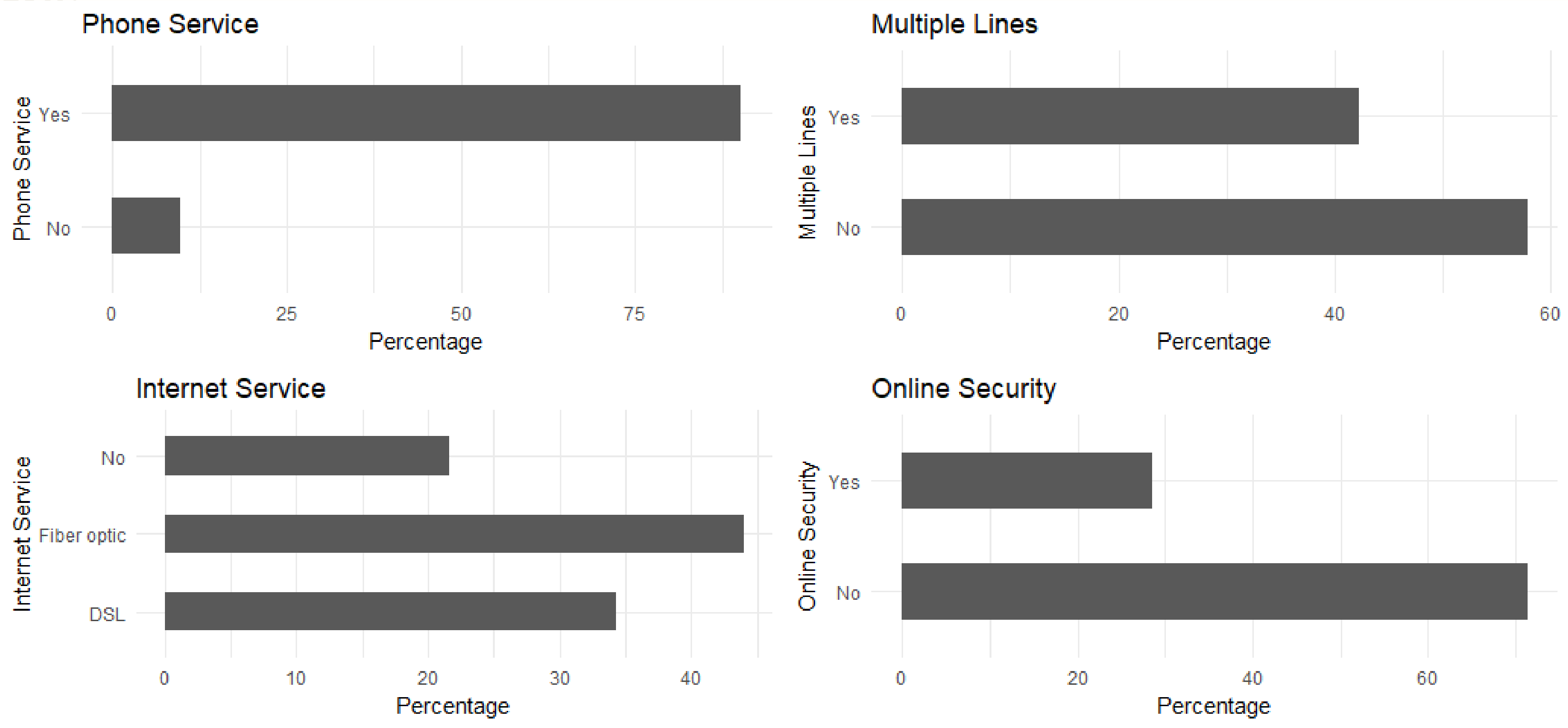
- As we can see from the correlation table the Monthly Charges and Total Charges are correlated. So one of them will be removed from the model. We remove Total Charges.

# Preprocessing cont…

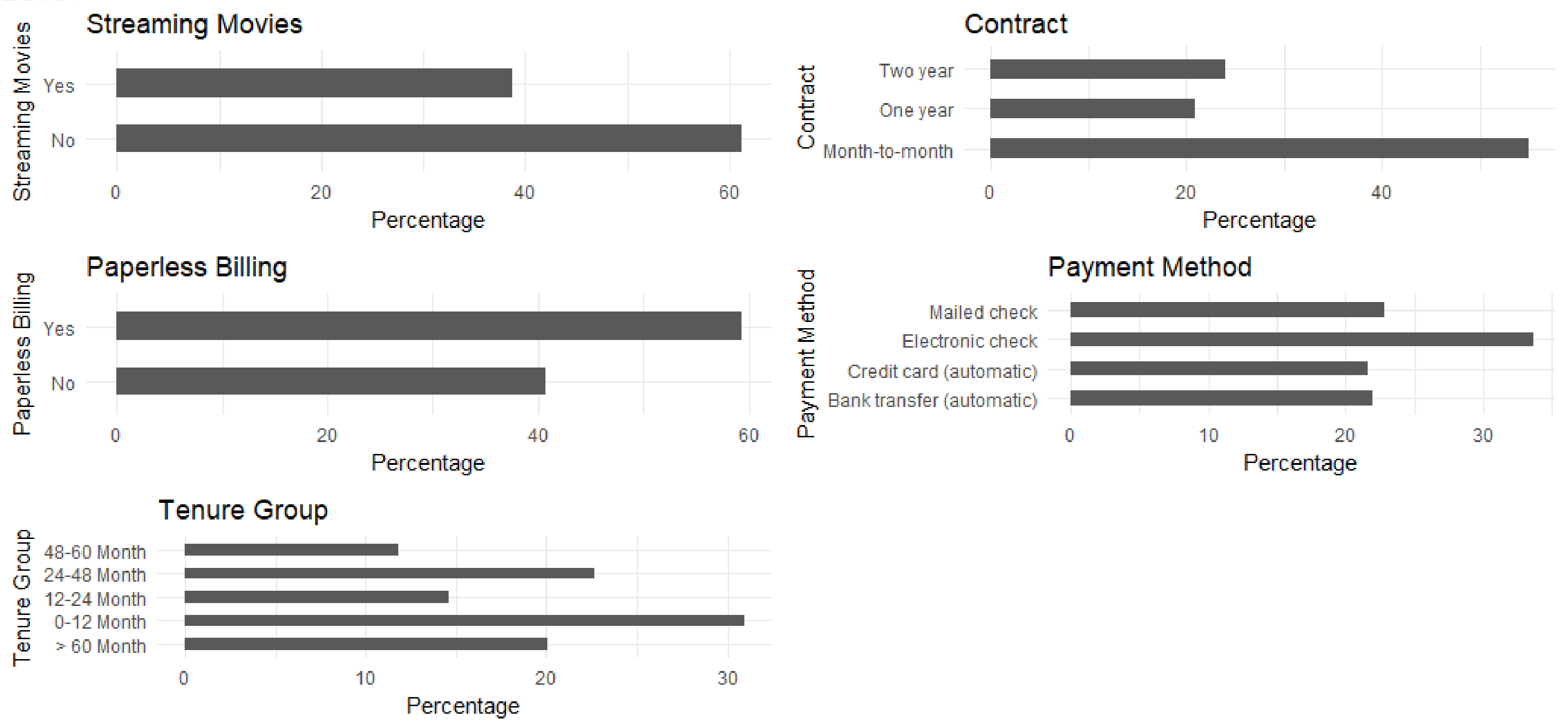
– **Bar plots of categorical variables:**



# Preprocessing cont…



# Preprocessing cont…



# Preprocessing cont…

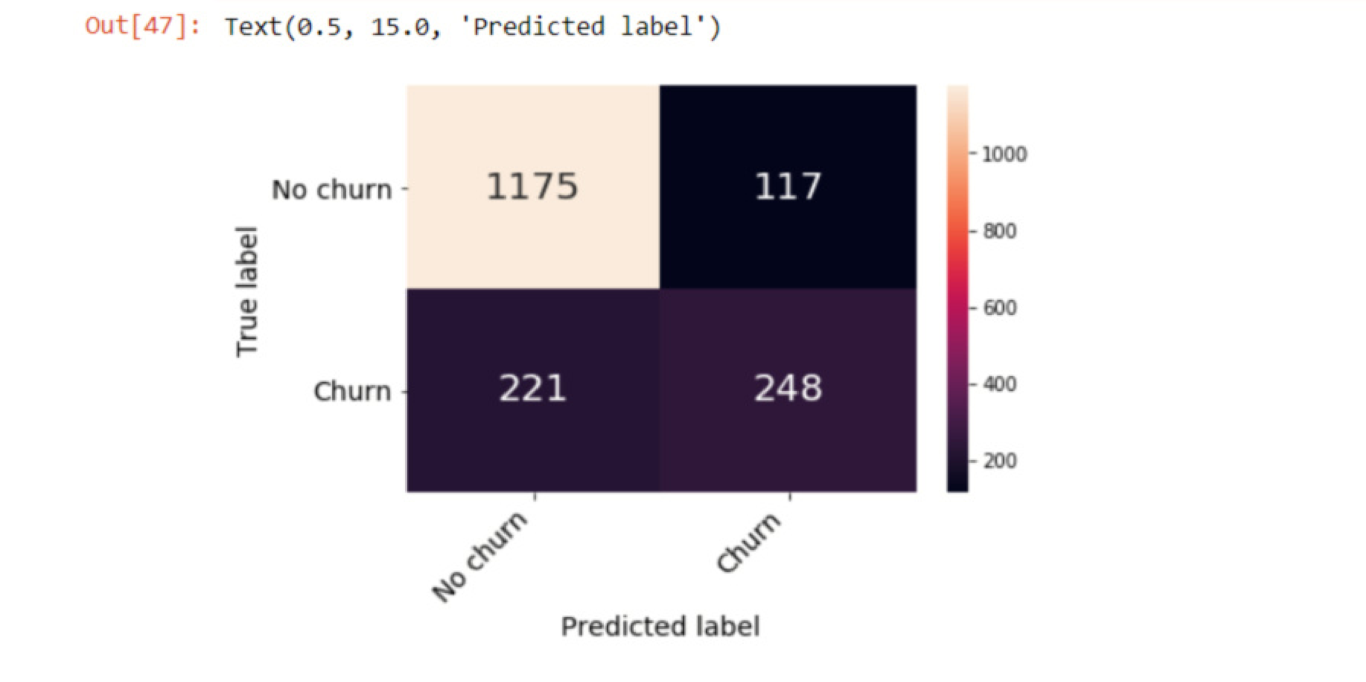
- As we can see all of the categorical variables seem to have a reasonably broad distribution, therefore, all of them will be kept for the further analysis after balancing classes.

# Splitting the data

- First, we split the data into training and testing sets.
- Training set consists of 75% of the instances and test set contains 25% instances.
- Confirm the splitting is correct.

# Applying Logistic Regression

– **Confusion Matrix:**



# Applying Logistic Regression

- Fitting the Logistic Regression Model
- Finally, the accuracy of Logistic Regression comes out to be **0.7964**

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Class 0 | 0.85 | 0.90 | 0.88 | 1305 |
| Class 1 | 0.66 | 0.55 | 0.60 | 456 |

# Applying Logistic Regression

- We got about 80% classification accuracy from our logistic regression classifier. But the precision and recall for predictions in the positive class (churn) are relatively low.
- This indicates that the data has **imbalanced classes.**

# Handling Imbalanced Classes

- There are 5174 instances in Class 0 (no churn) and 1869 instances in Class 1 (churn). Clear imbalance in classes.
- So minority class would be upsampled.
- Minority Class is upsampled to 5174 instances.
- Logistic Regression is applied in newly data created by balanced classes.
- Although accuracy drops to **0.76** but precision and recall for Class 1 are significantly increased.
- So further modelling would be done on balanced classes.

# Handling Imbalanced Classes

- Accuracy of logistic regression classifier on test set : **0.76**

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Class 0 | 0.78 | 0.73 | 0.75 | 1303 |
| Class 1 | 0.74 | 0.79 | 0.77 | 1284 |

# Applying Decision Tree

– **Decision Tree confusion matrix:**

| | Actual NO | Actual Yes |
|---|---|---|
| Predicted NO | 1395 | 346 |
| Predicted Yes | 153 | 214 |

# Applying Decision Tree

- Accuracy of unpruned Decision Tree on training set = **1.0**
- Accuracy of unpruned Decision Tree on test set = **0.73**

- Decision tree is giving high testing error and low training error, this means Decision tree is overfitting. So the tree has to be pruned.
- Accuracy of pruned Decision Tree on training set = **0.79**
- Accuracy of pruned Decision Tree on test set = **0.79**

- The accuracy has not improved from logistic regression.

# Applying Random Forest

- **Random Forest confusion matrix:**

| | Actual No | Actual Yes |
|---|---|---|
| Predicted No | 1381 | 281 |
| Predicted Yes | 167 | 279 |

# Applying Random Forest

- After applying random forest, we found that the accuracy is **0.78473**
- So it performs better than Decision Tree but it’s accuracy is slightly lesser than that of Logistic regression.

# Applying SVM

- Accuracy for SVM classifier comes out to be **0.81**.

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Class 0 | 0.85 | 0.77 | 0.81 | 1295 |
| Class 1 | 0.79 | 0.87 | 0.82 | 1292 |

# Applying kNN Classifier

– Accuracy of kNN Classifier comes out to be **0.77**.

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Class 0 | 0.84 | 0.69 | 0.76 | 1295 |
| Class 1 | 0.74 | 0.87 | 0.80 | 1292 |

# Applying Naïve Bayes Classifier

- Three Naive Bayes Classifiers, Gaussian NB, Bernoulli NB and Multinomial NB, are applied.

| | Accuracy |
|---|---|
| Gaussian NB | 0.73 |
| Bernoulli NB | 0.71 |
| Multinomial NB | 0.72 |

- Naïve Bayes Classifiers did not perform well compared to other models.

# Applying XGBoost

- To obtain better results, boosting is performed.
- XGBoost is used for modelling.
- Hyperparameter tuning is done because XGBoost is highly sensitive to hyperparameters.

# Applying XGBoost

– Accuracy of XGBoost with best tuned parameters comes out to be **0.7974**.

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Class 0 | 0.83 | 0.67 | 0.74 | 1278 |
| Class 1 | 0.73 | 0.86 | 0.79 | 1309 |

# Results

| Model Name | Accuracy |
|---|---|
| Logistic Regression | **0.7964** |
| Decision Tree | **0.79** |
| Random Forest | **0.7847** |
| Gaussian NB | **0.73** |
| Bernoulli NB | **0.71** |
| Multinomial NB | **0.72** |
| kNN | **0.77** |
| SVM | **0.81** |
| XGBoost | **0.7974** |

# Conclusion

- SVM classifier performs best for predicting customer churn based on given dataset.
- This model can work for both categorical and numerical data with small tweaking in pre-processing step based on dataset.
- Model can handle imbalance in classes.
- Based on distribution of data, at least on of these classifiers should do the job well enough.
- Due to it's modularity, this model can be used for predicting customer churn for almost every company dataset.

# Thank You!

Questions?