# **Customer Churn Analysis and Prediction**

****

## **Introduction**

Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.

Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can prioritize focused marketing efforts on that subset of their customer base.

## **Table of contents**

1) Problem Definition

* Context
* Content

2) Data Analysis

3) Descriptive analysis and EDA (Exploratory Data Analysis)

* Churn customer cluster analysis — by Demographic, Account Type, and Service Usage

4) Pre-processing Pipeline

5) Building Machine Learning Models

6) Concluding Remarks

## **Problem Definition**

**Context**

You have to predict if a customer is going to churn or not by analyzing customer behavior and all relevant customer data provided to you. In the end, produce an efficient model which could be used for customer retention programs.

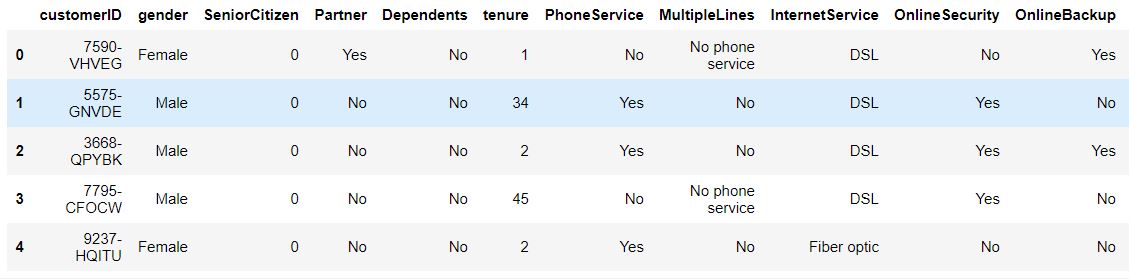
**Content**

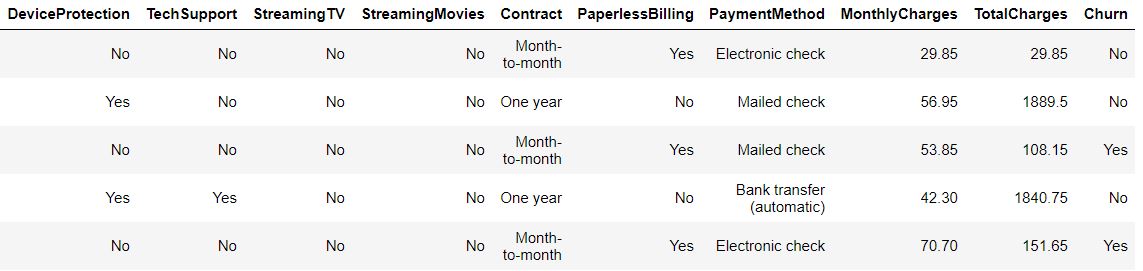
Dataset is provided by IBM Telecom Industry. Each row represents a customer, each column contains the customer's attributes described on the column Metadata.

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

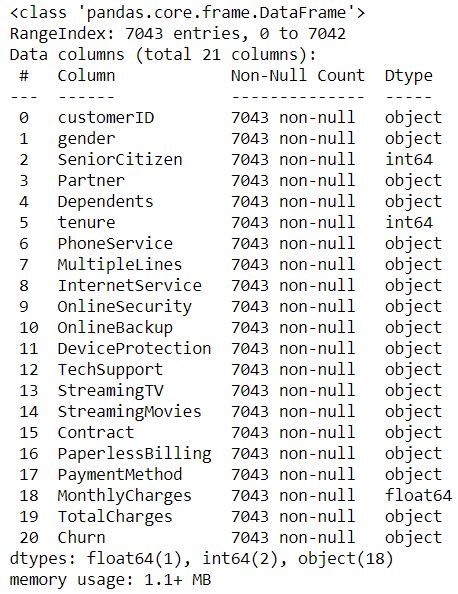
## **Data Analysis**





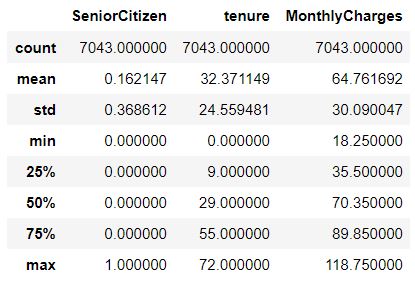
Our dataset looks like the above with all the columns for the first five customers.





Dataset has 7043 rows and 21 columns including the label column. There are no null values in the dataset. There are 3 columns with numerical columns that rest all object types including the Total charges column which contains numerical type values.



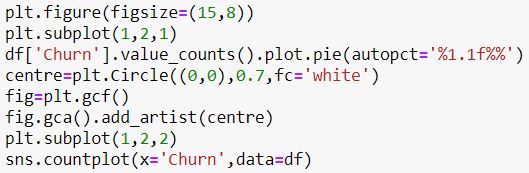


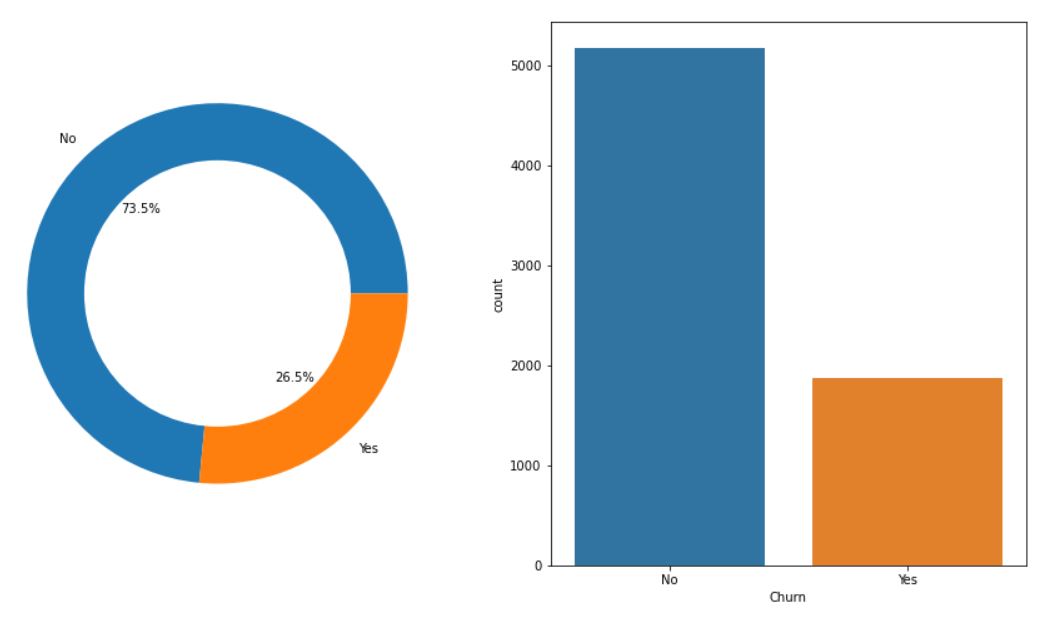
All the columns are no described here as they are of object datatype. We will have to encode them to carry out such statistical evaluation on them. From the above data, we see that count of every column is 7043. Mean is higher than the median in the tenure column, and lower in Monthly charges. Both the columns show a little skewness. The difference between the interquartile range, min, and max doesn't vary much which means there are fewer outliers. There is a high variance in the tenure and Monthly charges column.

## **Descriptive analysis and EDA (Exploratory Data Analysis)**

We’ll go through each column iteratively and see which ones are useful for ML modeling later on. Some columns may need more pre-processing than others to get ready to use an algorithm.

**Churn**

****

****

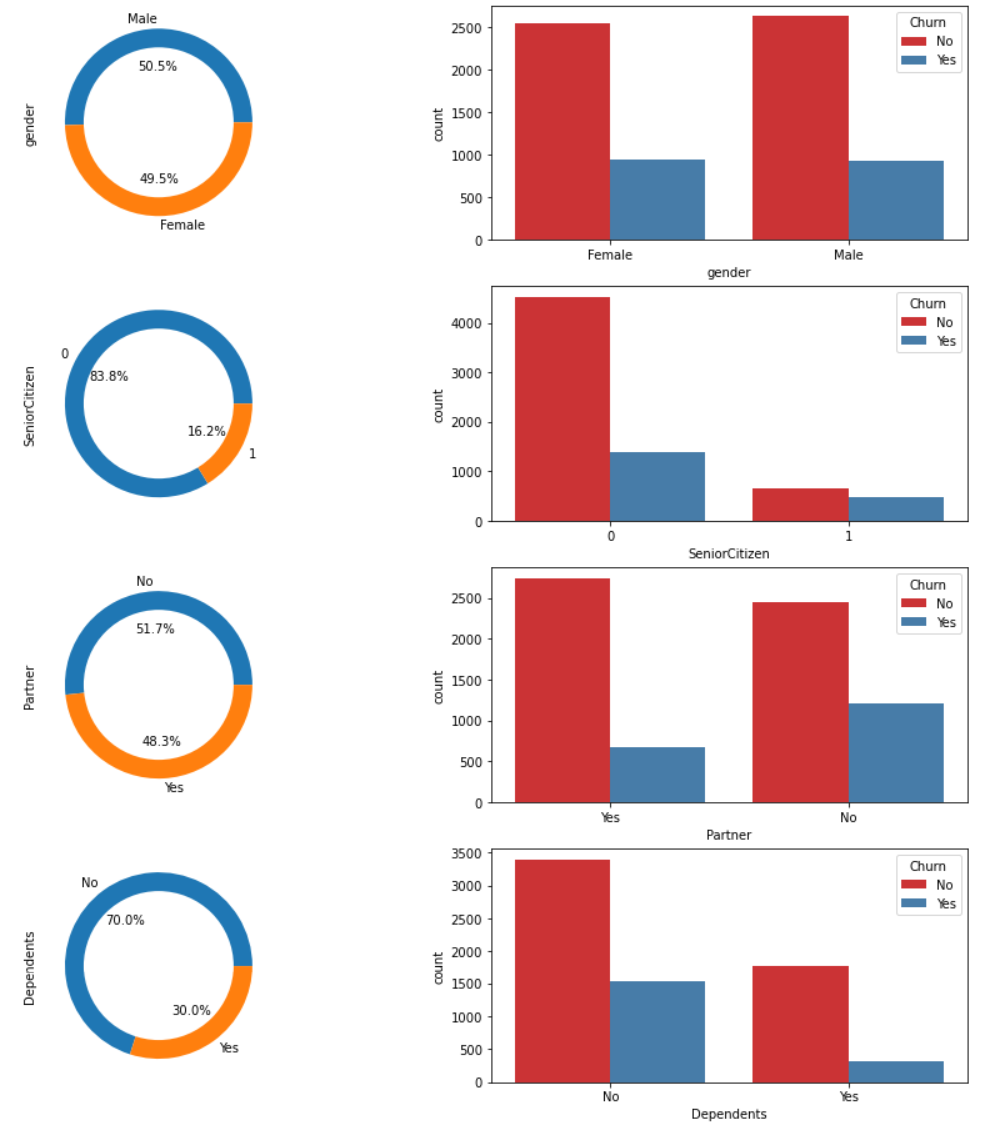
This is our label column which tells us if the customer is going to churn or not.

* The graph shows that around 26.5% of the customers have churned from the company while 73.5% haven't.

**Demographic info about customers**

Let’s explore how the demographic information of customers such as Gender, Partner, Dependents, etc affects the churning.



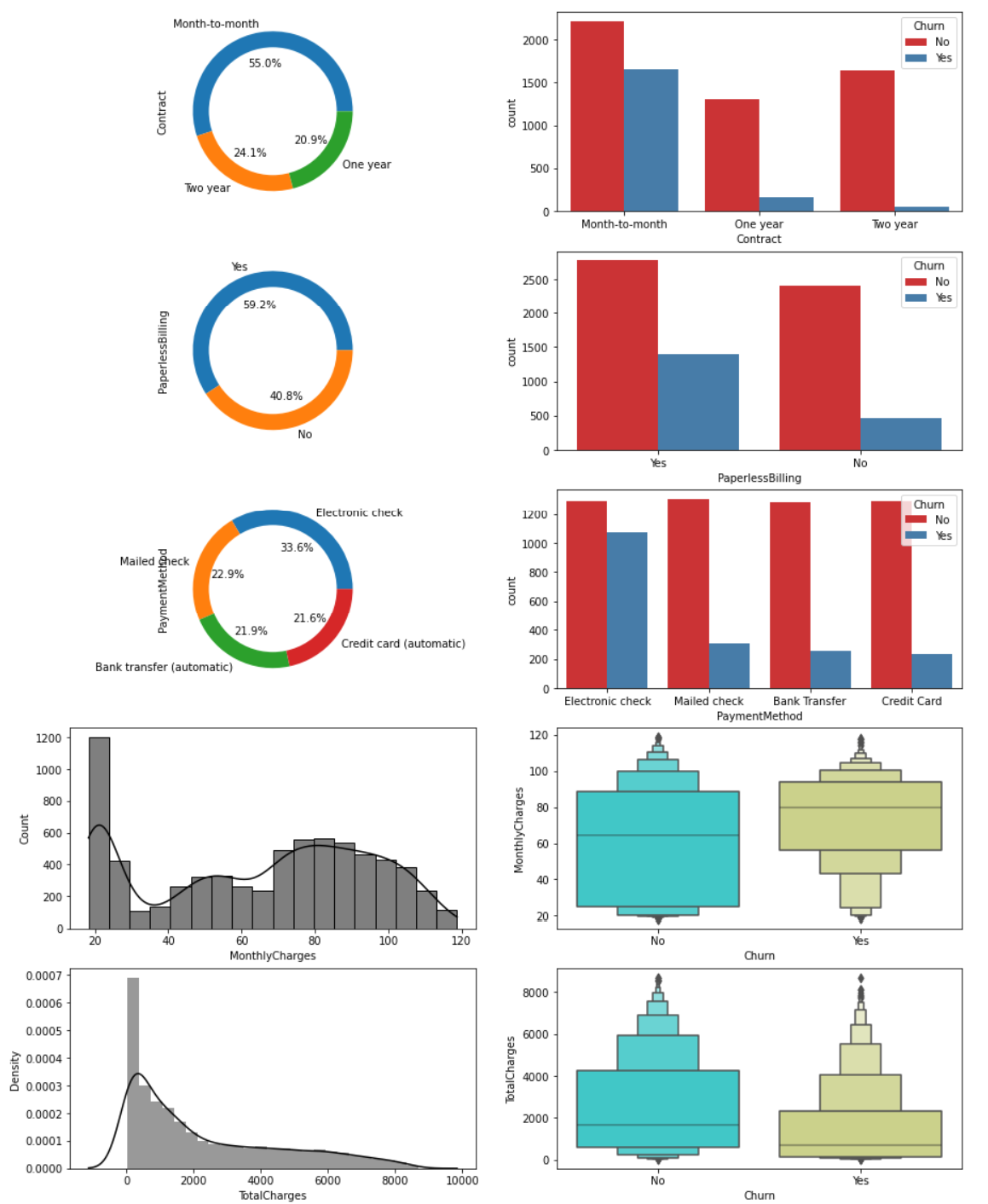


* There are the almost same number of males and females as customers and gender alone does not show any signs of the customer is going to churn or not.
* Around 16% of the customers are senior citizens. Chances of churning of senior citizens are way greater than the younger ones.
* Customers with or without customers almost are of the same numbers. Customers who do not have partners have a greater ratio of churning.
* There are 30% of customers with dependents. Customers with dependents have a lower ratio of churning than the customers without.

**Customer account information**

Account information of customers is an important detail which can tell us a great deal about them and help us in retaining them.

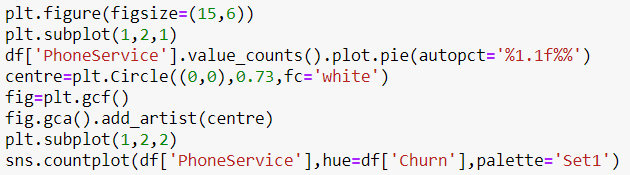


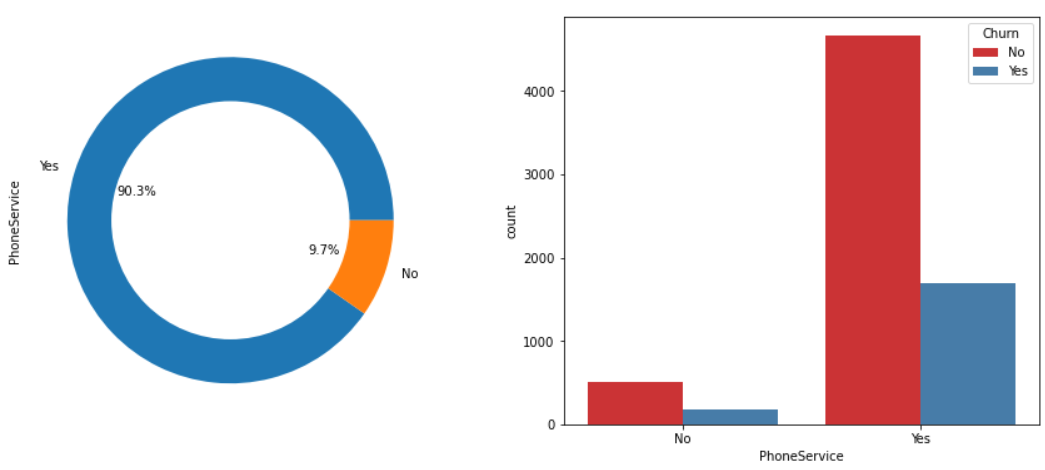


* The majority of the customers stick to the month-to-month contract and customer retention seems to more for a higher duration of contact.
* Around 60% of customers choose paperless billing while they are the ones who have a greater churning ratio than those who don't.
* Customers paying through electronic checks are slightly higher than other methods, also these customers have the highest ratio of churning.
* The majority of customers have lower monthly charges below 30. Monthly charges data is almost normally distributed. People paying higher monthly charges have a higher churn rate.
* There is a high peak of customers paying total charges less than 1000. Data is skewed towards the right. Customers who do not churn have higher total charges which are obvious.

**Major services provided are Phone service and Internet service**

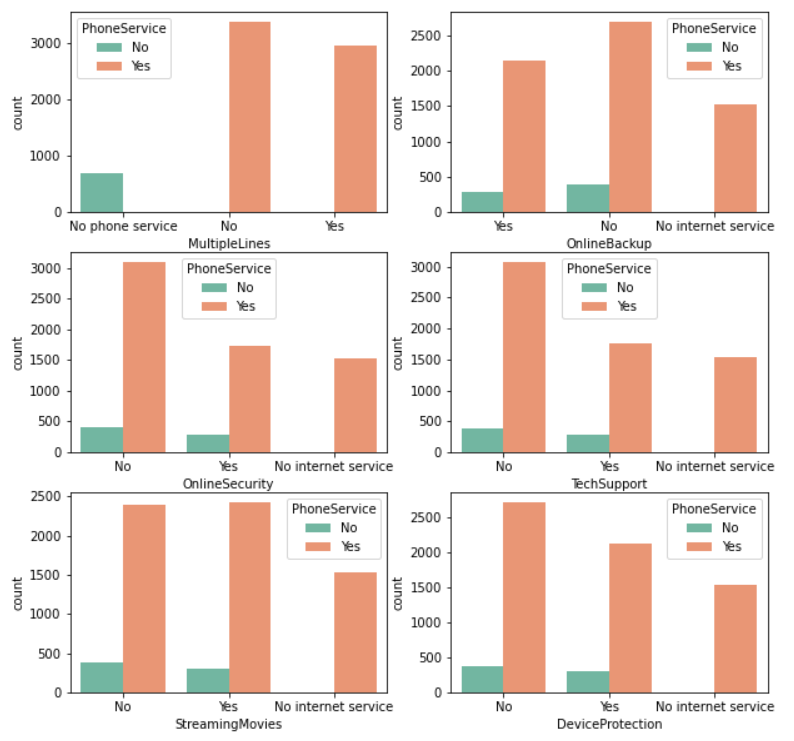
Phone service





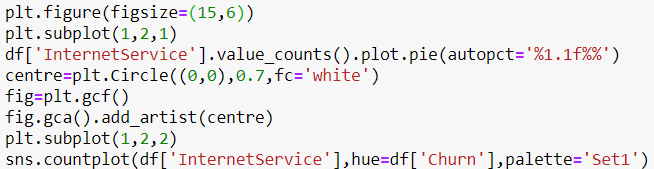
Customers who do not use Phone service are below 10% while customers using them have the same churning rate as that of customers who do not.

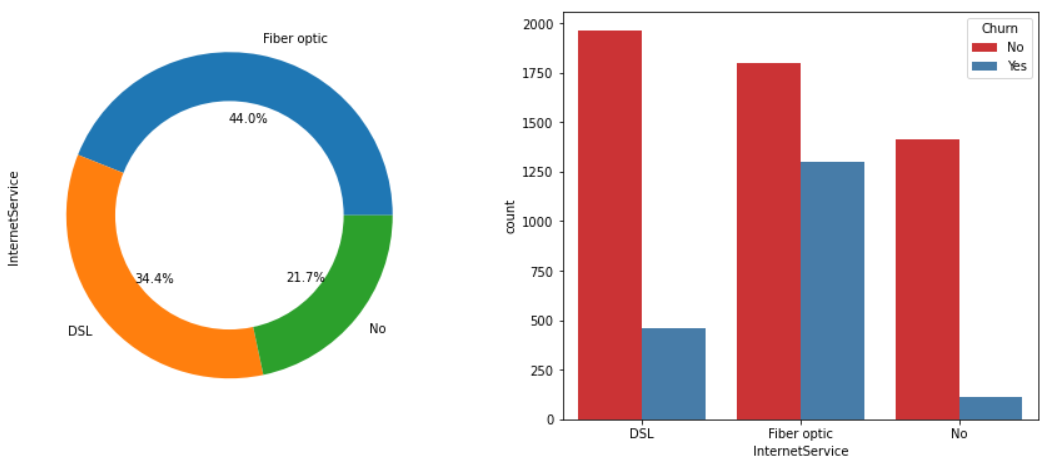




* There is an option of multiple lines if you use phone service which is taken by a lot of customers.
* The 10% of customers who do not use the phone service, use the internet service.
* Services other than Multiple lines are only meant for customers who use internet services.

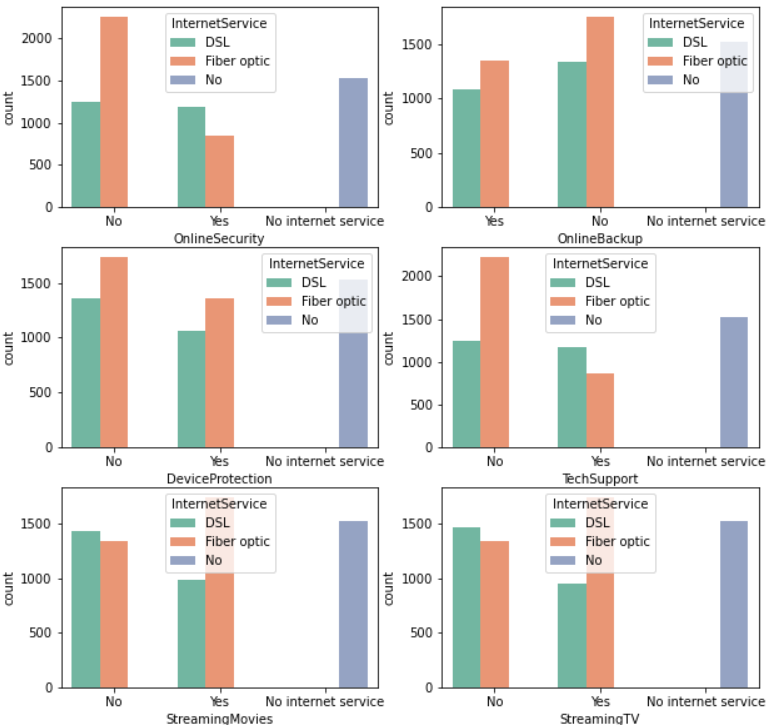
Internet service





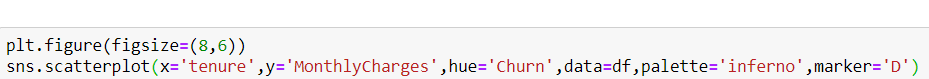
* There are three categories for internet service, among which fiber optic is the more popular one as it is a new technology and provides much high speed than the DSL.
* Customers using the fiber optic have the highest churning ratio there could be various reasons for it, higher price could be one of those as the fiber optic is a new service providing higher internet speed and hence costlier.

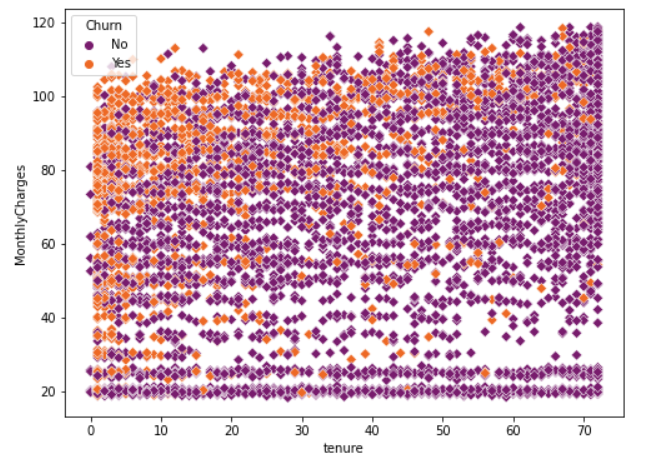




* Internet service doesn't seem to be an important feature as there is already a category in all the other features of no phone service so we are going to remove it in the pre-processing phase.
* People who do not have internet service do not use the above features.
* Customers having fiber optic have a higher ratio of not opting for Online Security, Device Protection, and tech Support while they show more interest in Online Backup, Streaming Movies, and Tv services.
* Customers having DSL internet service have a higher ratio of opting for Online Security, Device Protection, and tech Support while they do not opt for Online Backup, Streaming Movies, and Tv services.

Churning of customers based on charges paid by them and tenure

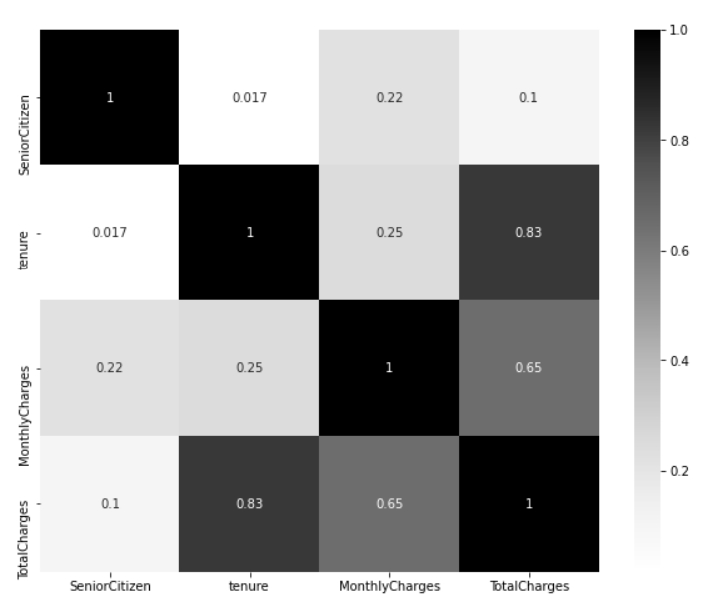




* Customers have a high churn rate at lower tenure and higher monthly charges
* As the tenure increases, very few customers leave the company.
* Monthly Charges affect the customers more as even at a higher tenure period if Monthly charges are high, the number of people churning increases.

Heatmap





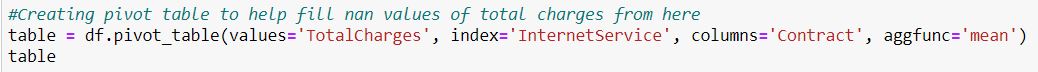
There is a high correlation between Total charges and tenure. Also a high correlation between Monthly charges and Total Charges. This correlation leads to multicollinearity which needs to be dealt with to avoid overfitting.

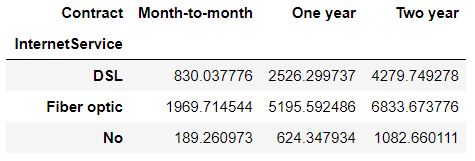
## **Pre-processing Pipeline**

First of all, we drop unnecessary columns. CustomerId is an identifier column so it is not useful in our prediction so, we remove it, and also as we discussed earlier there is already a category of no Internet Service in services provided by the company, hence we remove it too.

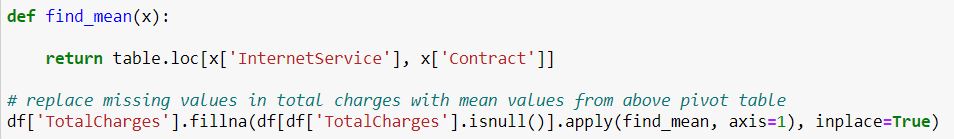


After removing unnecessary columns, we fill missing values in the columns. The only missing values are in the total charges column. We fill the missing values of this column with the help of Internet service and Contract columns by creating a pivot table between these two and filling the values with the mean of Total Charges.



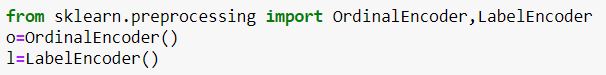


Suppose, a customer has a missing value for total charges and this customer uses DSL internet service and has a month-to-month contract then we will fill the missing value of total charges by 830.037776.

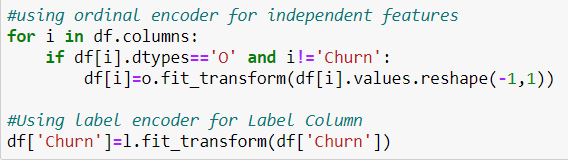


We use the above function which serves as mapping for the pivot table we have created and imputes the missing values.

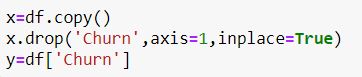
Since there were no outliers and skewness in our dataset as seen in the EDA phase, we further move on to converting the categorical features into numerical form.



We import both Ordinal and Label Encoder for encoding our categorical data. We use Ordinal Encoder to encode the independent features while Label Encoder to encode the dependent features.



We further separate dependent and independent features for simplicity.

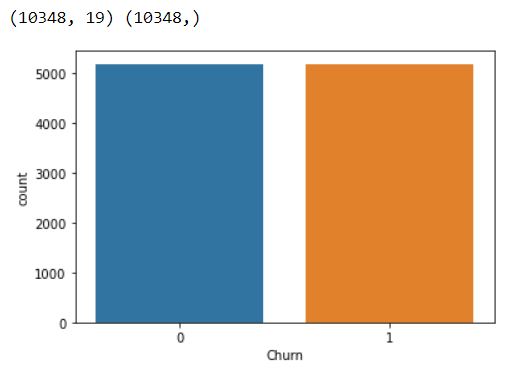


We now perform oversampling on the imbalanced dataset using SMOTE.









As we can see now that the count of customers churned and the customers who did not have become equal. Also, the rows in the dataset have increased to 10,348 as we have introduced new samples of churn customers.

Since our dataset contains features highly varying in magnitudes, units, and range, therefore we will scale them using Min Max Scaler.

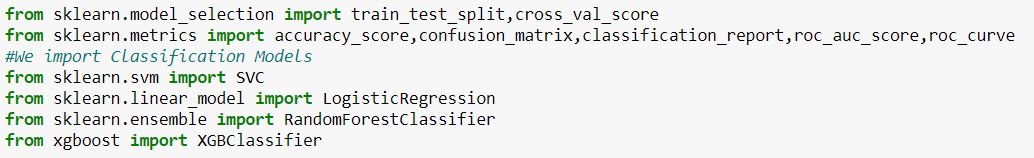




Min Max Scaler transforms a feature into a given range and by default, it transforms it between 0 to 1.

## **Building Machine Learning Models**

Importing the necessary libraries



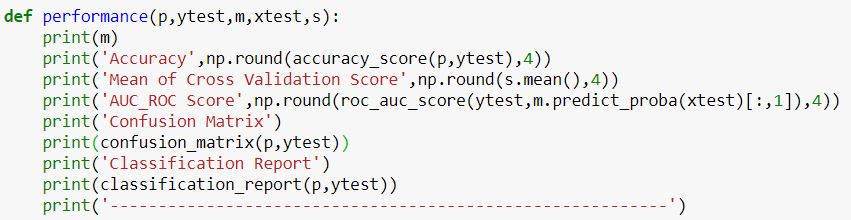
For our prediction, we make use of Support Vector Classifier, Logistic Regression models, and ensemble models such as Random Forest Classifier and Xtreme Gradient Classifier. We put these models on a list to evaluate them one by one.



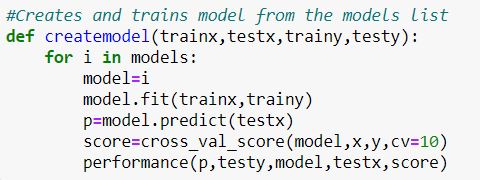
We split our data into train and test set



We create a display function that will display the performance metrics of each of the classification model

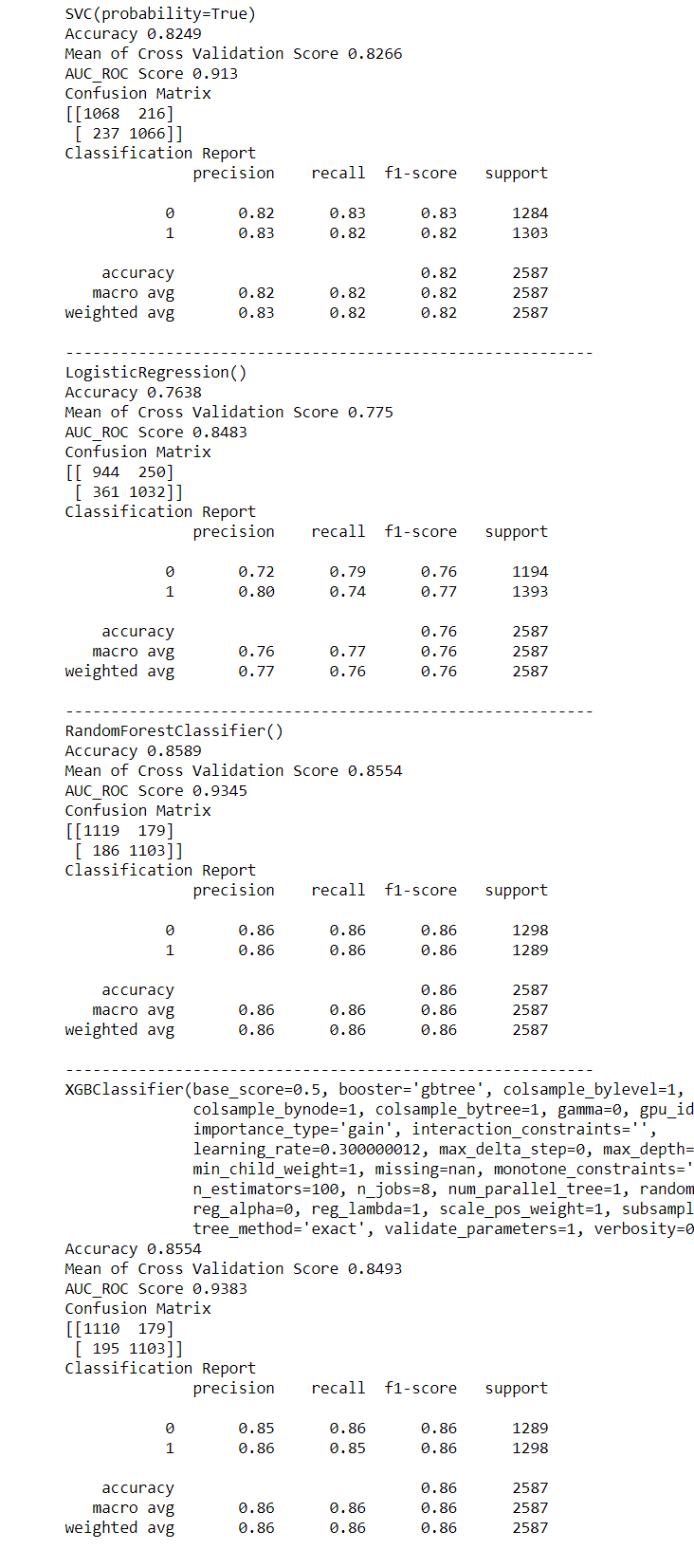


Now, we create a function that will create model one by one from the model list, predict the test set, calculate the cross-validation score and then call the performance function to display the model’s performance



Calling the model creating function.

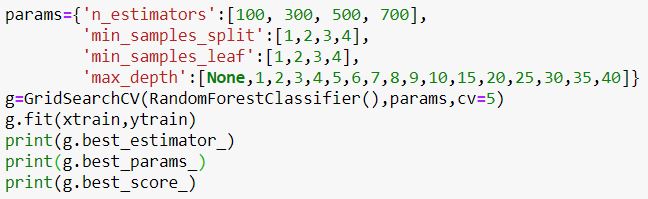


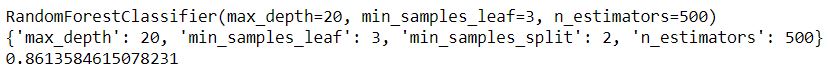


As we can see that Random Forest and XGBClassifier outperform other models and on the close analysis we find out that random forest has a better cross-validation score than XGB. A high cross-validation score states that the model is more generalized and has a low bias.

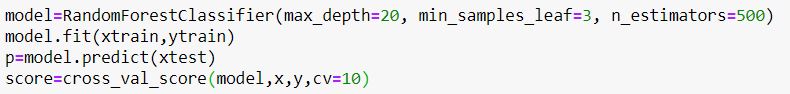
## **Hyperparameter Tuning**

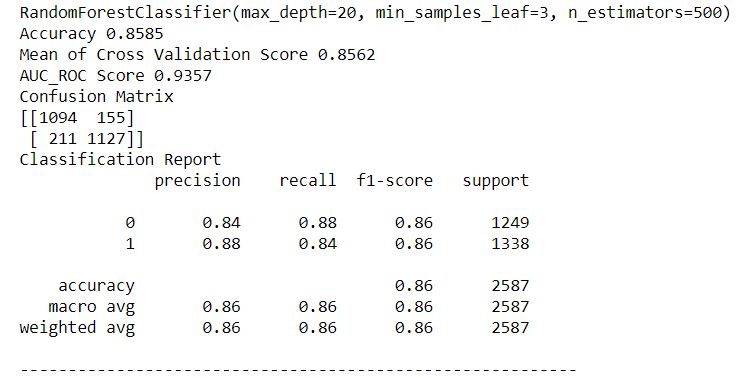
We perform hyperparameter tuning on Random Forest Classifier model to further increase the prediction efficiency using GridSearchCV.





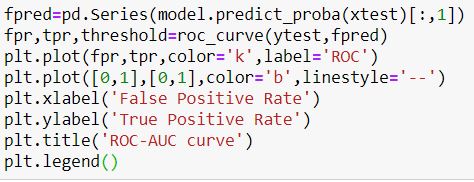
GridSearchCV provides us with the above parameters as the best parameters. So, we use these parameters to create our final model.

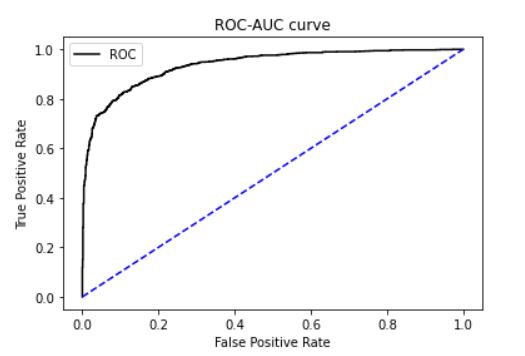




The cross-Validation score and AUC ROC score have increased slightly. There is an increase in precision for the 1st category, i.e. for people who churn which is a good thing as we want to discover people who are going to churn more accurately than those who are not.

We can further plot the AUC-ROC curve of the model to get a better understanding of the performance.





## **Conclusion**

To succeed at retaining customers who would otherwise leave the business, marketers and retention specialists must be able to (a) foretell in advance which customers are going to churn by churn analysis and (b) know which marketing actions will have the most prominent retention impact on each particular customer. Armed with this understanding, a large proportion of customer churn can be reduced.

While simple, in theory, the actualities involved with producing this "proactive retention" goal are extremely challenging.

We can use our model to calculate the probability of churn of a customer and based on this retention plans can be applied to those customers who have a high probability of churning.

To do this, we first calculate the churn probability of each customer and add it as a new feature in our Data Frame.



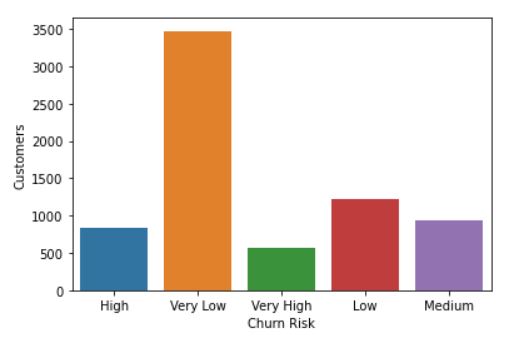


Now, we classify categories based on the churn probability of a customer.



Lastly, we have visualized the Risk Factor of customers. Customers who are at high risk of churn can be provided with special retention-focused offers or incentives to keep them happy and prevent them from churning.





**Acknowledgments**

[Student Support Team (datatrained.com)](https://support.datatrained.com/)

[Telco customer churn - IBM Documentation](https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=samples-telco-customer-churn)

[scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation](https://scikit-learn.org/stable/index.html)

[Recently Active 'machine-learning' Questions - Stack Overflow](https://stackoverflow.com/questions/tagged/machine-learning)

For more details, please check out the source code on [Github](https://github.com/bitsplease98/Datascience-Practice-Projects/tree/main/Doctor's%20Consultancy%20fees%20prediction)