ICPS GROUP 1 PROJECT

END EVALUATION REPORT FOR GROUP 1

NOVEMBER 15, 2023

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
SRICITY, ANDHRA PRADESH

TOPIC:

Telecom Industry Churn Analysis and Predictive modelling.

ABSTRACT:

Enhancing telecom industry customer retention through predictive churn analysis and modelling.

TEAM MEMBERS:

GROUP - 1

1) Utkarsh Vaish (Lead) (S20200020309)
2) Jeswin Sam (S20200020266)
3) Avinash Saroj (S20200010027)
4) Warish (S20200010232)

DELIVERABLE 1:

Data preprocessing, feature extraction, data visualization and drawing some insights out of the data.

DELIVERABLE 2:

Model Building, predicting churn and building interactive and insightful dashboard, Deploy the model using Flask.

DATASET:

- Telecom Industry Churn Dataset
- Shape – (7043, 21)

ATTRIBUTES:

customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges and Churn

DATA CLEANING/ PREPROCESSING

- No Null Values in Dataset
- No outliers
- No duplicate entries
- datatype of TotalCharge changed to float from string
- TotalCharge contained 11 missing values so those rows dropped

OUTLIER ANALYSIS:

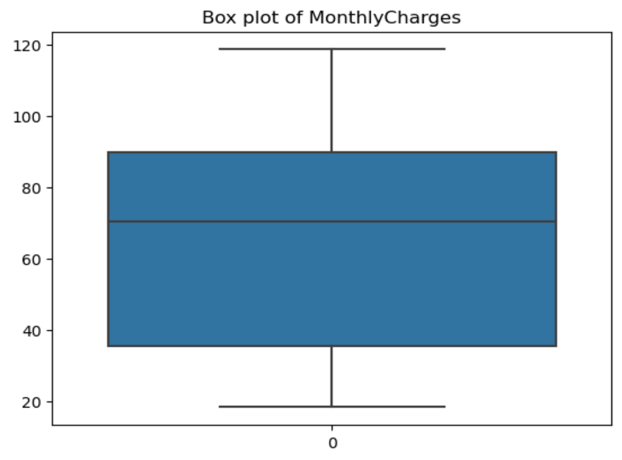
Q1 - 25th percentile; Q2 - 50th percentile; Q3 – 75th percentile

Upper Bound (top whisker) = Q3 + (1.5 x IQR)

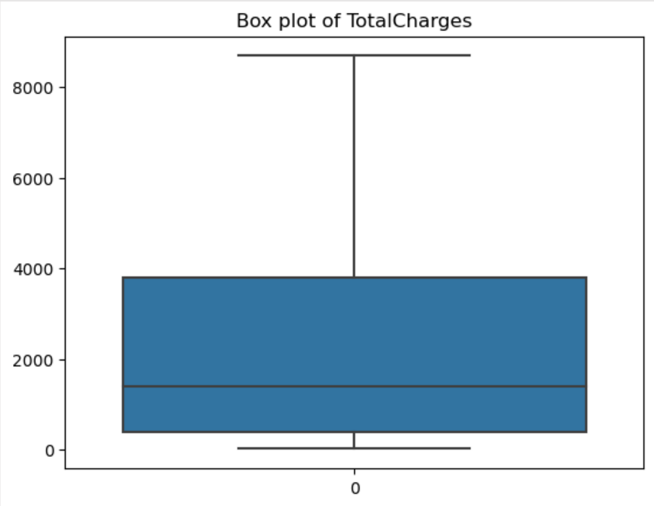
Lower Bound (bottom whisker) = Q1 – (1.5 x IQR)

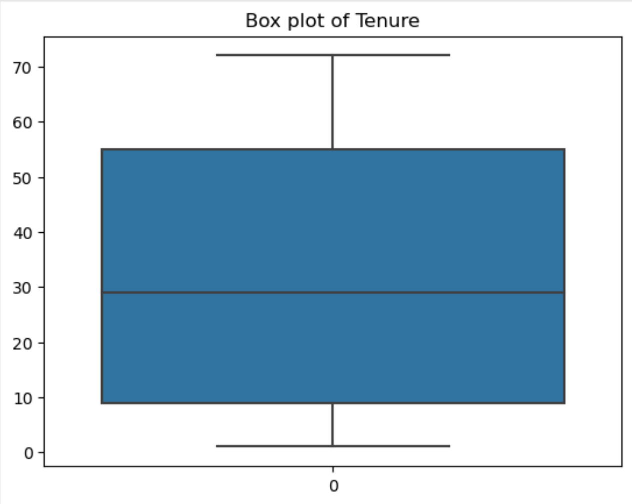
$$\boxed{IQR = Q3 - Q1}$$

Where IQR – Interquartile Range



1. **No outliers as visible from the plot**
2. **50 percentile pays 35-90 RS approx.**
3. **25 percentile pays 90-120 RS approx.**





1. **No outliers as visible from the plot**
2. **50 percentile pays between 500 and 4000 RS approx.**
3. **25 percentile pays 4000 – 8500**

1. **No outliers as visible from the plot**
2. **50 percentile has tenure from 10-55 months appx.**
3. **25 percentile has tenure from 55-70 months appx.**

Inference from table below:

- **75 percentile Customers have tenure of 55 months or lower, pays 89.86 RS or lower per month and have total charges of 3794.73 RS or lower.**
- **Mean tenure, Monthly charges and Total Charges are 32.42 months, 64.79 RS, 2283.30 RS.**

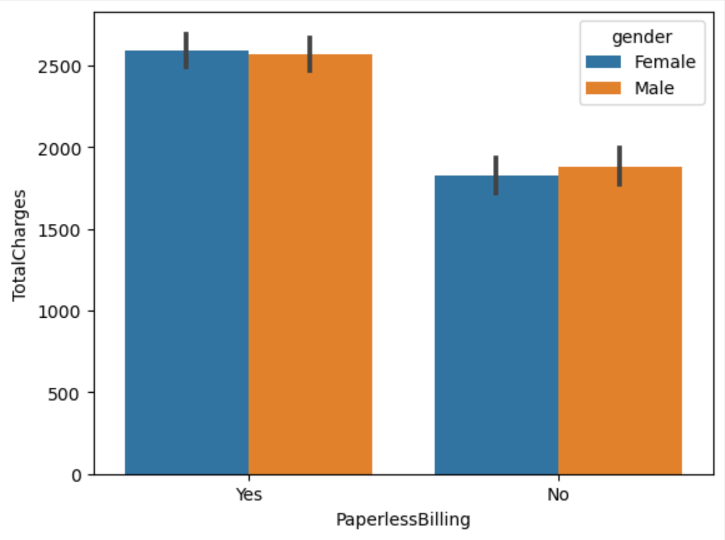
STATISTICS:

Stat	Tenure	Monthly Charges	Total Charges
Count	7032	7032	7032
Mean	32.422	64.798	2283.3
Std Dev	24.545	30.085	2266.771
Min	1	18.25	18.8
25 percentile	9	35.587	401.45
50 percentile	29	70.35	1397.475
75 percentile	55	89.862	3794.737
Max	72	118.75	8684.8

FEATURE EXTRACTION:

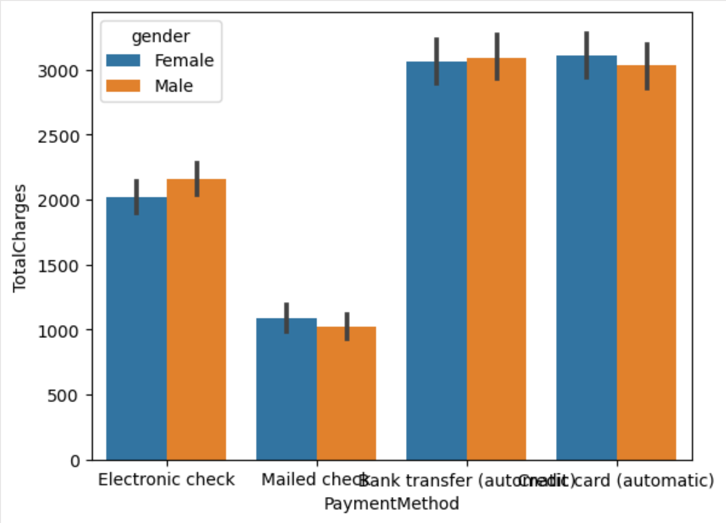
- The churn risk is assessed with the categorization of entries into 7 separate categories namely- High Churn Risk, Moderate Churn Risk, Low Churn Risk, No Churn, Premium Customers, Occasional Churn Risk, Seasonal Churn Risk.
- New category ChurnRisk is added and Churn removed
- The churn risk assessment is made easier by mapping the risk categories to numerical scale (Magnitude of 0-6)

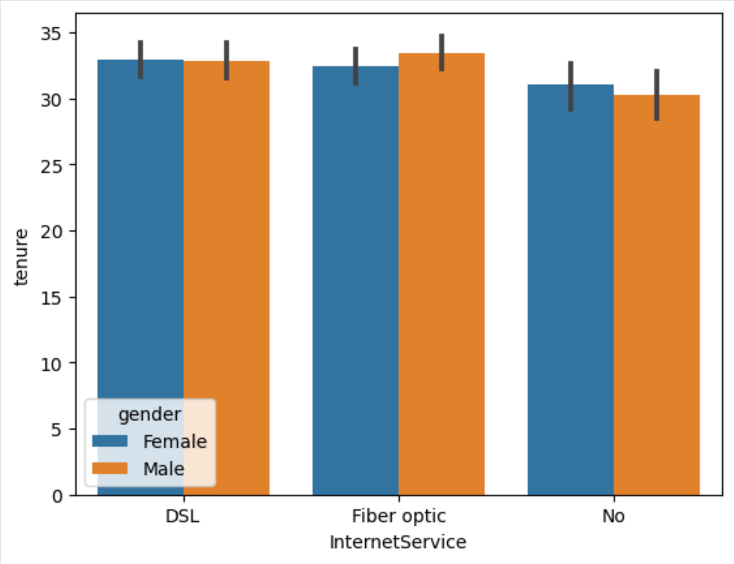
DATA VISUALIZATION AND INSIGHTS



- **There is no gender disparity in preference of paperless billing**
- **Those with higher total charges prefer paperless billing.**

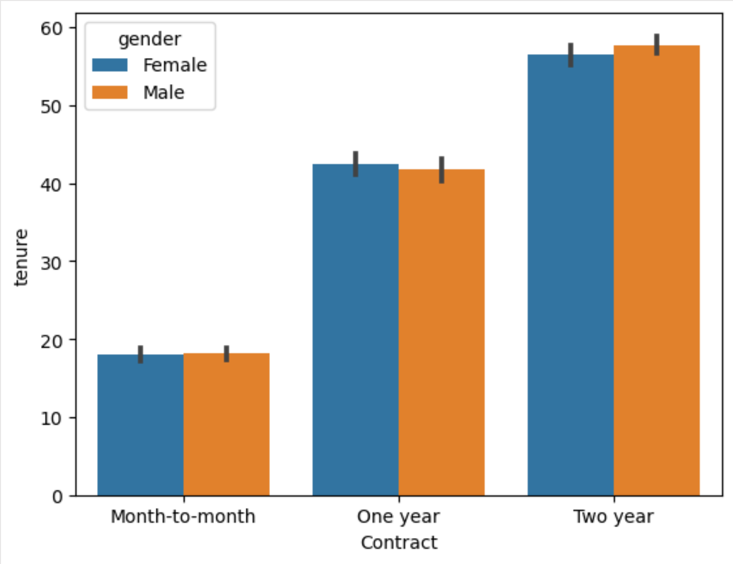
- **Majority of the Charges are collected from Bank transfer and Credit card payment methods.**

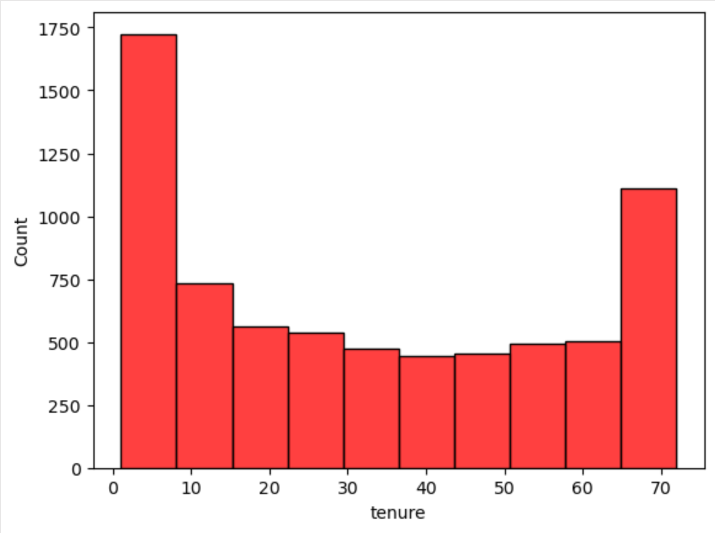




- **Men are slightly oversubscribed to the Fiber optic service as compared to women.**

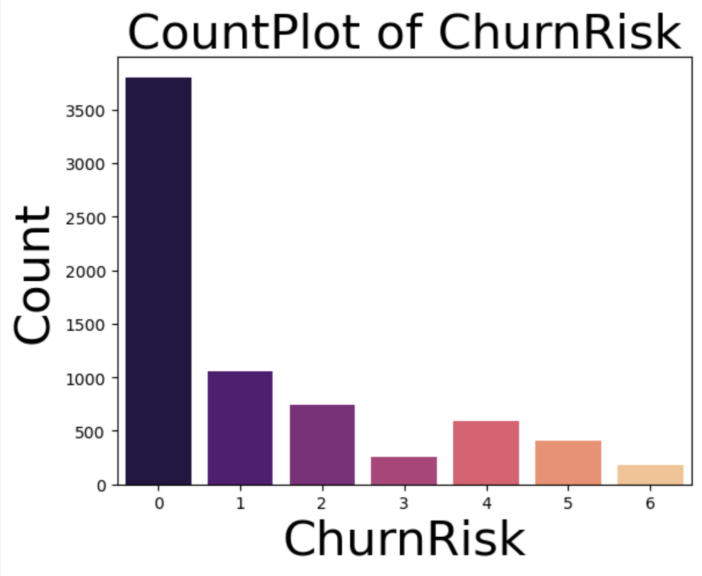
- **More people prefer the two-year contract over one year and month-to-month contracts, hinting about the considerable number of customers with no churn.**

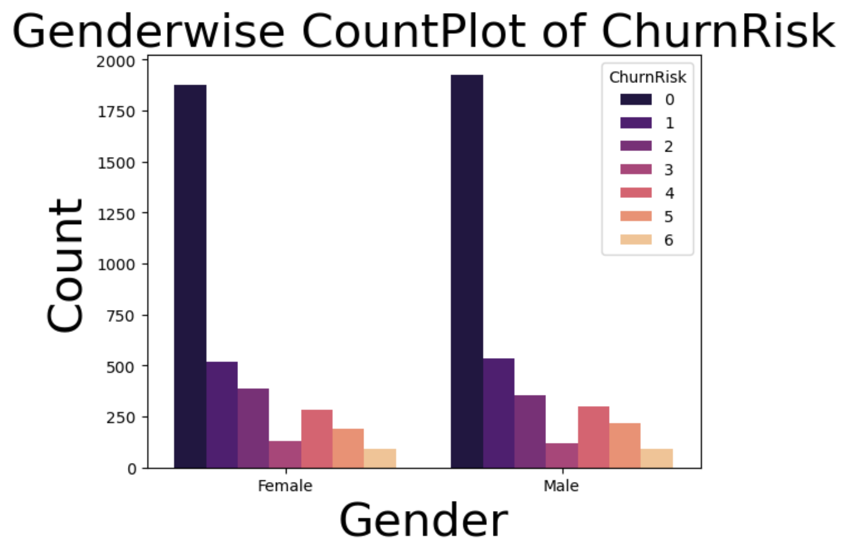




- **Majority of the customers, around 1750, has the tenure of (0-10) months**

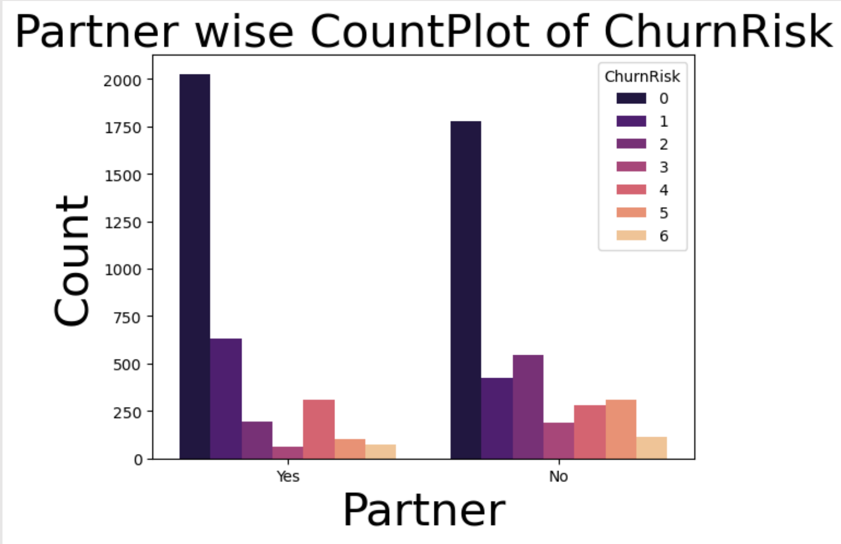
- **Churn risk is considerably lower than no churn as depicted by the bar graph.**

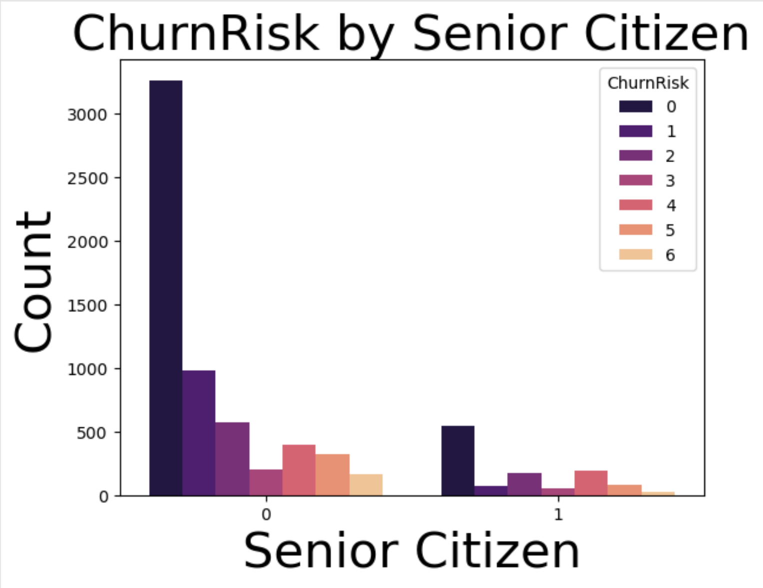




- **Comparing the two bar graphs tells us that the churn pattern is almost similar regardless of gender.**

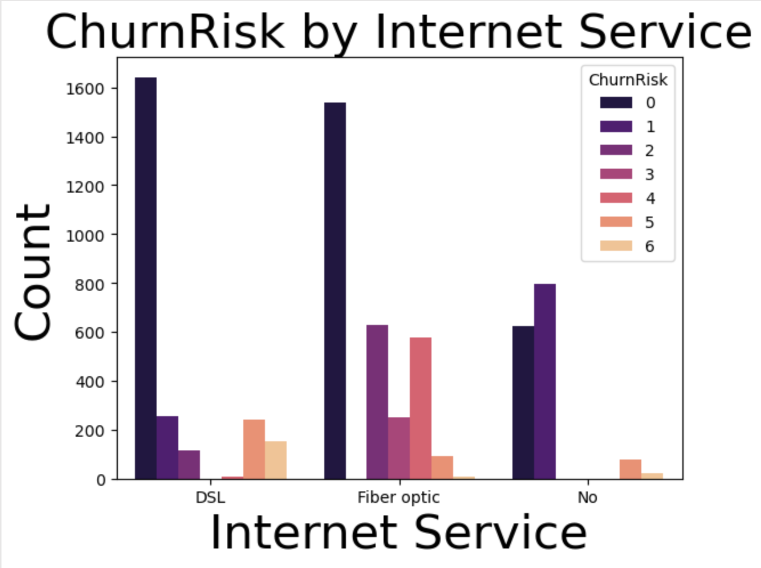
- **There is considerable number of singles who fall into High-risk churn category (singles).**

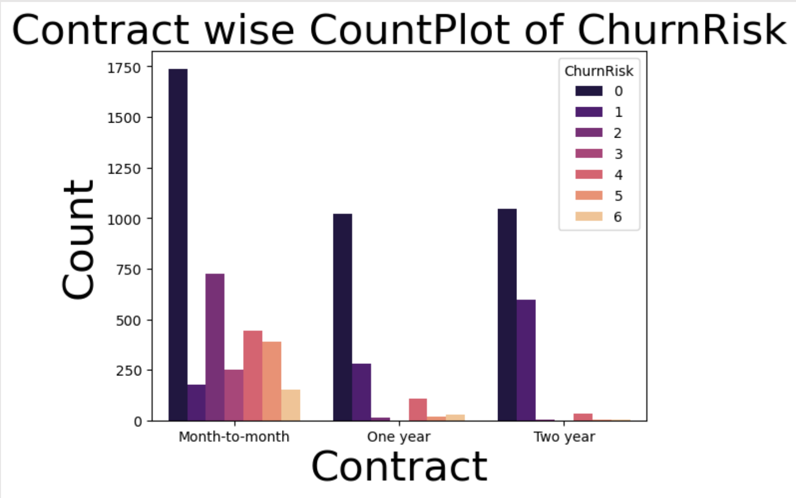




- **Churn is low in senior citizens as compare to youth.**
- **Majority of the customers are youth.**

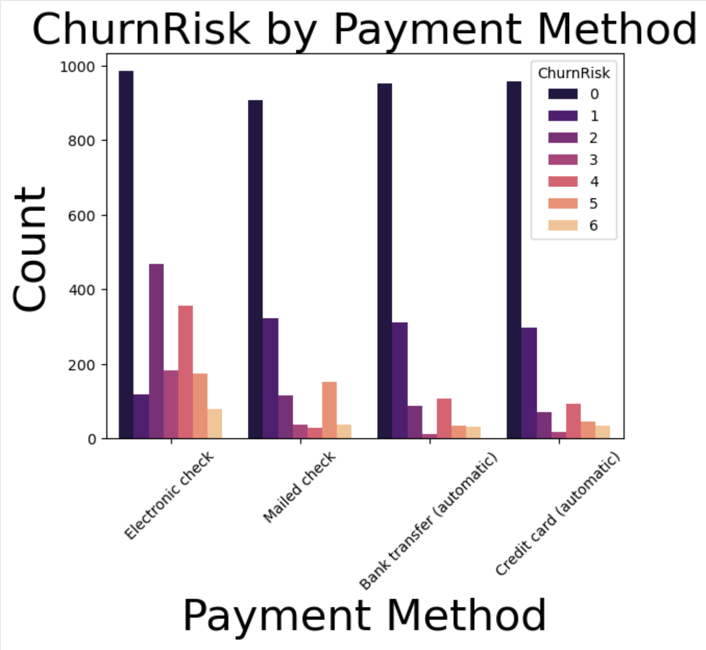
- Less churn in case of customers who do not have internet connection.
- Less churn in case of customers who have DSL connection.
- High churn risk is considerable in people using Fiber optic.

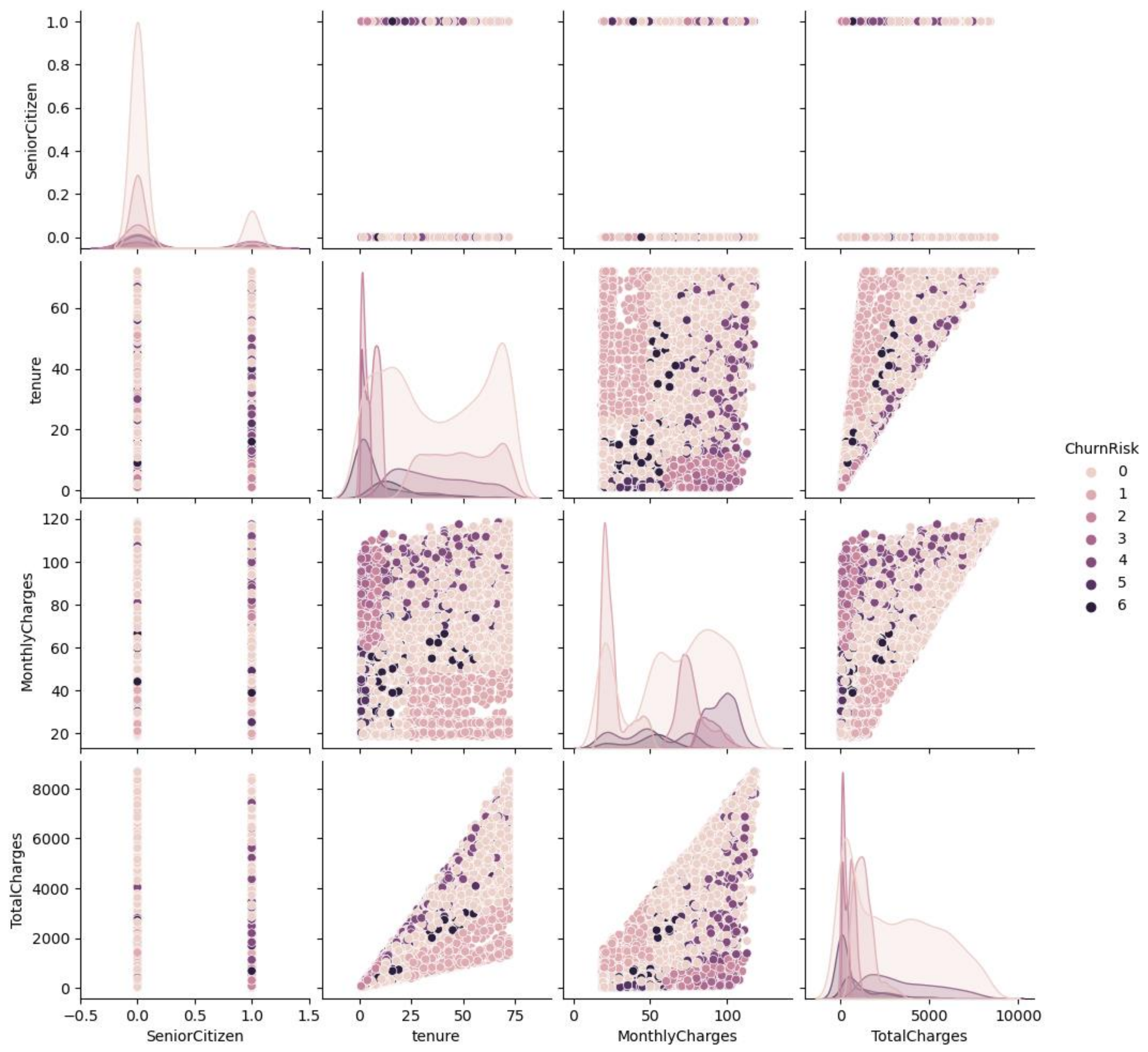




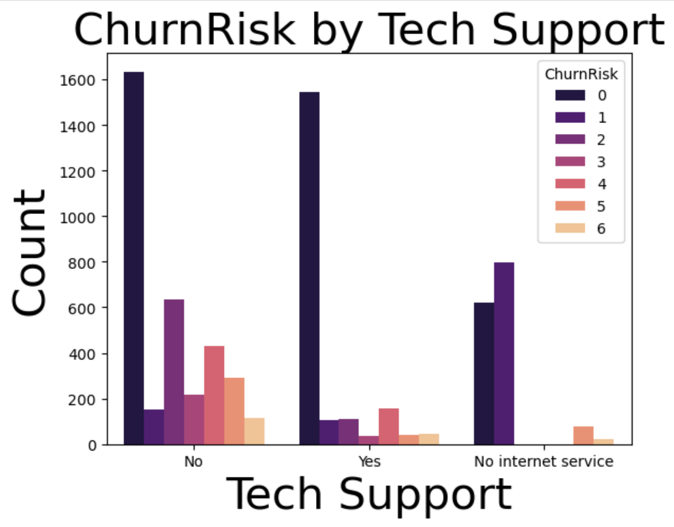
- High risk churn is significant in month-to-month contract category as compared to one- or two-year contracts.

- **High risk of churn in electronic check payment method as compared to other methods.**

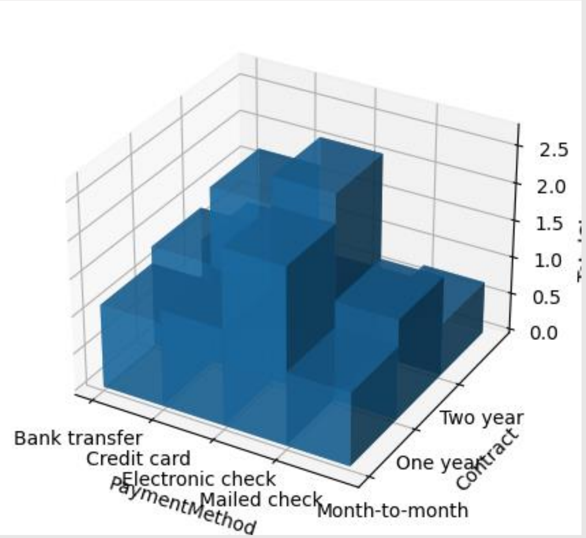




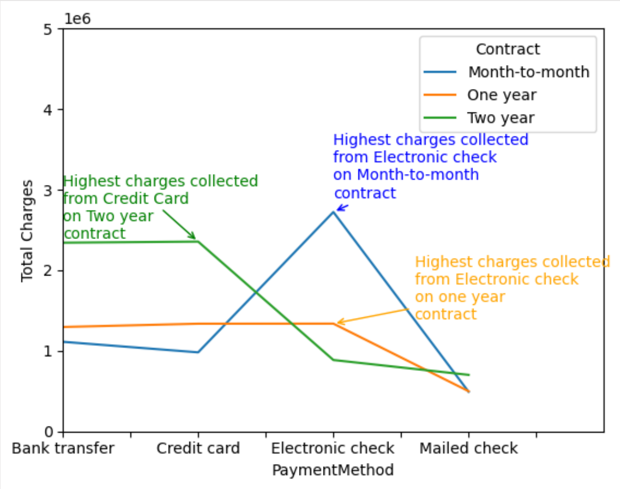
- **Churn is high at low tenure and decreases as tenure increases.**
- **Churn decreases with increase in total charges.**
- **Churn is low in senior citizens as compare to youth.**



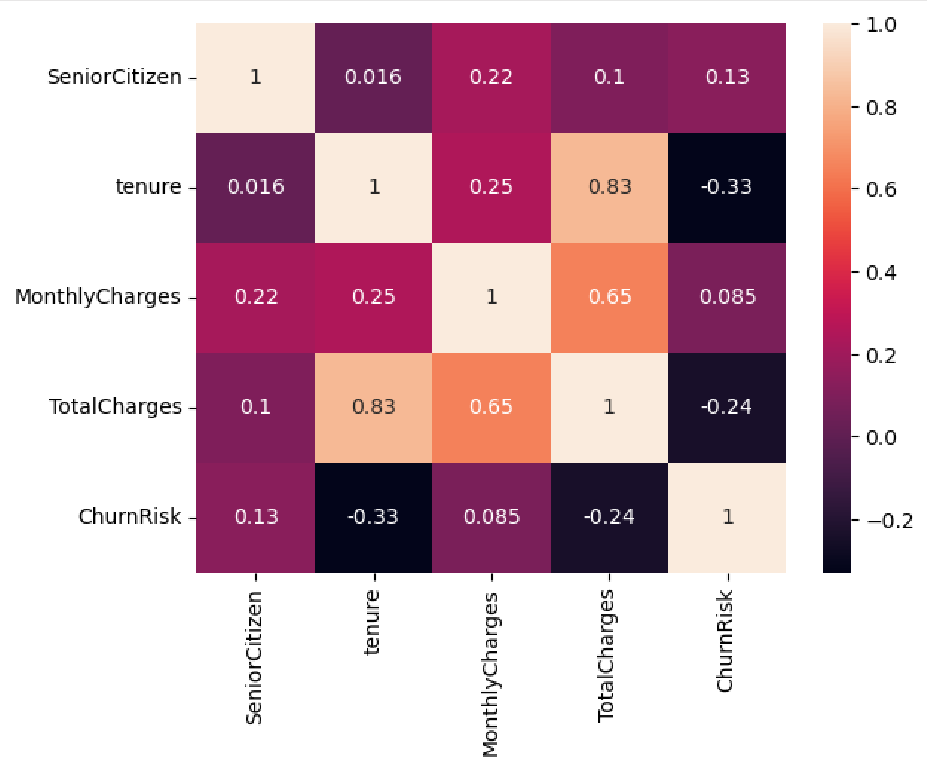
- There is high churn risk in no tech support category.

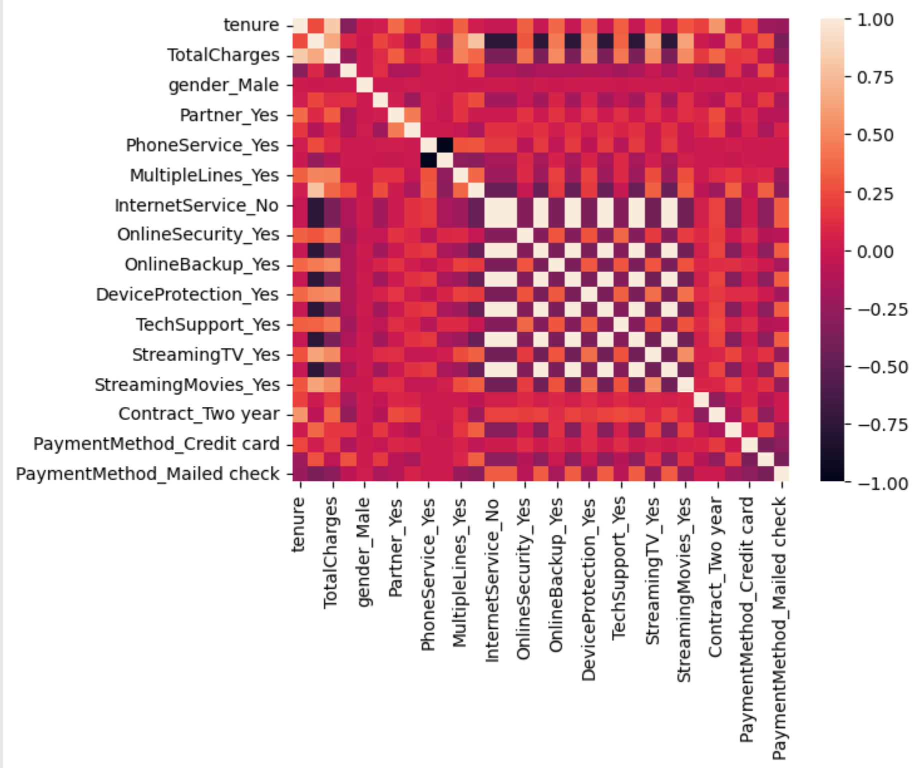


- In order to attract more customers Company should introduce some schemes in:
- Electronic check payment method for month to month and one year contract customers
- Credit card payment method for two years contract customers.



CORRELATION ANALYSIS:





- TotalCharges and Tenure has a good correlation of 0.83
- ChurnRisk has negative correlation with Tenure and TotalCharges

DATA TRANSFORMATION:

- Min-Max Normalization and Log transformation is applied but it does not bring any significant changes in the accuracy of prediction models.

ONE HOT ENCODING:

- Converting entries of categorical attributes into numerical values in order to feed data to model.

DATA BALANCING:

- Random Over Sampling is done to balance data so that model does not become biased towards one class.

```
Original Dataset shapeCounter({0: 4415, 5: 827, 1: 685, 6: 484, 4: 356, 2: 202, 3: 63})
Resampled datset shapeCounter({0: 4415, 5: 4415, 2: 4415, 4: 4415, 1: 4415, 6: 4415, 3: 4415})
```

MODELLING:

- Feature Selection:
  Target Feature: ChurnRisk
  Irrelevant Feature: Customer Id

K-Nearest Neighbours (KNN):

```
              precision    recall  f1-score   support

           0       0.96      0.54      0.69      1416
           1       0.97      1.00      0.98      1324
           2       0.99      1.00      1.00      1351
           3       1.00      1.00      1.00      1324
           4       0.90      1.00      0.95      1259
           5       0.86      0.98      0.92      1285
           6       0.84      0.99      0.91      1313

    accuracy                           0.93      9272
   macro avg       0.93      0.93      0.92      9272
weighted avg       0.93      0.93      0.92      9272

Confusion Matrix:
[[ 766   45    5    0  133  210  257]
 [   2 1322    0    0    0    0    0]
 [   0    0 1351    0    0    0    0]
 [   0    0    0 1324    0    0    0]
 [   0    0    0    0 1259    0    0]
 [  20    0    2    0    0 1263    0]
 [   7    0    0    0    0    0 1306]]


Accuracy Score:  92.65530629853322
```

Decision Tree Classifier:

```
              precision    recall  f1-score   support

           0       0.99      0.76      0.86      1416
           1       1.00      1.00      1.00      1324
           2       1.00      1.00      1.00      1351
           3       1.00      1.00      1.00      1324
           4       0.94      1.00      0.97      1259
           5       0.91      0.99      0.95      1285
           6       0.90      1.00      0.95      1313

    accuracy                           0.96      9272
   macro avg       0.96      0.96      0.96      9272
weighted avg       0.96      0.96      0.96      9272

Confusion Matrix:
[[1071    0    0    0   78  129  138]
 [   0 1324    0    0    0    0    0]
 [   0    0 1351    0    0    0    0]
 [   0    0    0 1324    0    0    0]
 [   0    0    0    0 1259    0    0]
 [  12    0    0    0    0 1273    0]
 [   4    0    0    0    0    0 1309]]


Accuracy Score:  96.10655737704919
```

Random forest Classifier:

```
              precision    recall  f1-score   support

           0       0.99      0.77      0.87      1416
           1       1.00      1.00      1.00      1324
           2       1.00      1.00      1.00      1351
           3       1.00      1.00      1.00      1324
           4       0.94      1.00      0.97      1259
           5       0.90      1.00      0.95      1285
           6       0.93      1.00      0.96      1313

    accuracy                           0.96      9272
   macro avg       0.97      0.97      0.96      9272
weighted avg       0.97      0.96      0.96      9272

Confusion Matrix:
[[1095    0    1    0   80  141   99]
 [   0 1324    0    0    0    0    0]
 [   0    0 1351    0    0    0    0]
 [   0    0    0 1324    0    0    0]
 [   0    0    0    0 1259    0    0]
 [   2    0    0    0    0 1283    0]
 [   4    0    0    0    0    0 1309]]


Accuracy Score:  96.4732528041415
```

Logistic Regression:

```
              precision    recall  f1-score   support

           0       0.52      0.24      0.33      1416
           1       0.96      0.99      0.98      1324
           2       0.68      0.99      0.81      1351
           3       0.49      0.94      0.64      1324
           4       0.77      0.84      0.81      1259
           5       0.00      0.00      0.00      1285
           6       0.60      0.62      0.61      1313

    accuracy                           0.66      9272
   macro avg       0.58      0.66      0.60      9272
weighted avg       0.58      0.66      0.60      9272

Confusion Matrix:
[[ 345   43   84  193  276    1  474]
 [   0 1310    0    0    0    0   14]
 [   0    0 1334   17    0    0    0]
 [   0    0   78 1246    0    0    0]
 [  48    0  140    0 1061    0   10]
 [  50    0   63 1074   38    0   60]
 [ 217    5  259   12    0    0  820]]


Accuracy Score:  65.96203623813632
```

Support Vector Machine:

```
              precision    recall  f1-score   support

           0       0.44      0.01      0.01      1416
           1       0.87      1.00      0.93      1324
           2       0.69      0.95      0.80      1351
           3       0.66      0.73      0.70      1324
           4       0.50      0.76      0.61      1259
           5       0.57      0.64      0.60      1285
           6       0.43      0.36      0.39      1313

    accuracy                           0.63      9272
   macro avg       0.60      0.63      0.58      9272
weighted avg       0.60      0.63      0.57      9272

Confusion Matrix:
[[   8  105   88   68  616  253  278]
 [   2 1318    0    0    0    0    4]
 [   0    0 1280   71    0    0    0]
 [   0    0   82  970    0  272    0]
 [   0    0   54    0  955    0  250]
 [   0    0   26  265   82  826   86]
 [   8   93  313   90  242  100  467]]


Accuracy Score:  62.81276962899051
```

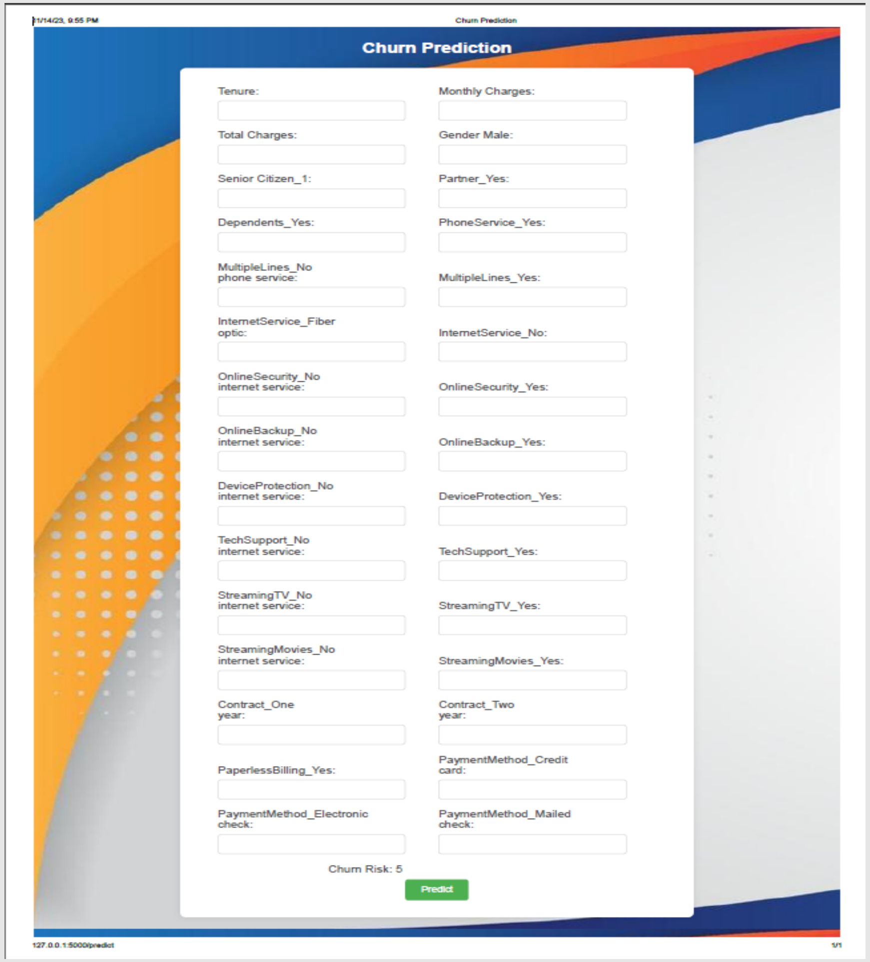
Naive Bayes:

```
              precision    recall  f1-score   support

           0       0.78      0.19      0.30      1416
           1       0.75      1.00      0.85      1324
           2       0.61      0.90      0.73      1351
           3       0.89      0.99      0.93      1324
           4       0.62      0.96      0.75      1259
           5       0.59      0.28      0.38      1285
           6       0.58      0.50      0.53      1313

    accuracy                           0.68      9272
   macro avg       0.69      0.69      0.64      9272
weighted avg       0.69      0.68      0.64      9272

Confusion Matrix:
[[ 265  238   75    6  399  113  320]
 [   0 1318    0    0    0    0    6]
 [   5    0 1216   52   71    0    7]
 [   0    0    0 1310   14    0    0]
 [   0    0   51    0 1208    0    0]
 [   3  137  382  112  148  363  140]
 [  66   76  262    0  115  144  650]]


Accuracy Score:  68.27006039689387
```

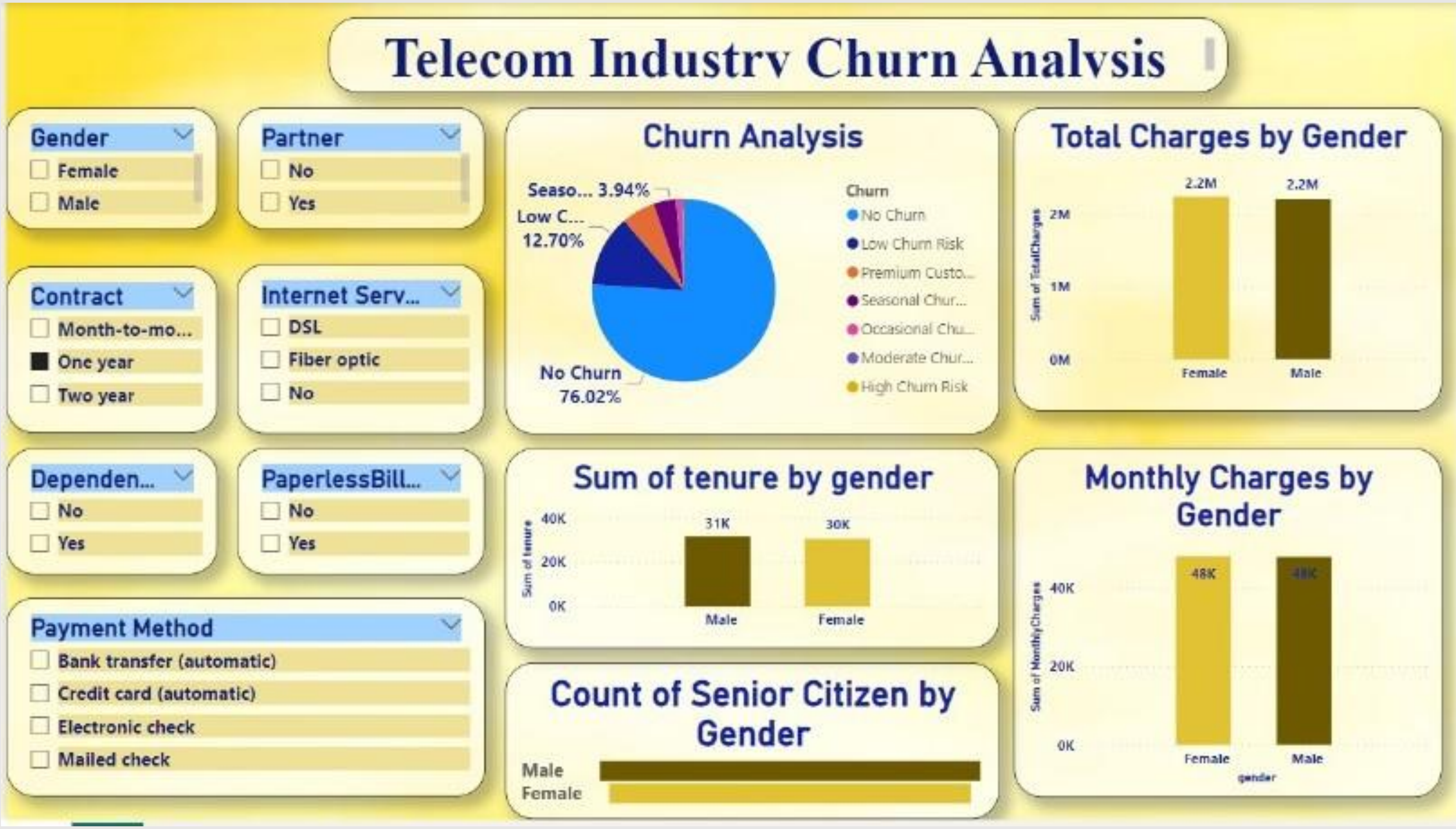
- Both the Decision Tree as well as Random Forest classifier is giving approximately same accuracy score but due to Random Forest being an ensemble technique and a more computationally expensive algorithm, we are going with **Decision Tree (96.11%)** to finally train our model.

MODEL DEPLOYMENT:

- Saving the model using Pickle
- Deploying the model using Flask.



Power BI Dashboard:



CONCLUSION:

- This analysis can be helpful in increasing revenue of the company by attracting new customers (by seeing trends) and avoiding contract terminations (ChurnRisk).

THANK YOU