REPORT ON TELECOM USERS DATASET

Submitted By:

Gurleen Oberoi

Submitted To:

Sally Weatherall

Introduction

Provided dataset is from a telecommunications company. The data contains information about almost six thousand users, their demographic characteristics, the services they use, the duration of using the operator's services, the method of payment, and the amount of payment. The task is to analyse the data and predict the churn of users (to identify people who will and will not renew their contract). If the company know in advance the customers who will churn, company can try to keep that customers by giving offers/discounts etc.

So from my analysis I will suggest which machine learning model is best in predicting churn and also check if the trained model is Fair as the dataset consist of some sensitive features i.e. gender, senior citizen.

Data Summary

Dataset contains 5986 rows and 22 columns.

Column Description:

1 customerID - customer id

2 gender - client gender (male / female)

3 SeniorCitizen - is the client retired (1, 0)

4 Partner - is the client married (Yes, No)

5 tenure - how many months a person has been a client of the company

6 PhoneService - is the telephone service connected (Yes, No)

7 MultipleLines - are multiple phone lines connected (Yes, No, No phone service)

8 InternetService - client's Internet service provider (DSL, Fiber optic, No)

9 OnlineSecurity - is the online security service connected (Yes, No, No internet service)

10 OnlineBackup - is the online backup service activated (Yes, No, No internet service)

11 DeviceProtection - does the client have equipment insurance (Yes, No, No internet service)

12 TechSupport - is the technical support service connected (Yes, No, No internet service)

13 StreamingTV - is the streaming TV service connected (Yes, No, No internet service)

14 StreamingMovies - is the streaming cinema service activated (Yes, No, No internet service)

15 Contract - type of customer contract (Month-to-month, One year, Two year)

16 PaperlessBilling - whether the client uses paperless billing (Yes, No)

17 PaymentMethod - payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

18 MonthlyCharges - current monthly payment

19 TotalCharges - the total amount that the client paid for the services for the entire time

20 Dependents - (Yes or no)

21 Unnamed:0

22 Churn - whether there was a churn (Yes or No)

PreProcessing

Dropping Column Unnamed:0 and customerID from the dataframe as there is no relation of these columns to Churn. Now we have dataframe which consist 20 columns.

```
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   gender            5986 non-null   object
 1   SeniorCitizen     5986 non-null   int64
 2   Partner           5986 non-null   object
 3   Dependents        5986 non-null   object
 4   tenure            5986 non-null   int64
 5   PhoneService      5986 non-null   object
 6   MultipleLines     5986 non-null   object
 7   InternetService   5986 non-null   object
 8   OnlineSecurity    5986 non-null   object
 9   OnlineBackup      5986 non-null   object
 10  DeviceProtection  5986 non-null   object
 11  TechSupport       5986 non-null   object
 12  StreamingTV       5986 non-null   object
 13  StreamingMovies   5986 non-null   object
 14  Contract          5986 non-null   object
 15  PaperlessBilling  5986 non-null   object
 16  PaymentMethod     5986 non-null   object
 17  MonthlyCharges    5986 non-null   float64
 18  TotalCharges      5986 non-null   object
 19  Churn             5986 non-null   object
dtypes: float64(1), int64(2), object(17)
```

Column TotalCharges datatype is object not int/float so converted the object data type into float. Also there are 10 null values in the column. Replaced these null values with 0 as the corresponding value in tenure column is 0. So it's simple that if tenure is 0 then totalcharges would also be 0.

For machine learning algorithm to work efficiently labels are converted into numeric form.

Label	**Replaced By:**
Yes	1
No	0
No internet service	0
No phone service	0
Male	1
Female	0

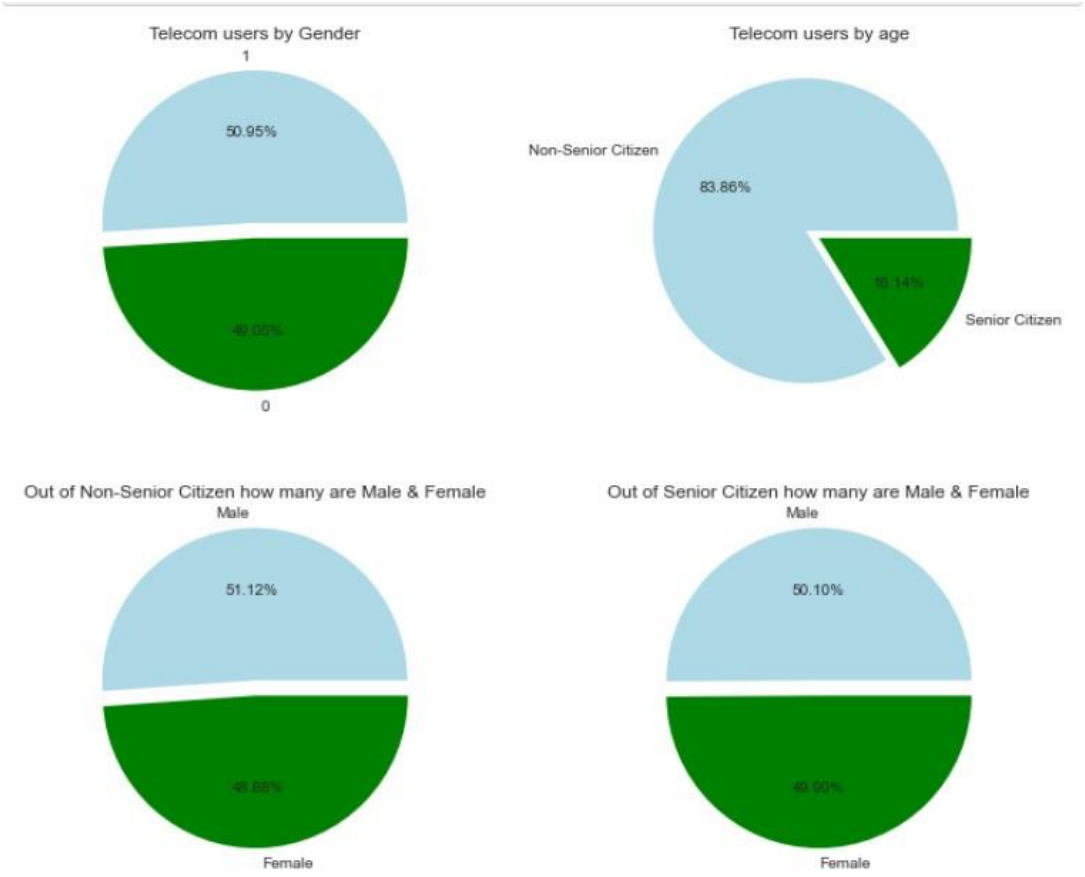
For categorical variables InternetService, PaymentMethod and Contract where there is no ordinal relationship, the integer encoding is not effective. So one hot coding is performed on them.

k-1 Dummies were created out of k categorical levels by removing the first level, for columns InternetService, PaymentMethod and Contract.

Exploratory Data Analysis

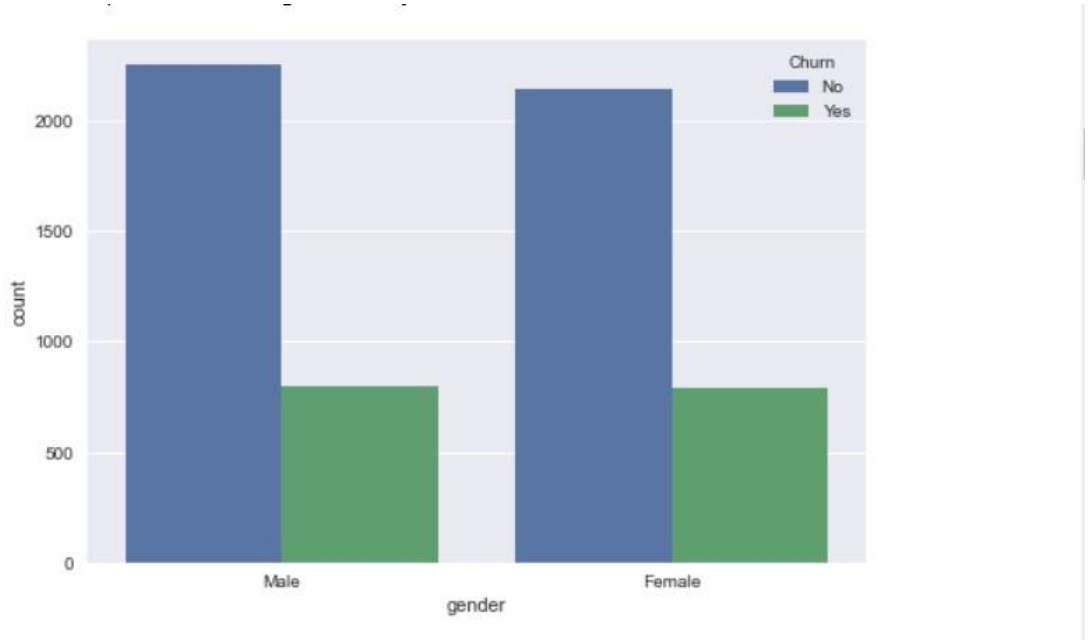
Data Analysis helps to get better understanding of the data.

1.



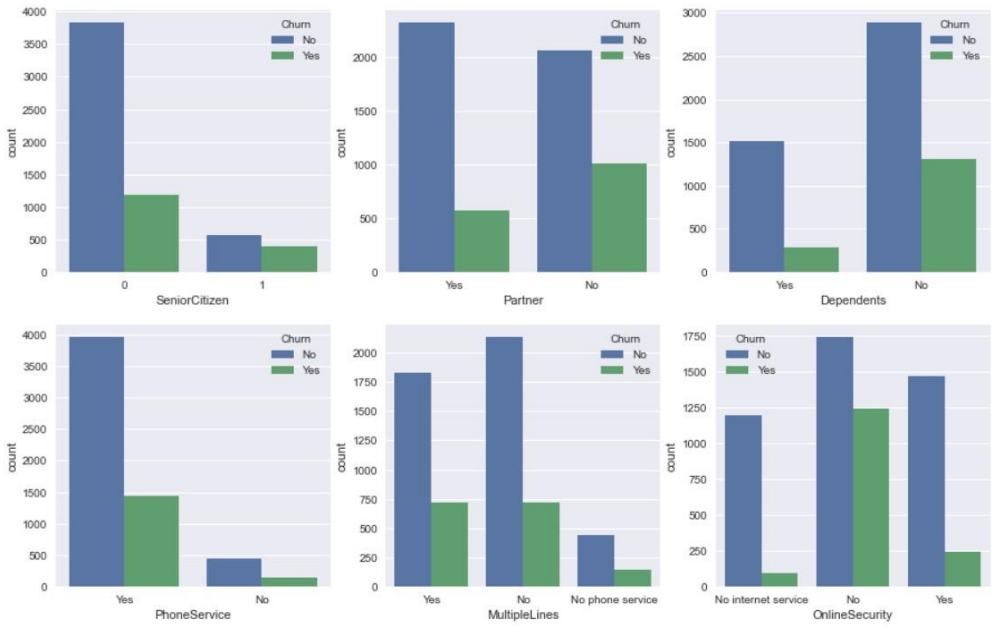
It is clear that male users are more than female but the difference is very small. Senior citizens are less than non-senior citizens.

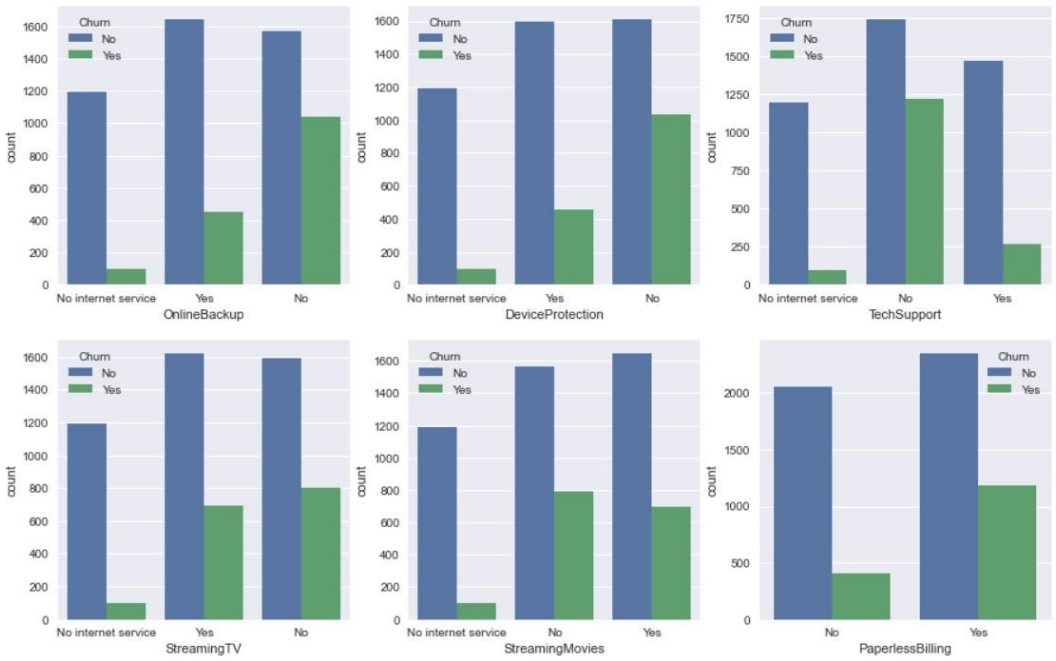
2.

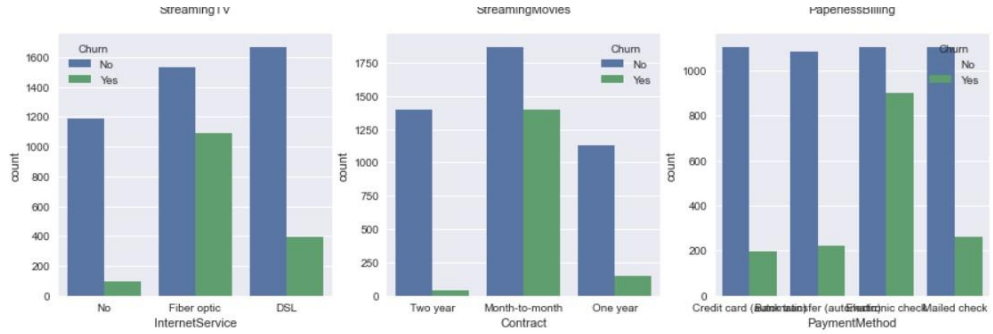


Male users and female users have same likelihood of renewing a contract or not renewing a contract.

3. Plotted other columns with churn to see any insights.







Key Observations:

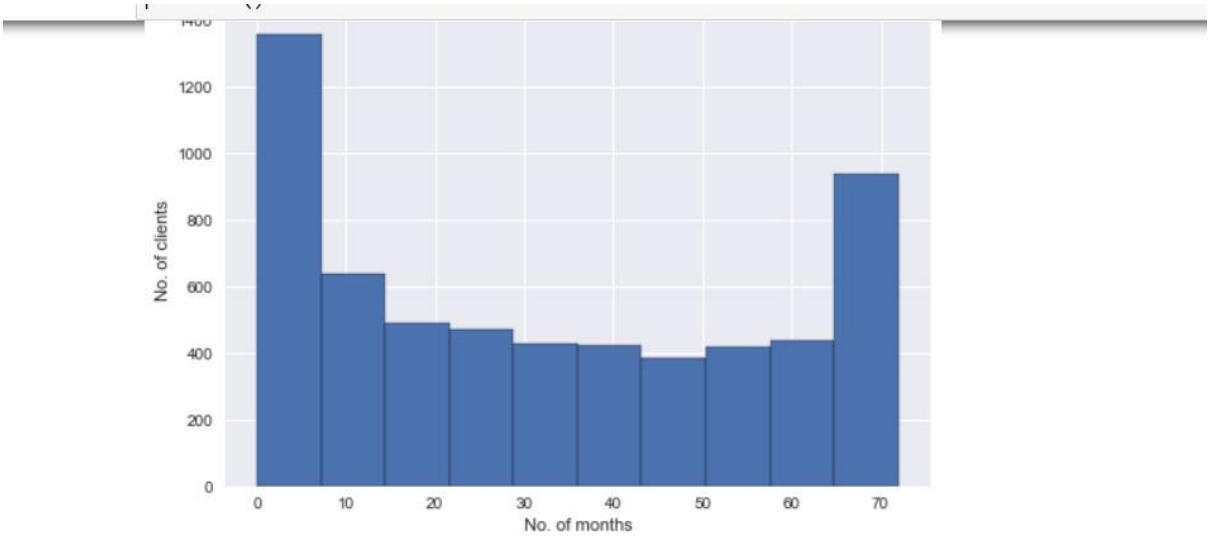
Customers which are using month to month contract have high churn as compared to other contract.

Customers which doesn't use tech support, online security, Device protection have high churn.

Customers who doesn't have dependents are also having high churn.

4.

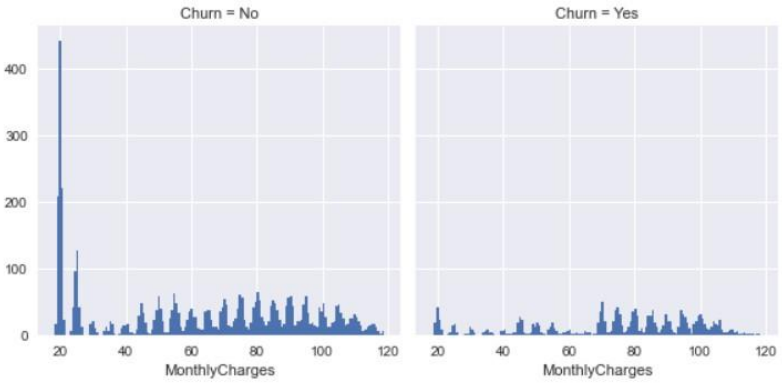
Now let’s visualise continuous variables.



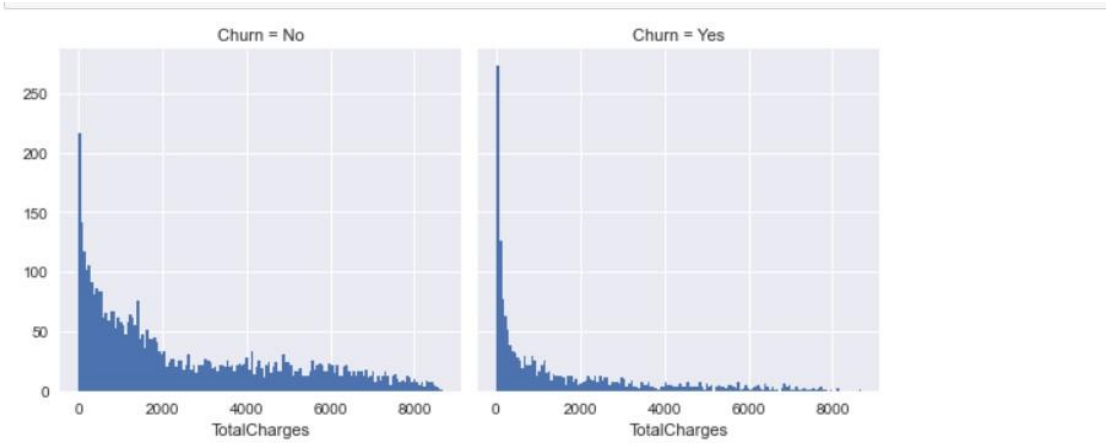
Average months a person has been customer of the company is 32.5 months (2.7 years).

From the graph it can observed that there is a sharp decrease in number of customers after 7 months.

5.

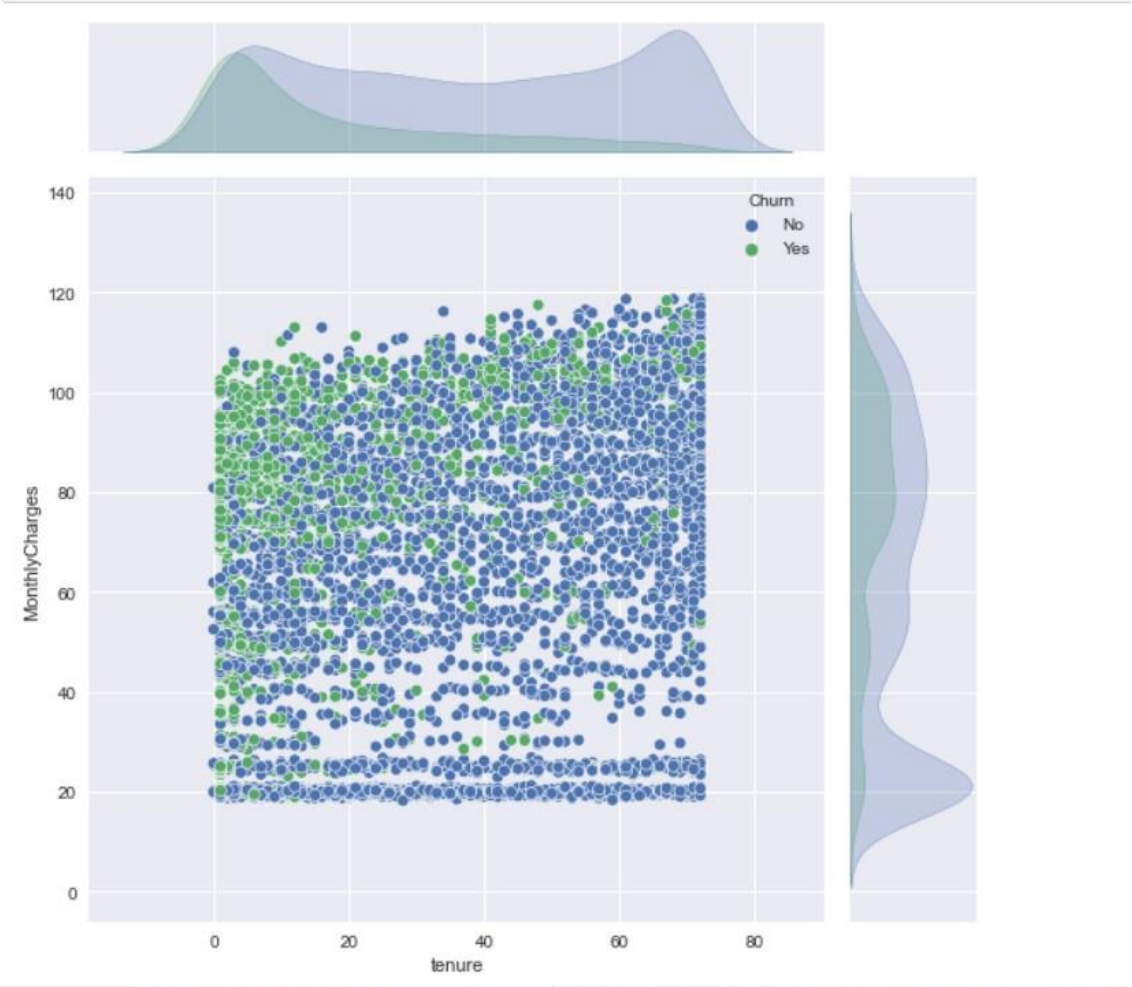


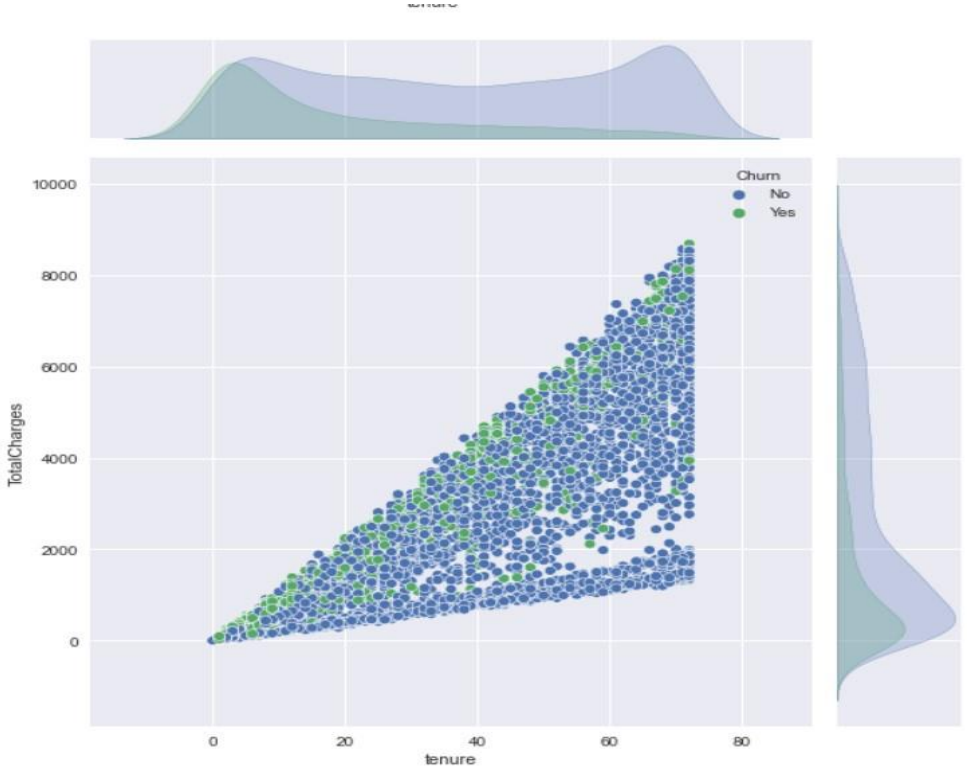
The resulting graphs show a rise in churn among customers whose monthly charges are more than $70.



Maximum customers leave company services before the total amount of payment of entire time surpasses $1000.

6.




Churn
No
Yes
TotalCharges
tenure
0
2000
4000
6000
8000
10000
0
20
40
60
80


Churn
No
Yes
MonthlyCharges
TotalCharges
0
20
40
60
80
100
120
140
0
2000
4000
6000
8000
10000

- A major aspect is unquestionably tenure; customers tend to be more likely to renew their contracts if they've been using the service for a while, regardless of how much they're paying or have paid in total.
- People are far more likely to retain their telecom contract if Total Charges or Monthly Charges are low.
- It's worth noting here that even though monthly charges are higher, people with a long tenure will most likely extend their contract and have low churn.

Implementation of Different Machine Learning Models

Target variable(churn) was extracted from the dataframe and when counted got to know that data is heavily imbalanced.

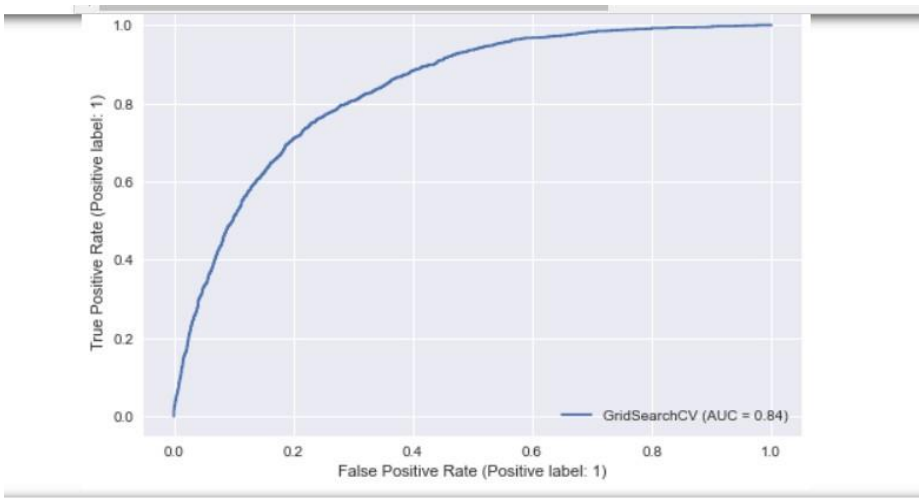
So to rectify this oversampling is done using SMOTE(Synthetic Minority Oversampling Technique).

Here 3 models logistic regression, KNN and SVM were trained using Nested Cross Validation.

Nested cross validation is used for hyperparameter tuning and model selection that attempts to overcome the problem of overfitting the dataset. Two k-fold cross validation loops are used, kfold cross validation hyperparameter tuning loop is nested inside the k fold cross validation loop for model selection.

Standard scaler is used in the pipeline to scale the data.

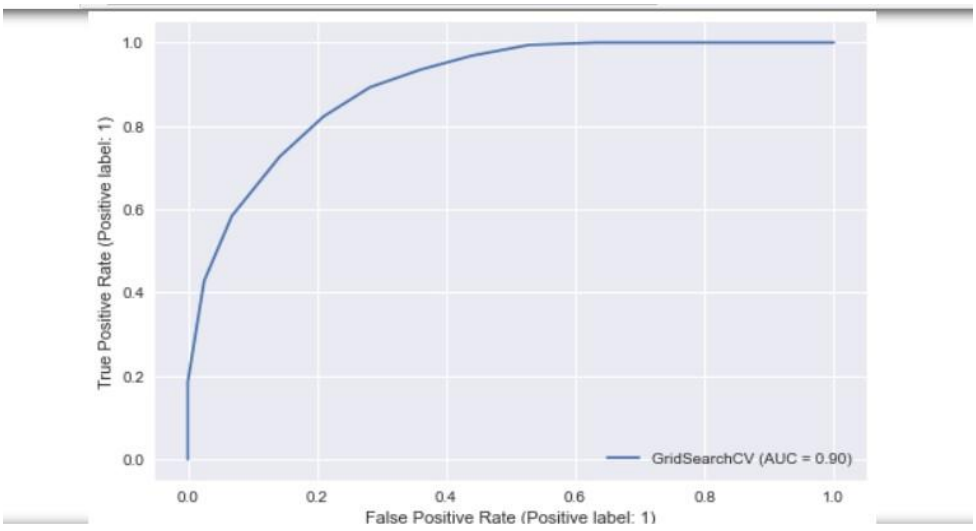
ROC curve and other Statistics for SVC :



```
svm
Accuracy in the outer folds: ['0.78', '0.76', '0.76', '0.77', '0.76'].
Average Accuracy: 0.76
```

	precision	recall	f1-score	support
0	0.89	0.78	0.83	4399
1	0.54	0.73	0.62	1587
accuracy			0.76	5986
macro avg	0.72	0.75	0.73	5986
weighted avg	0.80	0.76	0.77	5986

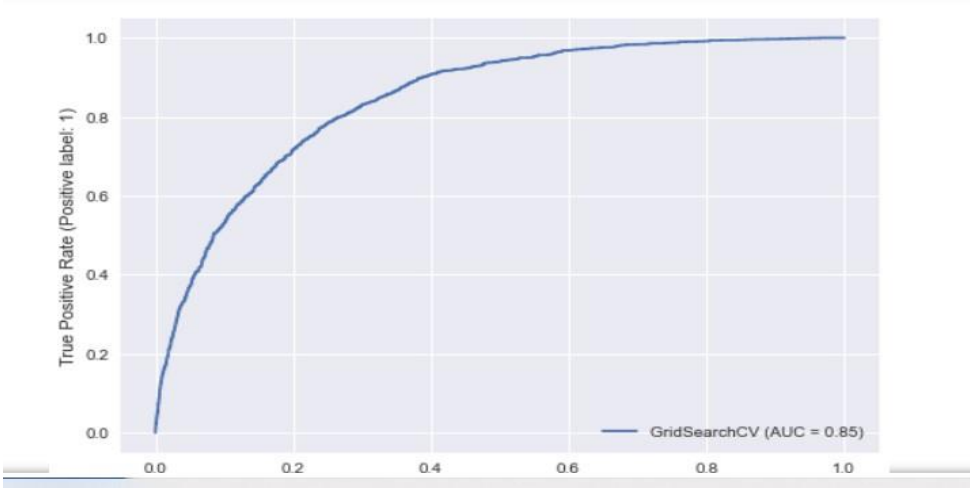
ROC curve and other statistics for KNN:




```
knn
Accuracy in the outer folds: ['0.74', '0.73', '0.73', '0.73', '0.74'].
Average Accuracy: 0.73
```

	precision	recall	f1-score	support
0	0.88	0.74	0.80	4399
1	0.50	0.71	0.59	1587
accuracy			0.73	5986
macro avg	0.69	0.73	0.70	5986
weighted avg	0.78	0.73	0.75	5986

ROC curve and other statistics for Logistic regression:




```
reg
Accuracy in the outer folds: ['0.75', '0.74', '0.74', '0.75', '0.74'].
Average Accuracy: 0.74
```

	precision	recall	f1-score	support
0	0.91	0.73	0.81	4399
1	0.51	0.80	0.62	1587
accuracy			0.74	5986
macro avg	0.71	0.76	0.72	5986
weighted avg	0.80	0.74	0.76	5986

The best model for predicting churn is SVM with average accuracy of 78.3%.

SVM also had highest precision for class 1 as compared to other models. That class represents customers who opted out from services. This is the most important metric to evaluate in this type of problem.

SVM also have a very descent AUC(Area Under Curve) score.

Is Trained Model Fair?

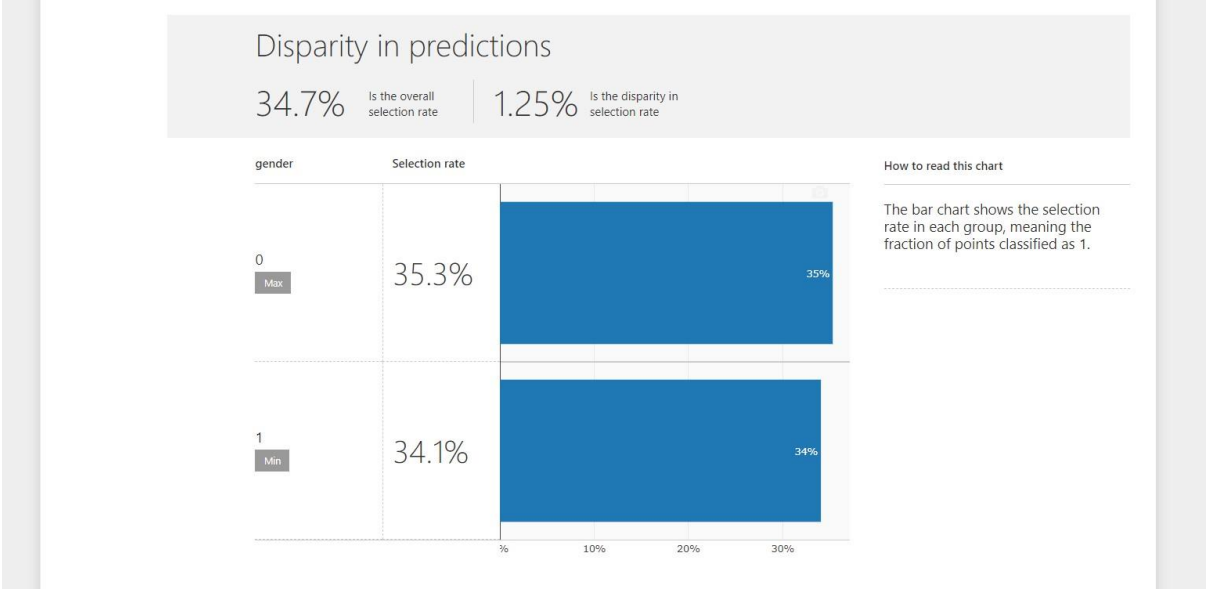
To check whether are trained model is Fair here I used python package Microsoft Fairlearn.

The fairlearn package contains a component called FairlearnDashboard. It's a widget that we can use within a Python notebook which measures and visualizes two metrics for our model:

Metric disparity: a measure that we can use to see what the difference in performance is between different groups of users for our model.

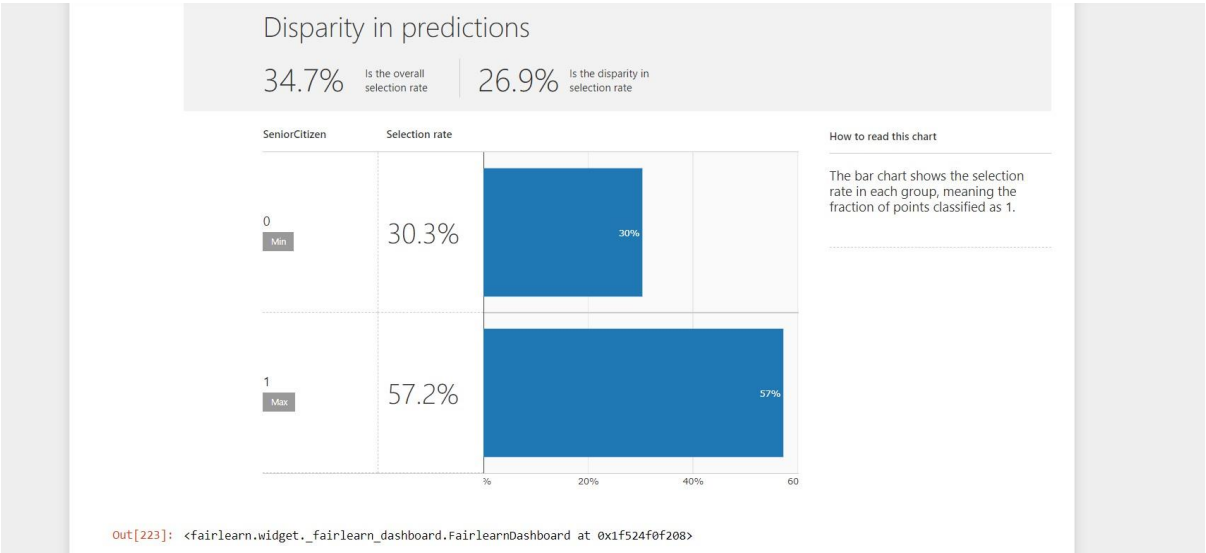
Prediction disparity: a measure that shows the difference of true positives between groups for binary classification models. For regression models it shows the difference in distribution of predicted outputs for different groups. (mediumfairlearn)

1. First checked model for sensitive feature i.e. gender.



Model was almost fair in case of gender.

2. Now let's see for sensitive feature SeniorCitizen.

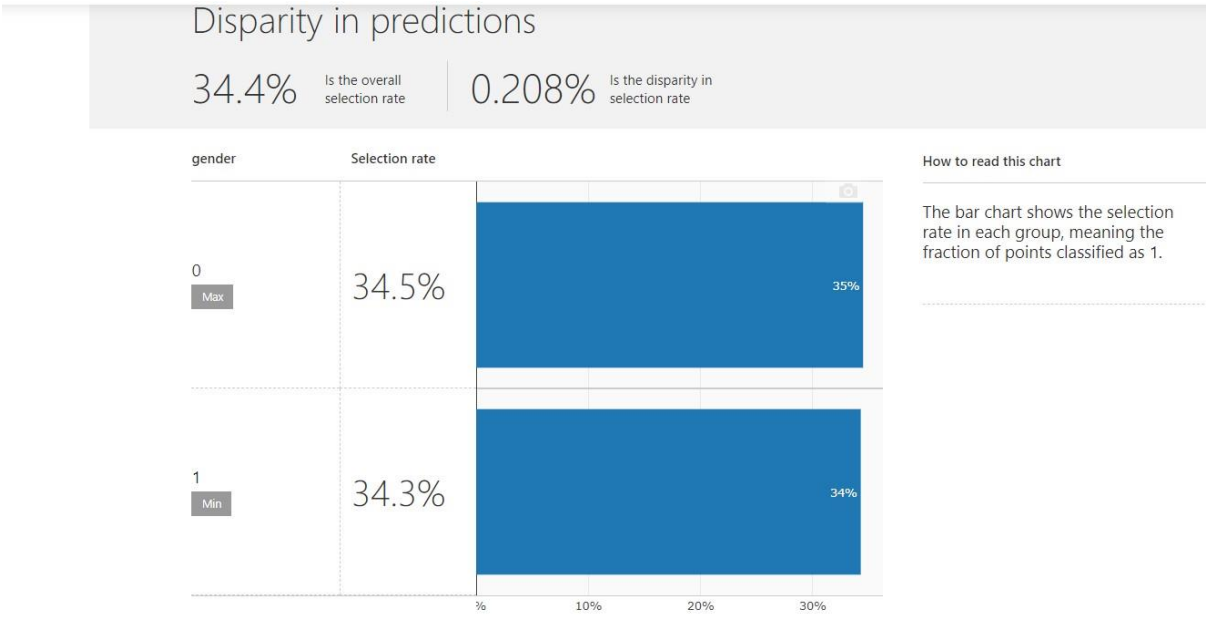


There is 26.9% disparity in selection rate. It means if a client is SeniorCitizen, model is more likely to predict that client will churn.

To improve fairness we will mitigate unfairness using threshold optimizer. Threshold optimizer is used to improve a existing model.

The TresholdOptimizer is based on a paper called “Equality of Opportunity in Supervised learning”. It tries to correct the model so that it no longer discriminates against specific groups of users, based on a set of sensitive features.(mediumfairlearn)

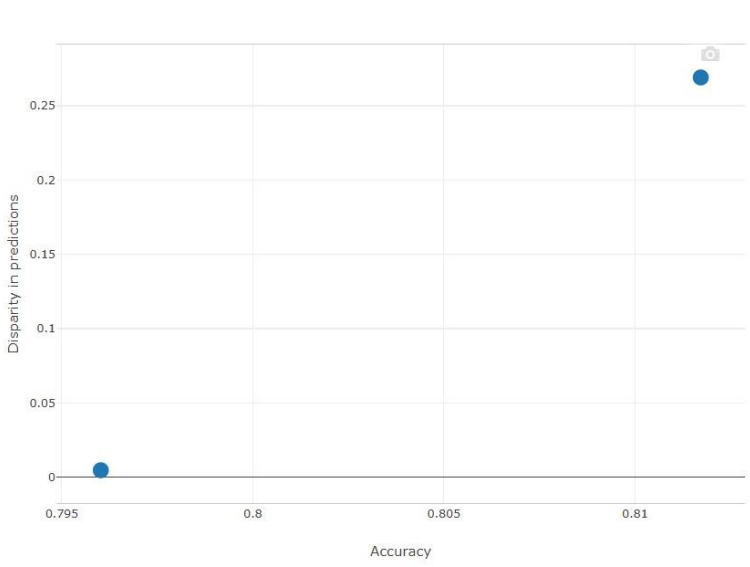
Using optimizer created a new model and let’s compare the new model with old model.




```
Out[34]: <fairlearn.widget._fairlearn_dashboard.FairlearnDashboard at 0x28483081c08>
```

We can see disparity in selection rate is reduced drastically.

3. Plotting graph of new model and old model.



How to read this chart

This chart represents each of the 2 models as a selectable point. The x-axis represents accuracy, with higher being better. The y-axis represents disparity, with lower being better.

INSIGHTS

Accuracy ranges from 79.6% to 81.2%. The disparity ranges from 0.466% to 26.9%.

The most accurate model achieves accuracy of 81.2% and a disparity of 26.9%.

The lowest-disparity model achieves accuracy of 79.6% and a disparity of 0.466%.

Interpretation:

Here didn't tried removing features before feeding to the model. Could have done comparative study of couple of sampling methods(to rectify problem of imbalanced data). Here used nested cross validation which drastically increases processing time. In place of GridSearchCV we can also use RandomizedSearchCV which takes less running time.