Strategic Thinking – CA2

Semester 2

CCT College Dublin

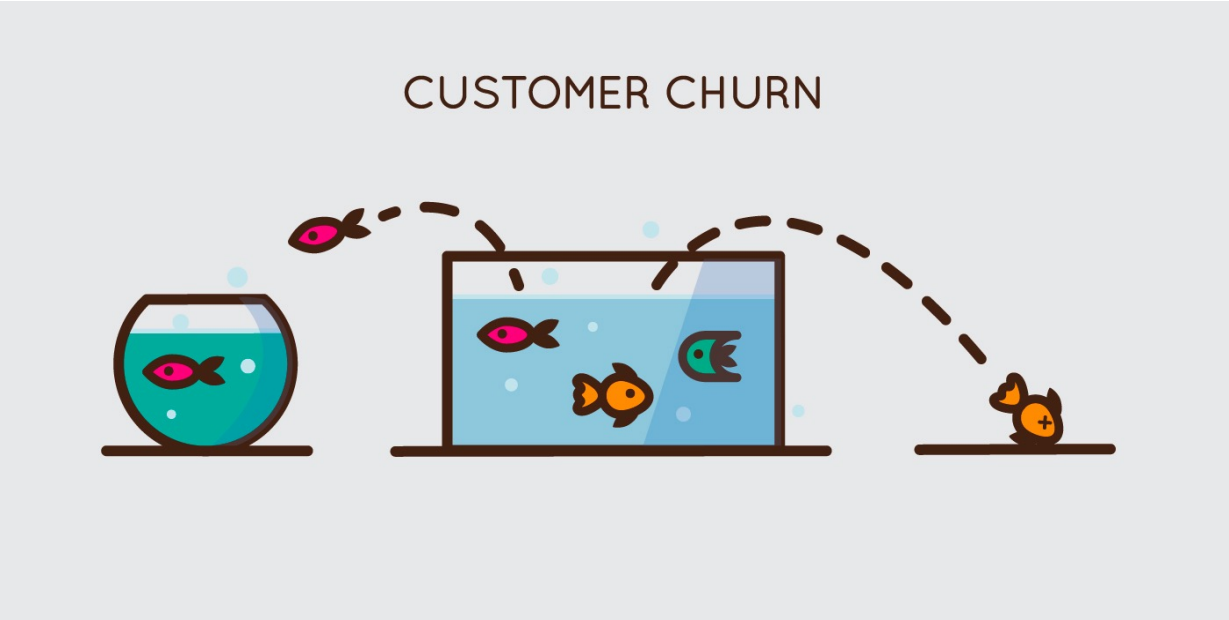
Predicting Churn in Telecommunication

Arthur Claudino Gomes de Assis (2023146)

Business understanding and data visualization

Churn is a common problem in the telecommunications business and refers to customers who cancel or do not renew their contract with a telecommunications company in a given period. Churn is a very important indicator for telecommunications companies since it is much more expensive to attract new customers than to retain existing ones.

- Price
- Product/Market Fit
- User Experience
- Customer Experience
- Other Causes



The dataset

```
1 df_churn.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   customerID        7043 non-null   object 
 1   gender            7043 non-null   object 
 2   SeniorCitizen     7043 non-null   int64  
 3   Partner           7043 non-null   object 
 4   Dependents        7043 non-null   object 
 5   tenure            7043 non-null   int64  
 6   PhoneService      7043 non-null   object 
 7   MultipleLines     7043 non-null   object 
 8   InternetService   7043 non-null   object 
 9   OnlineSecurity    7043 non-null   object 
 10  OnlineBackup      7043 non-null   object 
 11  DeviceProtection  7043 non-null   object 
 12  TechSupport       7043 non-null   object 
 13  StreamingTV       7043 non-null   object 
 14  StreamingMovies   7043 non-null   object 
 15  Contract          7043 non-null   object 
 16  PaperlessBilling  7043 non-null   object 
 17  PaymentMethod     7043 non-null   object 
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object 
 20  Churn             7043 non-null   object 
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

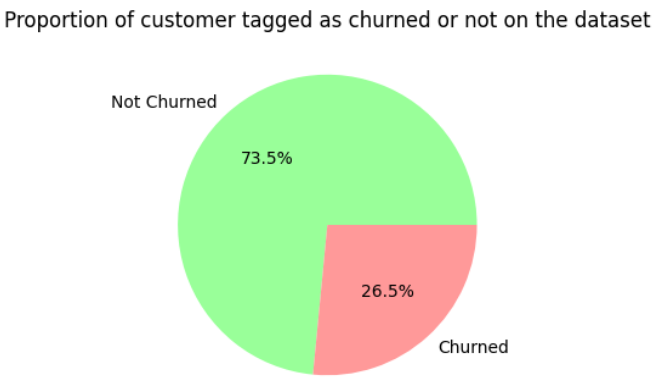
```
1 df_churn.head()
```

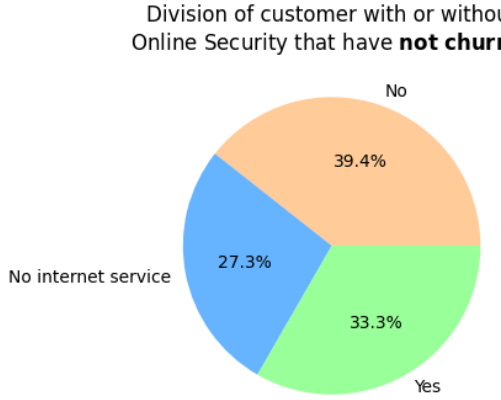
	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSu
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	

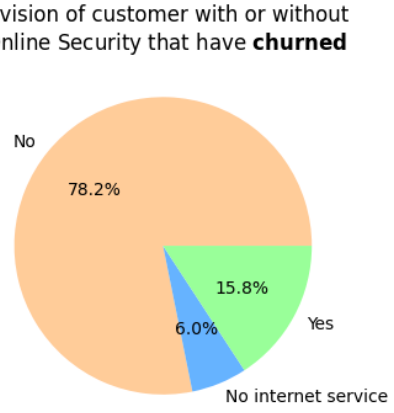
Security	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	...	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
Yes	...	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
Yes	...	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
Yes	...	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	...	No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

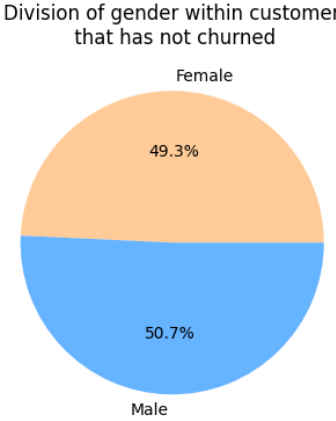
- Dataset was very much clean, only the column Total Charges had some blank spaces identified in the same rows where tenure was 0. Therefore, they was replaced by 0 also.

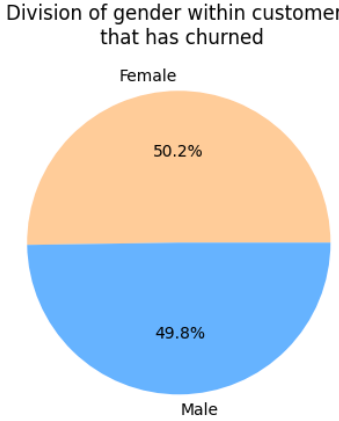
Exploratory data Analysis

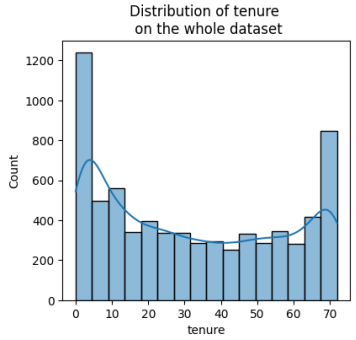


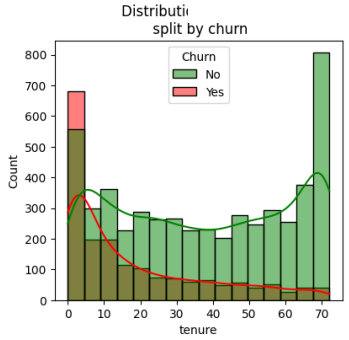


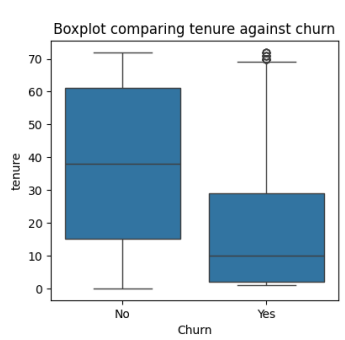


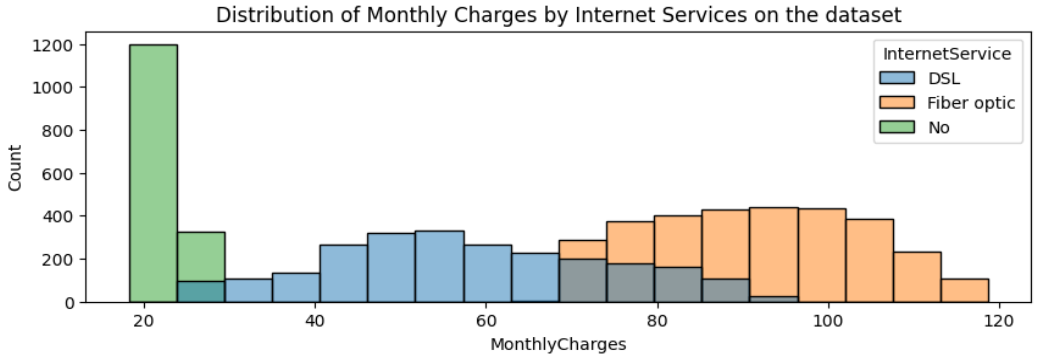






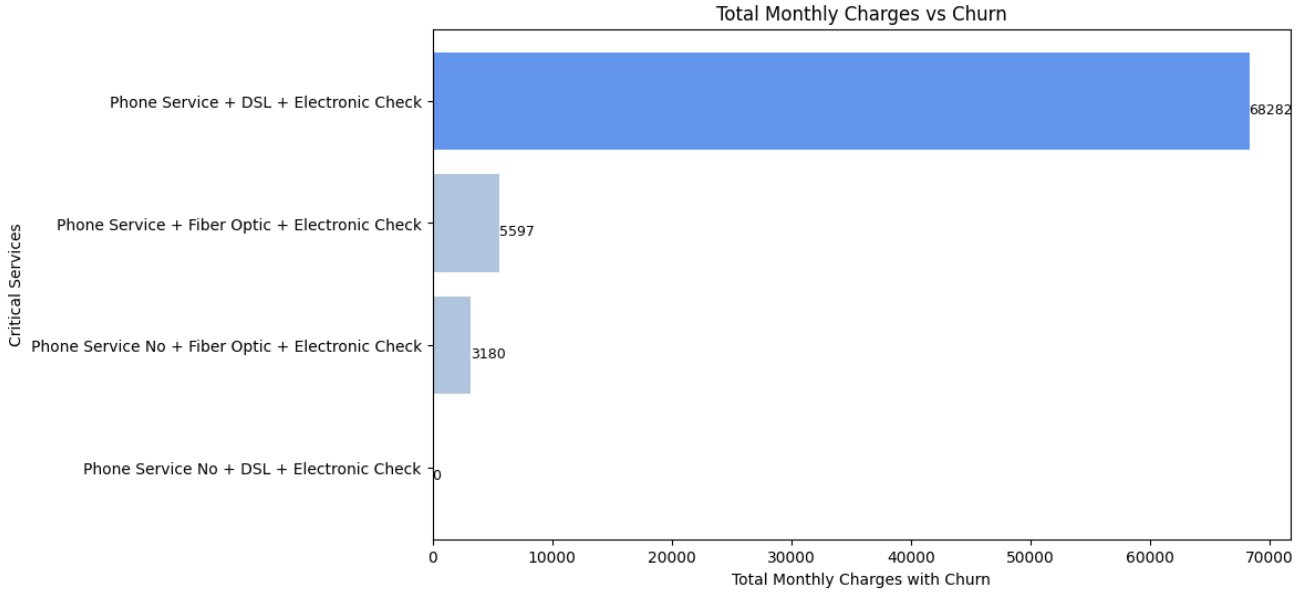


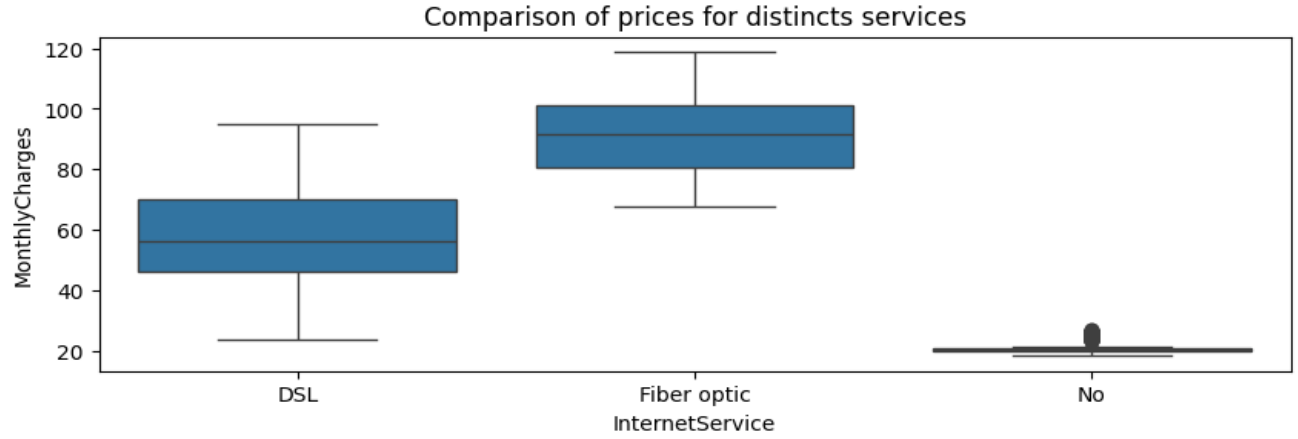




- Extra services reduces the chances of churn.
- 2/3 of customer that have churned had fiber optic service.
- Online Security, Tech Support and Device Protection seem to influence churn the most.

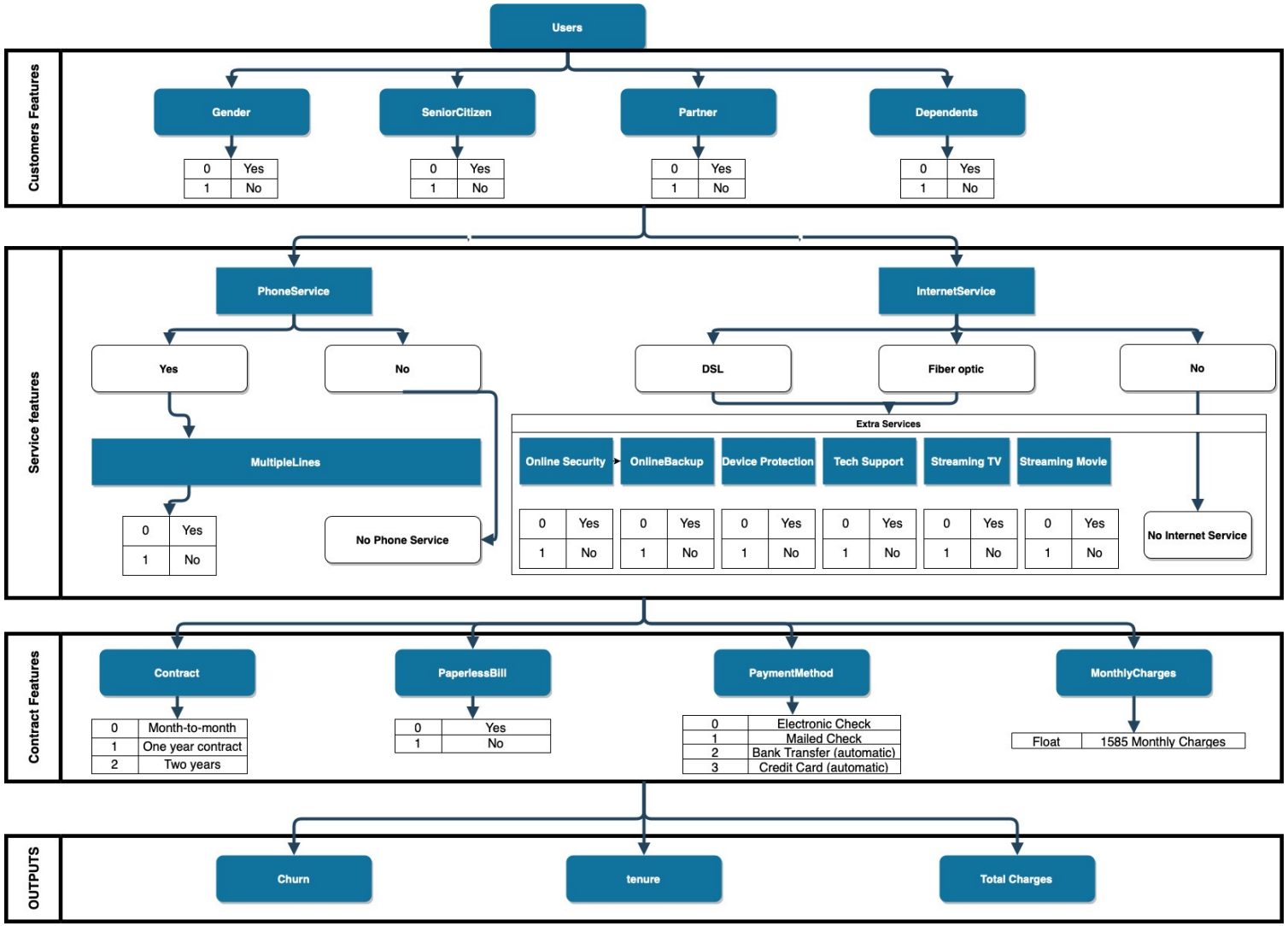
Exploratory Data Analysis



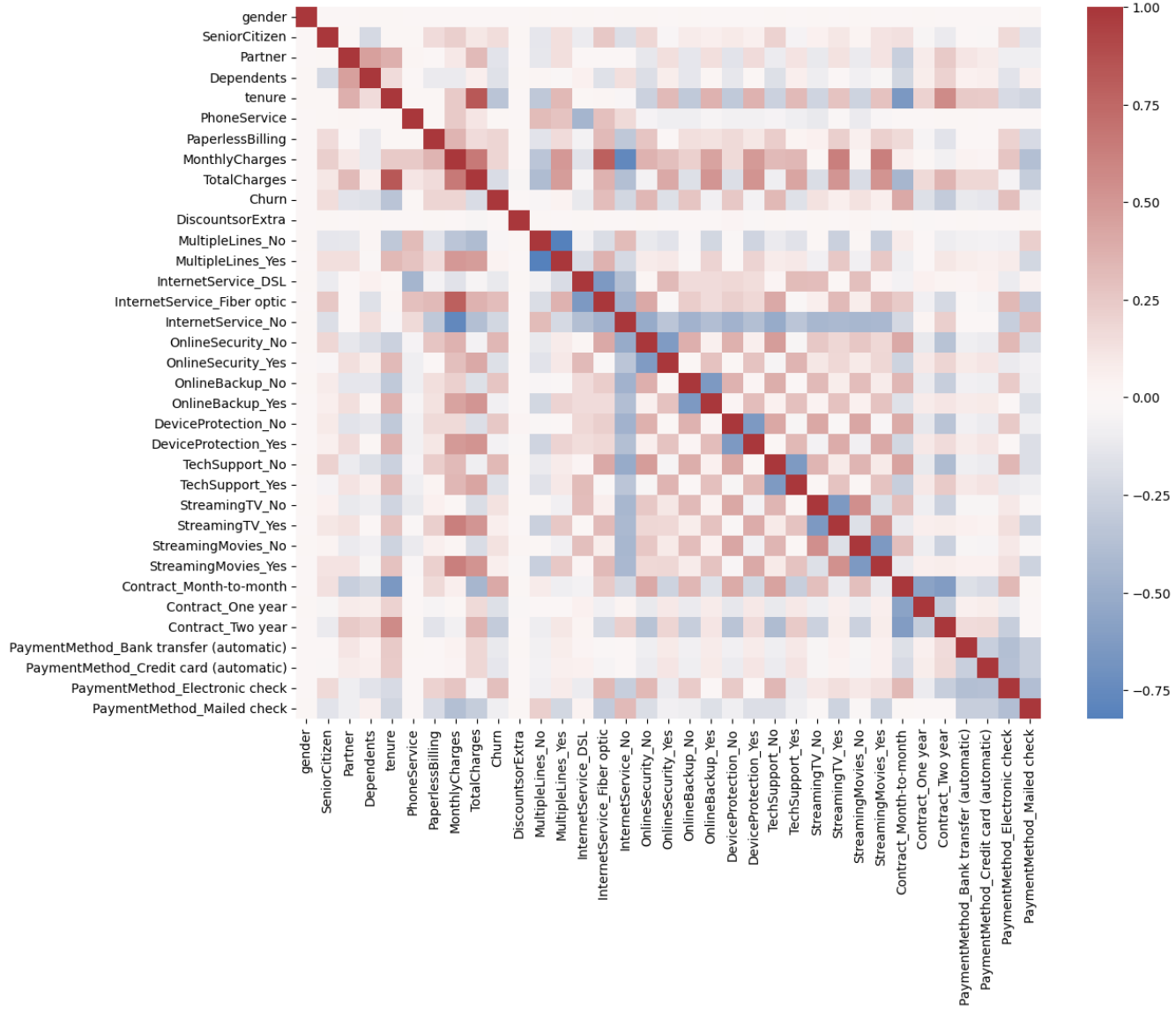


- Total loss for churn: 139,131.
- Loss just in customer with Phone Service, DSL internet and Electronic Check: 68,282. (49%)

Data Structure



Correlation Matrix




```
StreamingMovies_Yes                0.061382
MultipleLines_Yes                  0.040102
PhoneService                       0.011942
DiscountsorExtra                  -0.000307
gender                            -0.008612
MultipleLines_No                  -0.032569
DeviceProtection_Yes              -0.066160
OnlineBackup_Yes                  -0.082255
PaymentMethod_Mailed check        -0.091683
```

- Correlation with Churn

- Correlation with Total Charges

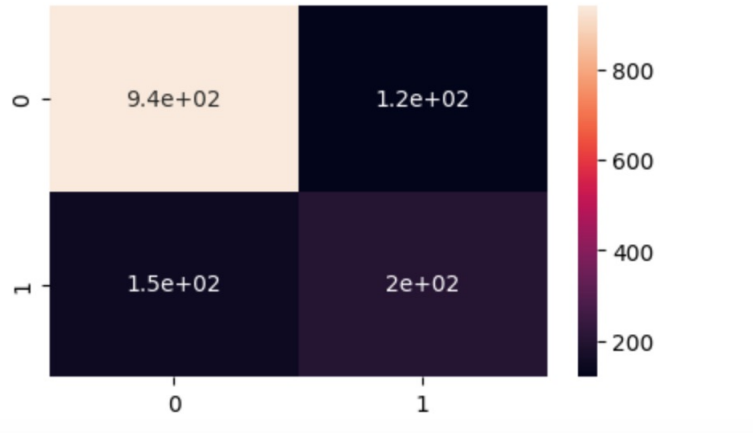
```
TotalCharges                       1.000000
tenure                             0.826178
MonthlyCharges                     0.651174
DeviceProtection_Yes               0.521983
StreamingMovies_Yes                0.520122
StreamingTV_Yes                    0.514973
OnlineBackup_Yes                   0.509226
```

Modelling

- Changing on business understanding

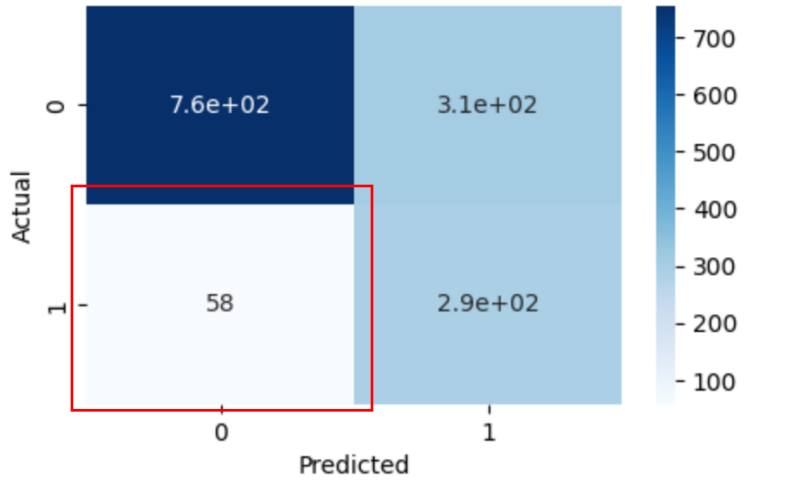
```
Accuracy Score:  0.8105039034776437

Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.89      0.88      1061
           1       0.63      0.58      0.60       348

    accuracy                           0.81      1409
   macro avg       0.75      0.73      0.74      1409
weighted avg       0.81      0.81      0.81      1409
```

<Axes: >



- Reduce Churn!
- Focus on false negatives.
- Focus on Recall
- False positives are target

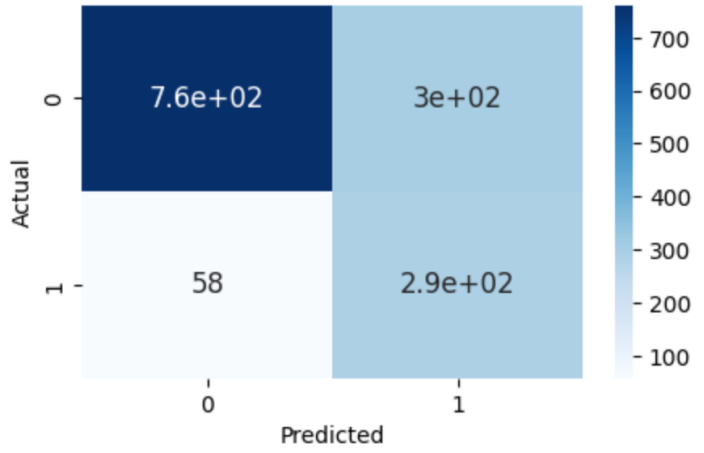
```
Accuracy Score:  0.7416607523066004

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.71      0.81      1061
           1       0.49      0.83      0.61       348

    accuracy                           0.74      1409
   macro avg       0.71      0.77      0.71      1409
weighted avg       0.82      0.74      0.76      1409
```



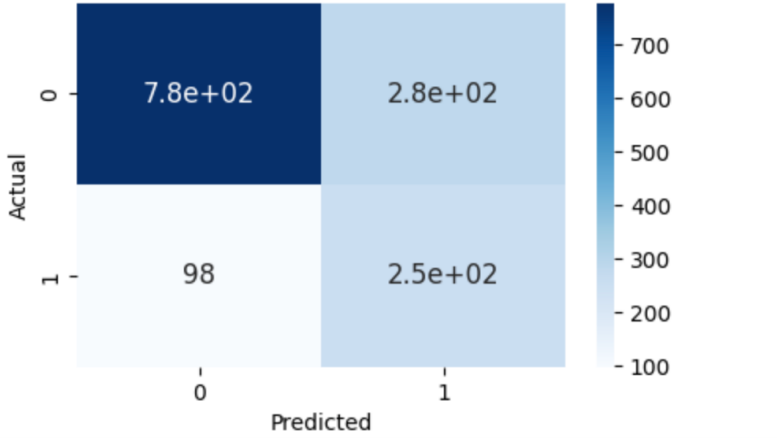
Modelling – Applying Smote

```
NB: 0.827834 (0.030637)
SVM: 0.880835 (0.022693)
KNN: 0.917563 (0.018560)
RFC: 0.844153 (0.020044)
CART: 0.812535 (0.030309)
LR: 0.826634 (0.022127)
```

```
Accuracy Score:  0.7459190915542938

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.72      0.81      1061
           1       0.49      0.83      0.62       348

    accuracy                           0.75      1409
   macro avg       0.71      0.78      0.71      1409
weighted avg       0.82      0.75      0.76      1409
```



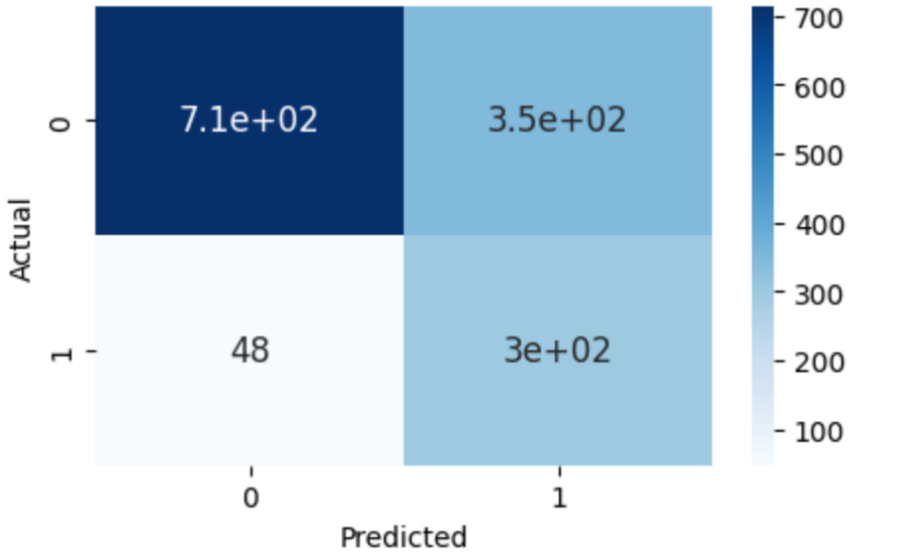

```
Accuracy Score:  0.7295954577714692

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.73      0.80      1061
           1       0.47      0.72      0.57       348

    accuracy                           0.73      1409
   macro avg       0.68      0.73      0.69      1409
weighted avg       0.78      0.73      0.75      1409
```



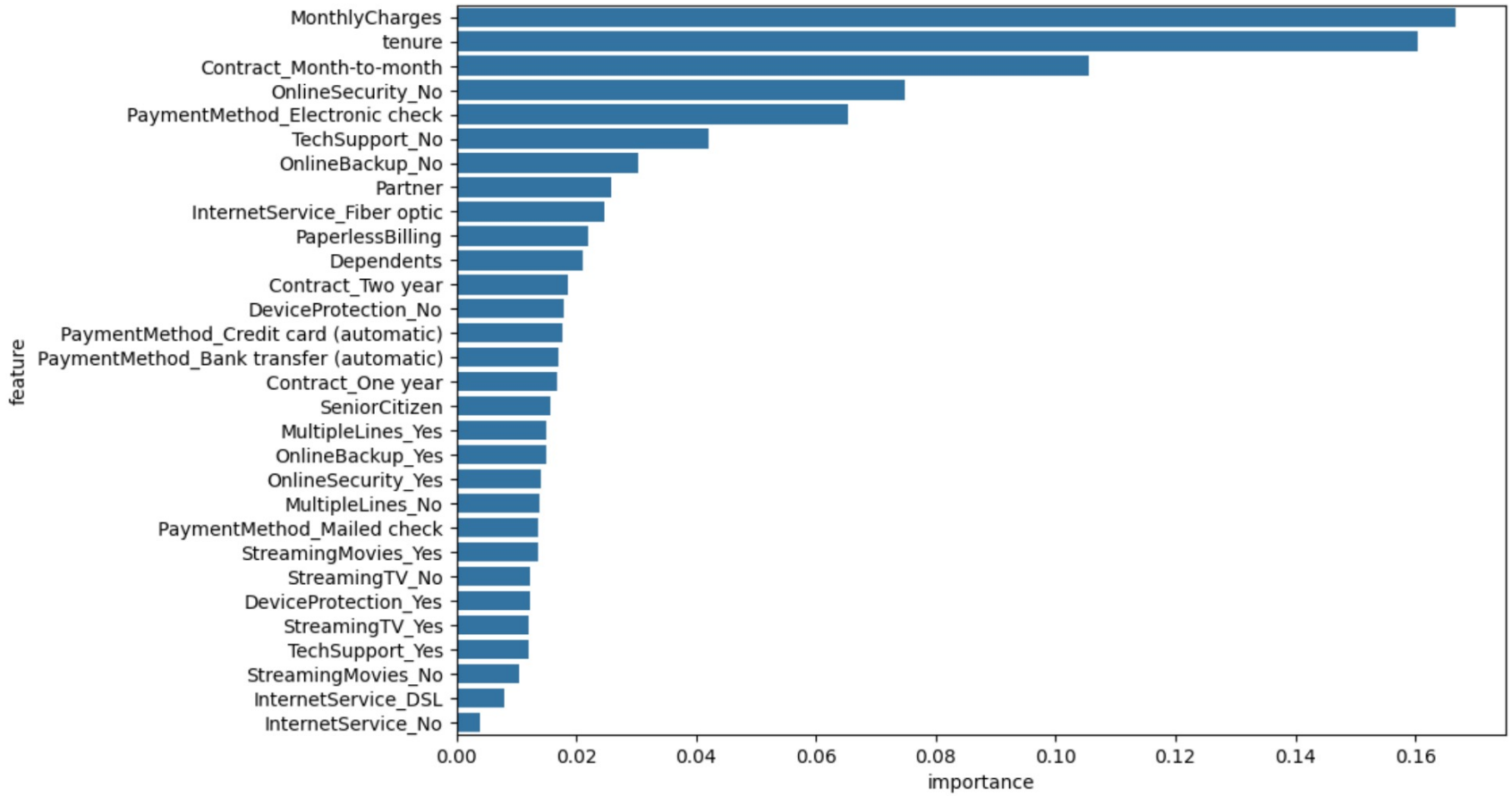
Modelling – Hyperparameter Tunning

```
#Setting parameters to be tunned and fold for cross validation
parameters = {
    'var_smoothing': np.logspace(0,-9, num=100)
}
k_folds = StratifiedKFold(n_splits=20, random_state=1, shuffle=True)

# Instantiate the grid search model
grid_search_rf = GridSearchCV(estimator = GaussianNB(), param_grid = parameters,
                              cv = k_folds, n_jobs = -1,verbose = 1, scoring=scoring)
```

```
Accuracy Score:  0.7196593328601846

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.67      0.78      1061
           1       0.46      0.86      0.60       348

    accuracy                           0.72      1409
   macro avg       0.70      0.77      0.69      1409
weighted avg       0.82      0.72      0.74      1409
```



Features Importance



Conclusion and next steps

- Most important features to churn:
  - Type of internet
  - Contract
  - Payment type
  - Extra service
- Best model: Gaussian Naïve Bayes optimized with Grid Search
- Split the dataset by internet type and analyze it separately.
- Correlations among variables.
- Verify Another models.

References

- Dino, L., 2022. *Medium.* [Online]
  Available at: https://medium.com/@24littledino/two-sample-chi-square-test-in-python-b9f2db89dc2b
  [Accessed 16 11 2023].
- Jain, K., 2021. *Analytics Vidhya.* [Online]
  Available at: https://medium.com/analytics-vidhya/how-to-improve-naive-bayes-9fa698e14cba
  [Accessed 18 11 2023]
- Jain, R., 2020. *Medium.* [Online]
  Available at: https://medium.com/@ritesh.110587/correlation-between-categorical-variables-63f6bd9bf2f7
  [Accessed 15 11 2023].
- Kübler, R., 2021. *Towards Data Science.* [Online]
  Available at: https://towardsdatascience.com/learning-by-implementing-gaussian-naive-bayes-3f0e3d2c01b2
  [Accessed 17 11 2023].
- Sharma, A., 2021. *Analytics Vidhya.* [Online]
  Available at: https://www.analyticsvidhya.com/blog/2021/01/gaussian-naive-bayes-with-hyperpameter-tuning/
  [Accessed 15 11 2023].
- Stack Overflow [online] (May 10, 2021) Available at: https://stackoverflow.com/questions/67474348/how-do-i-create-a-bar-chart-with-percentage-values-in-python-plotly-express
- GeeksforGeeks [online] (16 Jul, 2020) Available at: https://www.geeksforgeeks.org/python-binomial-distribution/?ref=gcse
- Scribbr [online] Published on November 5, 2020 by Pritha Bhandari. Available at: https://www.scribbr.com/statistics/standard-normal-distribution/
- Kaggle, Updated 5 Years Ago by Blastchar [online] Available at: https://www.kaggle.com/datasets/blastchar/telco-customer-churn
- W3Schools, Copyright 1999-2023 by Refsnes Data [online] Available at: https://www.w3schools.com/statistics/statistics_standard_normal_distribution.php

Thank you!