MASTER OF TECHNOLOGY INTELLIGENT SYSTEMS (PART-TIME) 2020

SEMESTER 1 – REASONING SYSTEMS FINAL PROJECT REPORT

CHURN FORTUNETELLER

TEAM MEMBERS

1. Anirban Kar Chaudhuri (A0108517H)
2. Maradana Vijaya Krishna (A0178453W)
3. Putrevu Manoj Niyogi (A0213557E)
4. Sivasankaran Balakrishnan (A0065970X)

CONTENT OVERVIEW:

1.0 Executive Summary

The Telecommunication industry mostly depends on subscription-based services. The profitability of the Organization mainly depends on its market share or its customer base. The customer acquisition and retention are two important factors that will directly impact the organization's profitability [2].

Churn rates depict the rate at which customer base shrink over a measured period and are often used to indicate the strength of a company's customer service division and its overall growth prospects. Lower churn rates suggest a company is in a stronger and competitive state. Customer loss impacts carriers significantly as they often make a significant investment to acquire customers. The ability to identify customers who will abandon services, while there is still time to do something about it, represents a huge additional potential revenue source for every online business. Furthermore, it is always more difficult and expensive to acquire a new customer than it is to retain a current paying customer.

Achieving right balance between customer acquisition and retention is not easy task. Statistics suggests that acquiring new customer is 5 times costlier than retaining the customer. Due to this, many companies are spending hugely to identify those customers who are likely to churn and taking the necessary steps to retain them and reduce the churn rate. Therefore, this project's aim to finding vital factors that increase customer churn is important to take necessary actions to reduce this churn.

2.0 Objective

Our team aims to create an **efficient** and **accurate** algorithm to identify the most prominent reasons for customer churn. The algorithm considers many factors such as customer's demographic, subscribed services and usage details. The prediction model assists the Telecom company to predict their customers who are more likely stop using their services. The prediction model developed in this work uses machine learning technique.

Building a machine learning model using the past customer churn data and their characteristics & behavior. The customer churn prediction model which assists telecom operators to predict customers who are most likely subject to churn based on various attributes related to demographics of customer, facilities offered, customer tenure and their charges. This machine learning project embodied in advanced analytics that leverage the existing customer's data to reveal hidden patterns and new insights that will enable the business users to take better decisions.

The supervised tree-based machine learning model will predict the response for existing customers. Upon completion of data analysis and feature engineering using if-else rules, we build tree-based machine learning methods and algorithms for predicting the customer churn. We have analyzed and implemented the Random Forest and Extreme Gradient Boost (XGBoost) tree-based models to build the predictive model.

3.0 Our Approach

Our team's approach relies on CRISP-DM process. The CRIPS-DM process provides an example of how to structure the Customer churn project. It consists of five stage, refer below.

i) Understand the Business Requirement

ii) Understanding & Analyzing the Data

iii) Build Machine learning model.

iv) Verify / Optimize the model.

v) Analyze the results and reveal the insights.

The data is one of the most critical / valuable asset of an organization. We are using publicly shared dataset Kaggle: Telcom Customer Churn Dataset contains more than 7000 customer details with several features. The collected data was full of columns, since there is a column for each service, product, and offer related to calls, SMS, MMS, and internet, in addition to columns related to personnel and demographic information.

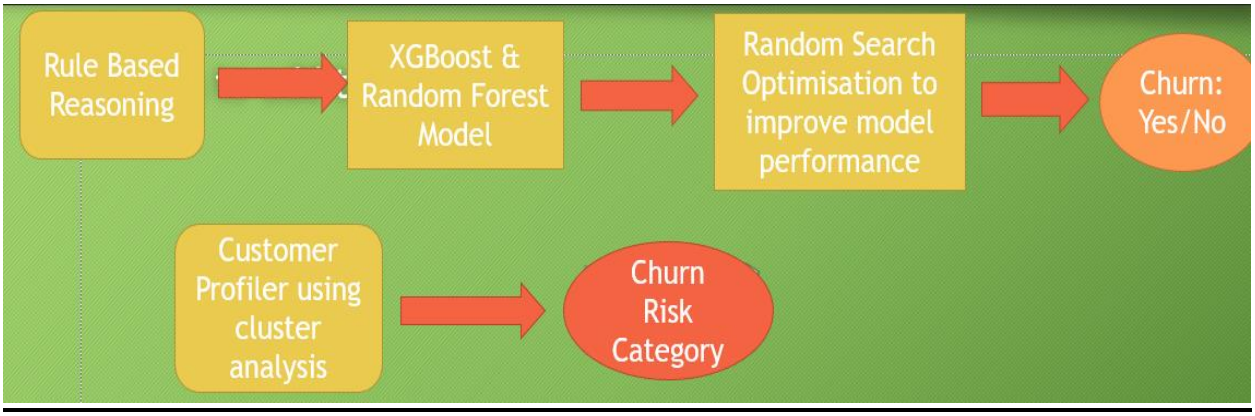
4.0 Technologies Used

The model developed in this work uses machine learning techniques using Python and its open source libraries,

- Pandas – Data preprocessing and manipulation
- Numpy – Statistical and Scientific Computing
- Matplotlib - Visualisation
- Seaborn – Visualisation
- Plotly – Interactive Visualisation
- Sklearn – machine learning preprocessors and model
- XGB Classifier – Package for extreme gradient boosting models (tree and linear)
- Flask – Deploying system on web server

We use Jupyter Notebook, that is an open-source application enables live coding and it allows us to tell a story with the code.

5.0 Churn Fortuneteller Architecture



Supervised Machine Learning Models Implemented in Churn Fortuneteller

- Random Forest Classifier (bagging) – Sum of individual decision trees and averaging over independent subsets of data (sampled with replacement)
- XGBoost Classifier (boosting) - Each decision tree learns from errors from previous learners which are assigned higher weights, sequential learners, optimization over error function
- Chosen Model: Random Forest (Lower false negatives, higher F1-score and accuracy)

Random Search Optimization

- Begin by selecting certain hyperparameter search space
- Randomness or probability (typically in the form of a pseudorandom number generator) in its methodology.
- The random element may be introduced through sampling specifications of the algorithm, or through noise in the function observation
- Finally, the outcome is given by optimal result for a given hyperparameter of a model to maximize it's predictive power
- Randomized Search Cross-Validation (CV) API in python scikit learn

Unsupervised K-Means Clustering

- Initialize k number of clusters and initial centroids randomly assigned
- Each instances assigned to nearest centroid which varies but eventually converges
- Euclidean distance used to measure and keep track of each dataset instance to shifting nearest centroids

6.0 Results and discussion

Visualizations

The following visualizations are employed in Churn Fortuneteller.

- Heatmap
- Boxplot analysis
- Pie Charts
- Countplots

Original Numeric predictors:

- monthly charge
- tenure
- total charge = monthly * charge

Original Categorical Influencers:

- Demographics - senior citizen, dependents, partner

- Security – Online Security, Online Backup, Device Protection, Tech Support
- Premium Services – Stream TV & Stream movies
- Payment & Billing - Payment Method, Paperless Billing, Contract

6.1 Exploratory Data Analysis - Customer Attrition based on categorical influencers

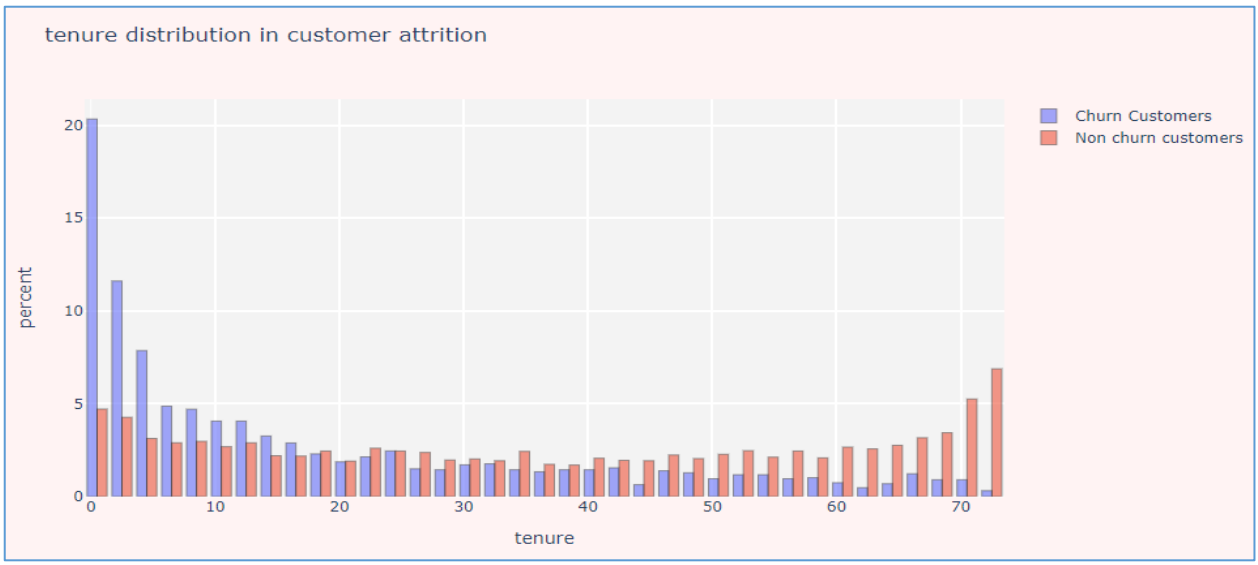
Refer *Annex A - Exploratory Data Analysis - Customer Attrition based on categorical influencers* for **complete** list of charts and more detailed explanation.

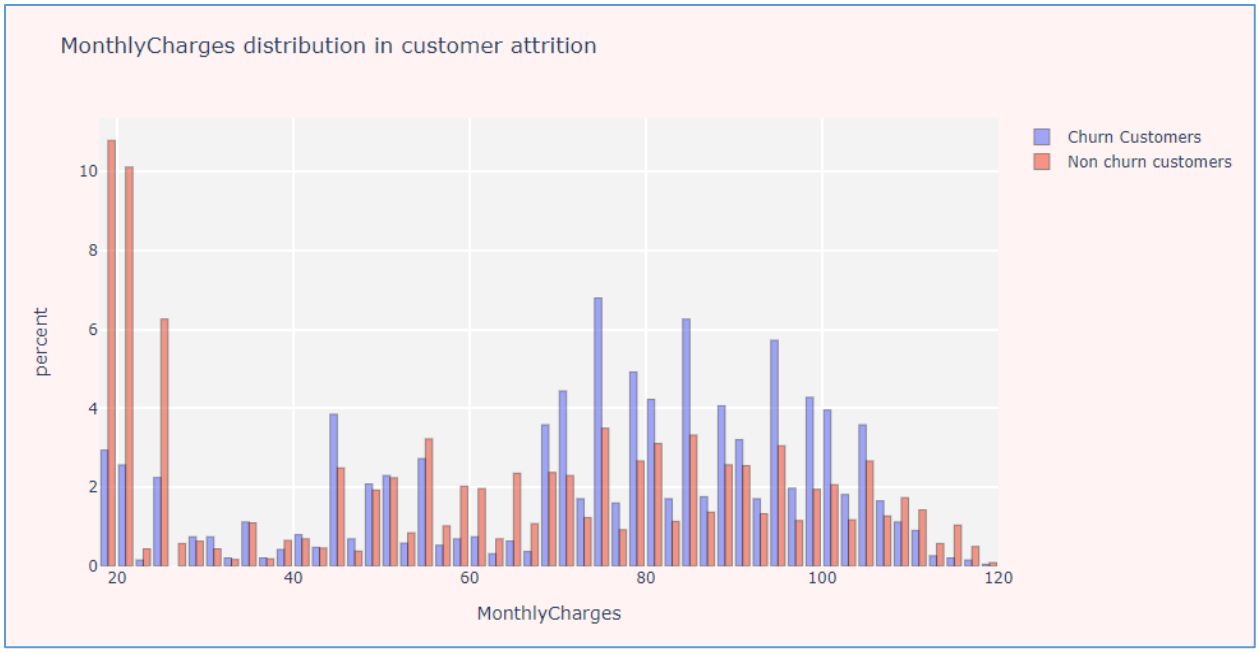
Inferences from pie charts:

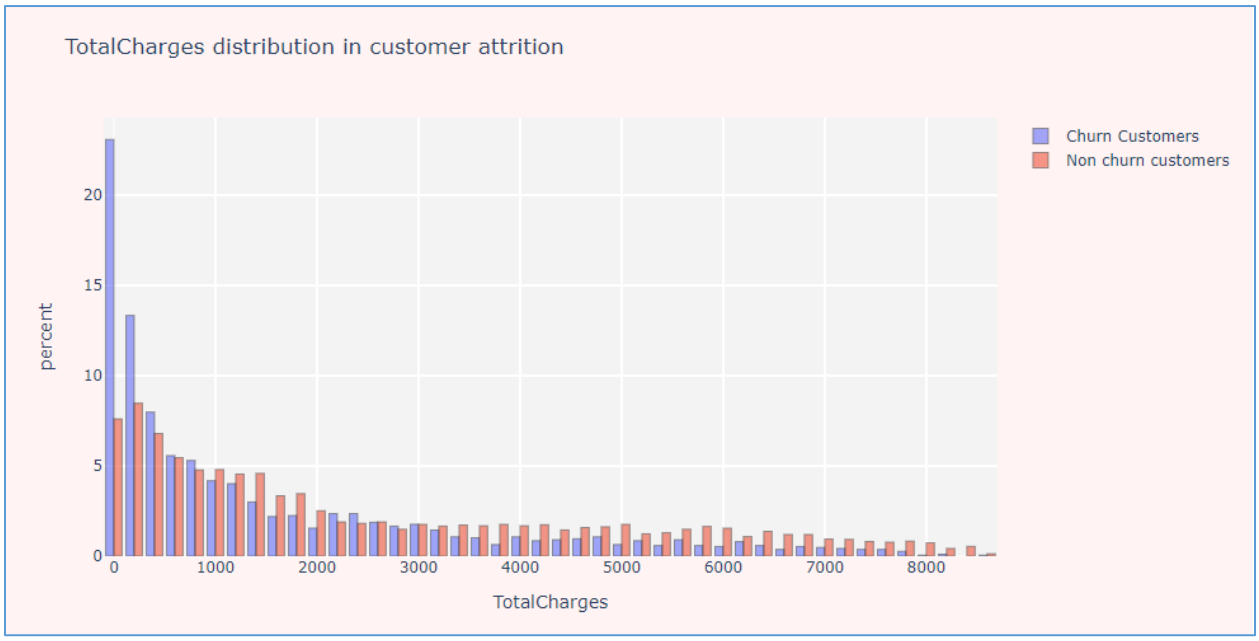
- Gender is not a good indicator of churn
- Customers that does not have partners are more likely to churn
- Customers without dependents are also more likely to churn
- Customers who are on month-to-month contract are likely to abandon company services
- Customers who have internet available, opt for paperless billing and automatic payment services are more likely to churn. These groups of customers tend to be tech savvy, read widely and be updated on latest market trends and rates.
- Customers who enjoy premium stream services are likely to leave, if they are lured by competitors offering similar services whose prices are competitive and offer better quality.
- Customers also tend to leave because of lack of technical support and online security as they're unlikely to find success in a company's products.
- Presence of phone service, especially multiple lines drive churn
- Without internet, you can't enjoy security, protection and premium streaming services
- Senior citizens have a higher probability of having dependents.
- People without partners generally do not have dependents

6.2 Exploratory Data Analysis – Customer attrition based on numeric influencers

Histograms of customer distribution according to charges and tenure

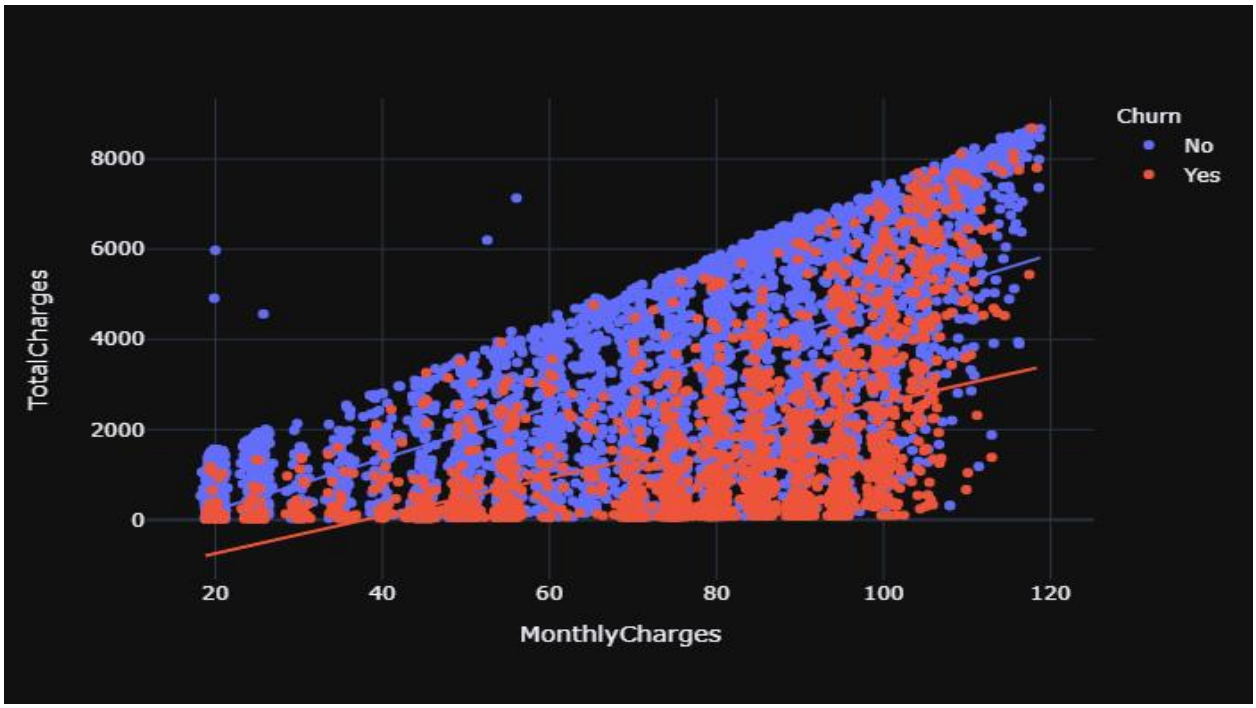






Inferences from histogram diagrams:

- 39% of the churn customers have a tenure of about 5 months.
- Number of churn customers have monthly charges peaked at around $75 per month while for no-churn it's $20
- Approximately 55% of churn customers cumulatively in frequency have maximum total charge of 900 dollars
- Monthly charge and tenure can be good predictors of churn given variations of counts in certain regions of value



Inferences from Scatterplot diagrams:

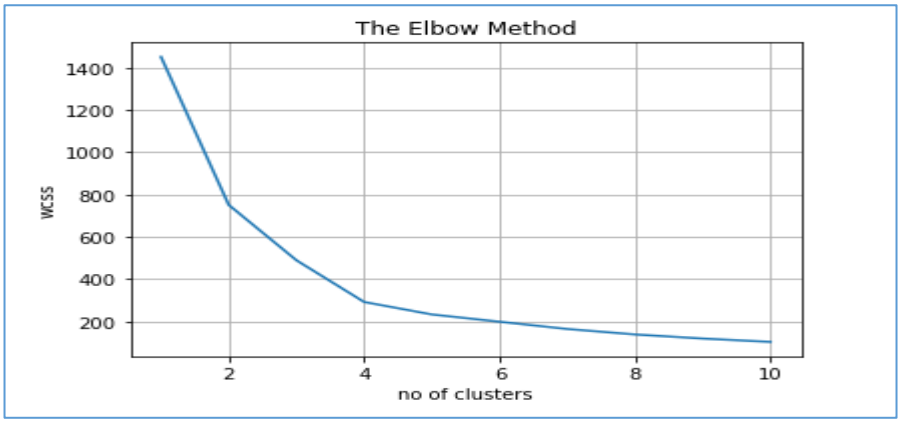
- Clients with lower tenure are more likely to churn as most of the churning customers subscribe for month-to-month subscription
- Clients with higher monthly charges are also more likely to churn
- Tenure and monthly charges are incredibly significant features in determining churn outcome
- **It's easier to predict non-churn customers compared to churn customers given how good fit a linear trend line is for churn and non-churn customers**

7. K-Means Clustering Based On Monthly Charges and Tenure

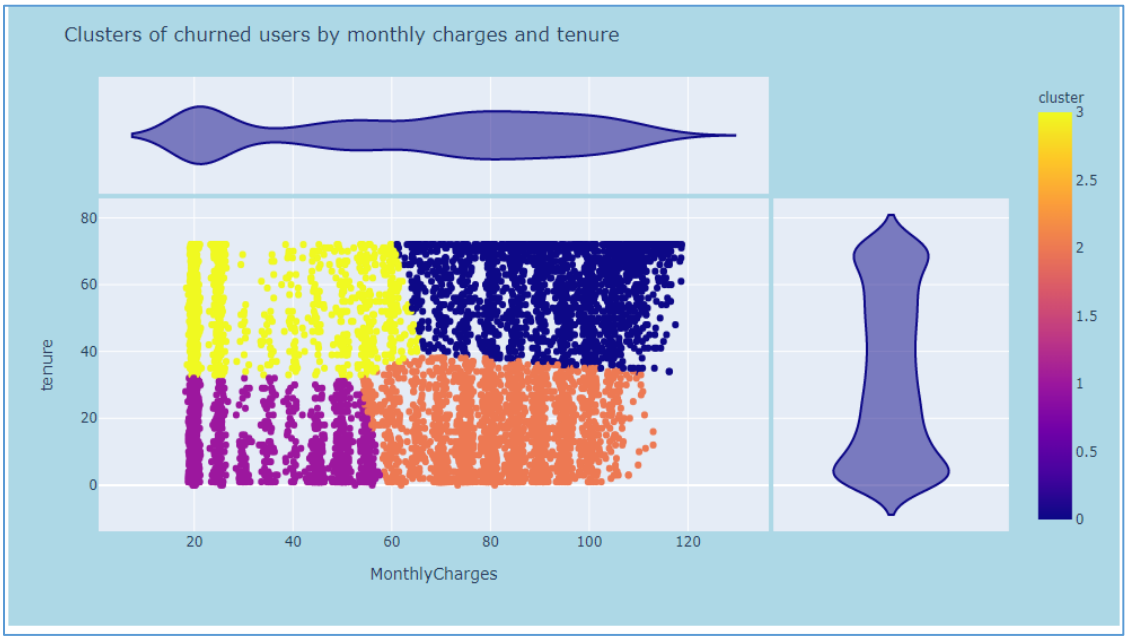
K-means clustering can be used to partition the dataset based on tenure and monthly charges, the significant numeric variables that significantly influence customer churn. Firstly, scale both features to using standard scaling technique (Standard Scaler API in Python Scikit Learn API). K-Means clustering uses distance metrics like Euclidean from cluster centroids to group instances together.

Purpose is to group instances of similar traits together. The K in K-Means denotes the number of clusters.

This algorithm initializes cluster centroids and assigns an instance to nearest centroid that shifts and varies periodically but randomly converges to a solution after some point in time.



- **Inertia** is the sum of squared error for each **cluster**. Therefore, the smaller the inertia the denser the cluster (closer together all the points are)
- Tip for choosing optimal number of clusters is looking at rate of decrease in inertia for addition of a cluster
- Optimal number of clusters is 4 since inertia does not decrease noticeably after additional clusters are added



Overall, clusters are well segregated as seen above.

Cluster 0: High tenure, high monthly charge

Cluster 1: Low tenure, low monthly charge

Cluster 2: Low tenure, high monthly charge

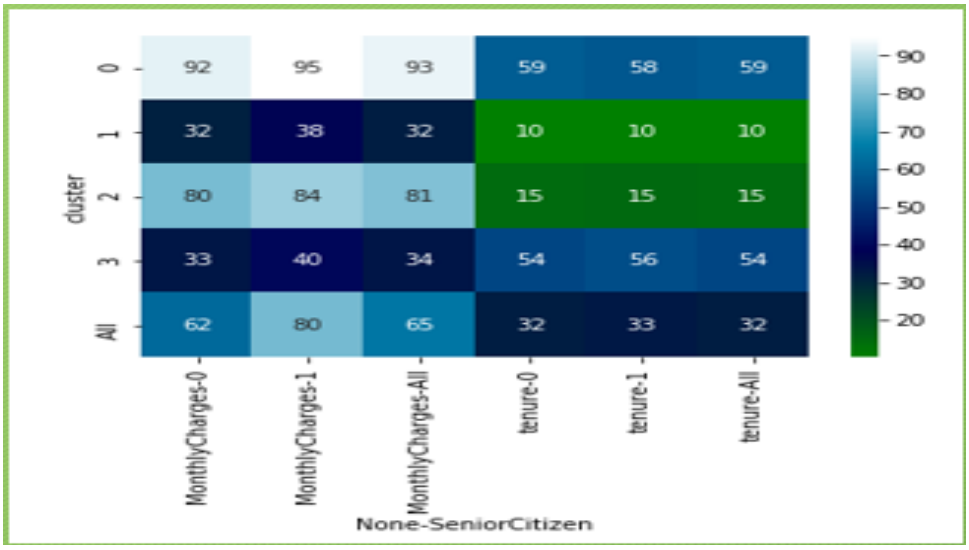
Cluster 3: High tenure, low monthly charge

Good clustering results in members of similar clusters close together while different clusters being segregated.

8. Cluster Analysis Visualizations

Monthly charges and Tenure of Senior Citizens

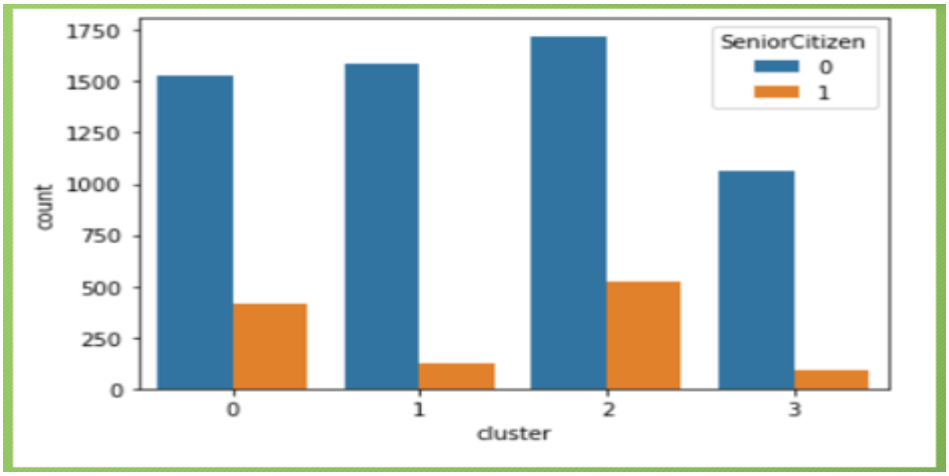
The pivot table below shows mean monthly charges and tenure of senior citizens in a cluster. Clusters where people churn have lower mean tenure and high monthly charge, even for senior citizens. This shows senior citizens have high likelihood of churning.



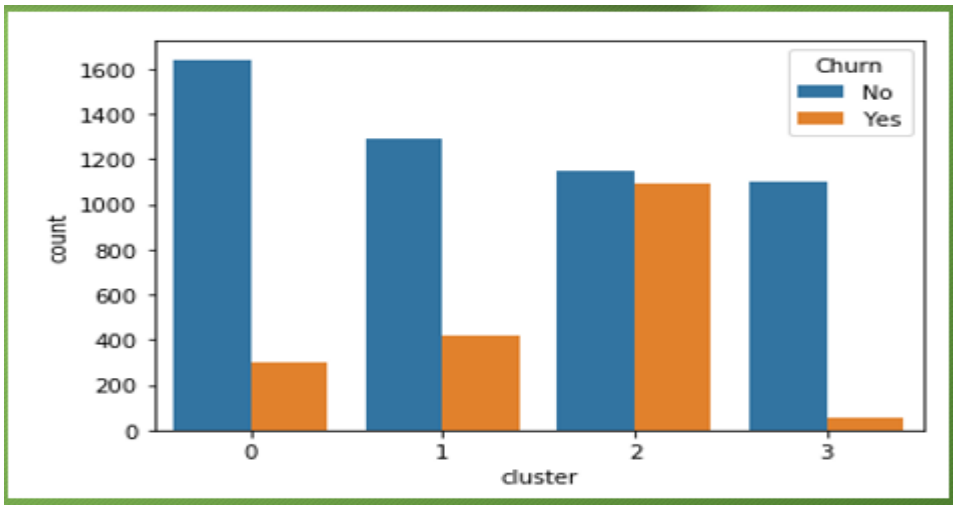
The figures in the table can be verified by hovering the cursor over the interactive graph above for cluster analysis. Joint effect of two random variable can be deduced as well.

Cluster Analysis – Monthly charges and Tenure

Refer *Annex B Cluster Analysis Based On Monthly Charges and Tenure* for complete list of charts.



Senior citizens fall under clusters 0 and 2, clusters with high monthly charge. The number of senior citizens in low tenure clusters outnumber those in high tenure clusters.

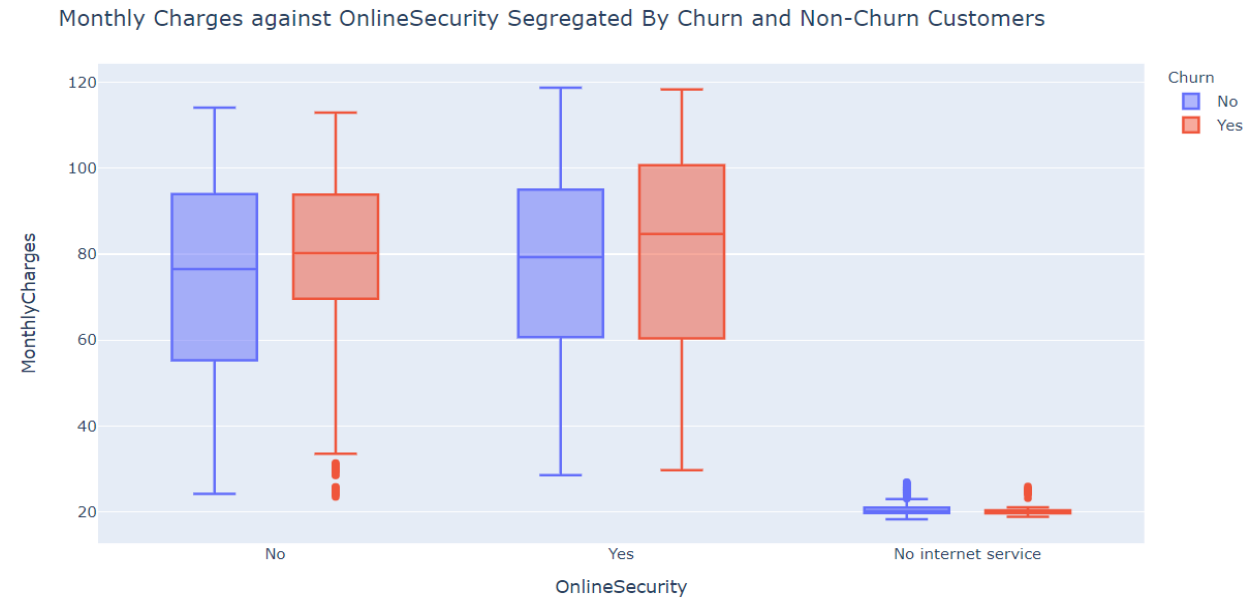


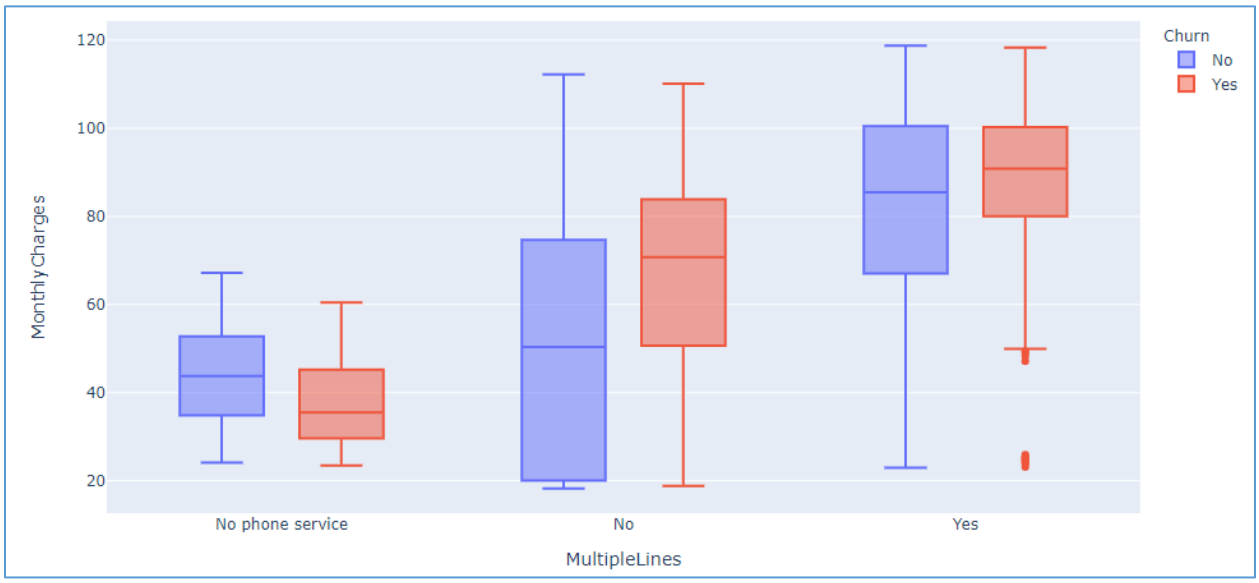
Clusters with descending order of churning probability: 2, 1, 0, 3

Cluster 3 is defined by high tenure and low, monthly charges, **ideal for retaining customers**.
Customers in category 2 (low tenure and high monthly charges), have highest probability of churning.
Cluster 0 customers have high tenure but high monthly charge. This shows monthly charge also an important predictor.

9. Boxplots Measuring Monthly Cost Against Amenities and Securities

Refer *Annex A - Exploratory Data Analysis - Customer Attrition based on categorical influencers* for **complete** list of charts and more detailed explanation.





Boxplot Inferences:

- Senior citizens tend to have higher cost monthly, even for those in churn groups. They are likely to churn but bring great benefits in revenue.
- Manual payment through checks are cheaper. Mailed check payment has largest range.
- Presence of internet increases monthly cost significantly. Addition of streaming cost poses greater costs.
- Having phoneline increases cost. Multiple phonelines raises monthly cost. High increases in costs due to internet and phone related services increases probability of churn.

10. Feature Engineering Using If-Else Rules inferred from Visualizations

The dataset underwent scrutiny and extensive analysis with the help of plotted visualizations and own domain knowledge of individual to create new, simplified features that are better interpretable and easier to understand using forward inference if-else rules. This process took the longest time due to the huge numbers of columns. The idea came from aggregating values of columns per month (average, count, sum, max, min ...) for each numerical column per customer, and the count of distinct values for categorical columns.

Level 1 Rules:

- If Dependents= “Yes” or Partner = “Yes”, then Family_Person = 1
- If Contract = “One year” or Contract = “Two year”, then Committed = 1
- If TechSupport = “Yes” or OnlineSecurity = “Yes” or OnlineBackup = “Yes” or DeviceProtection = “Yes” then Protection = 1
- If InternetService = “Fiber Optic” or InternetService = “DSL” then Has_Internet = 1
- If StreamingTV = “Yes” or StreamingMovies = “Yes” then Streaming =1
- If PaymentMethod not “MailedCheck” then Tech_Payment = 1

Level 2 Rules:

- If Tech_Payment = 1 and Streaming = 1, then Techie is 1.
- If Multiple_Lines = “Yes” and Internet_Service = “Fiber Optic” then Premium_Services = 1

Lastly, label encode other categorical features to be fed to machine learning models which usually takes in integers instead of strings.

11. Supervised Machine Learning Models Training & Evaluation

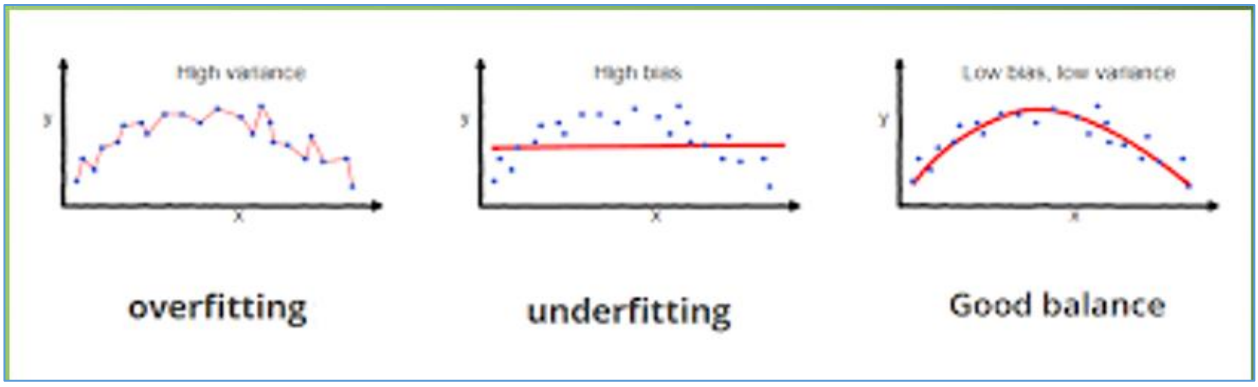
Positive class (churn) is 1 while negative class (non-churn) is 0. Churn instances in the entire dataset are under-represented in original dataset. Model may be biased towards non-churn instances.

```
print(churn_data["Churn"].value_counts()/len(churn_data)*100)

0    73.463013
1    26.536987
Name: Churn, dtype: float64
```

To mitigate the problem of biasness, take note of hyperparameters of various machine learning models that allow you to assign weight of importance for an imbalanced class underrepresented, including Random Forest Classifier and Extreme Gradient Boosting classifier we will be exploring.
Definitions of recall, precision and f1-score given below:

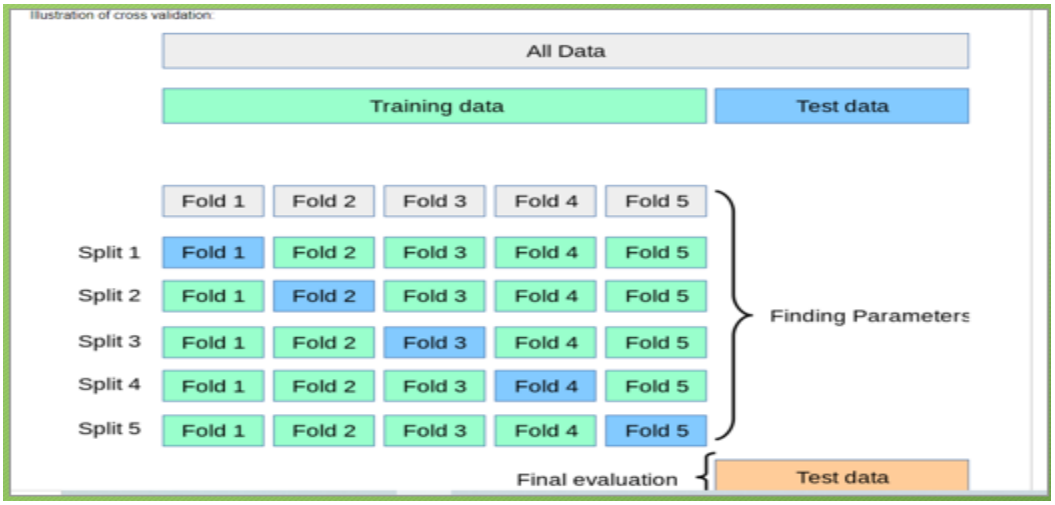
Accuracy = (TP+FP) / (TP+FP+TN+FN) (How many instances correctly identified out of all)
Precision = TP / (TP + FP) (Fraction of a particular class retrieved that are accurate)
Recall = TP / (TP +FN) (Fraction of a particular class that are successfully retrieved)
F1-Score = 2 *((*Precision* Recall) / (Precision + Recall)) (Mean harmonics of precision and recall)



Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model, leading to high error on training and test data. (underfitting).

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data (overfitting).

Error = Bias + Variance + Irreducible Error



The solution we proposed divided the data into two groups: the training group and the testing group. The training group consists of 75% of the dataset and aims to train the algorithms. The test group contains 25% of the dataset and is used to test the algorithms.

Cross-validation carried out further on training dataset that's split into k folds and each fold tested once and trained (k – 1) times. Purpose is to evaluate performance of model on seen data.

First, split data into train and test

	fit_time	score_time	test_score	train_score
0	0.243926	0.009978	0.773559	0.815499
1	0.234973	0.010008	0.790508	0.817873
2	0.237095	0.008974	0.786974	0.820448
3	0.235932	0.008970	0.783582	0.821465
4	0.238915	0.009969	0.790366	0.819939

The difference between mean test and train score used to determine if a model overfitting on training data or not. Alternatively, learning curves showing variation of test and train.

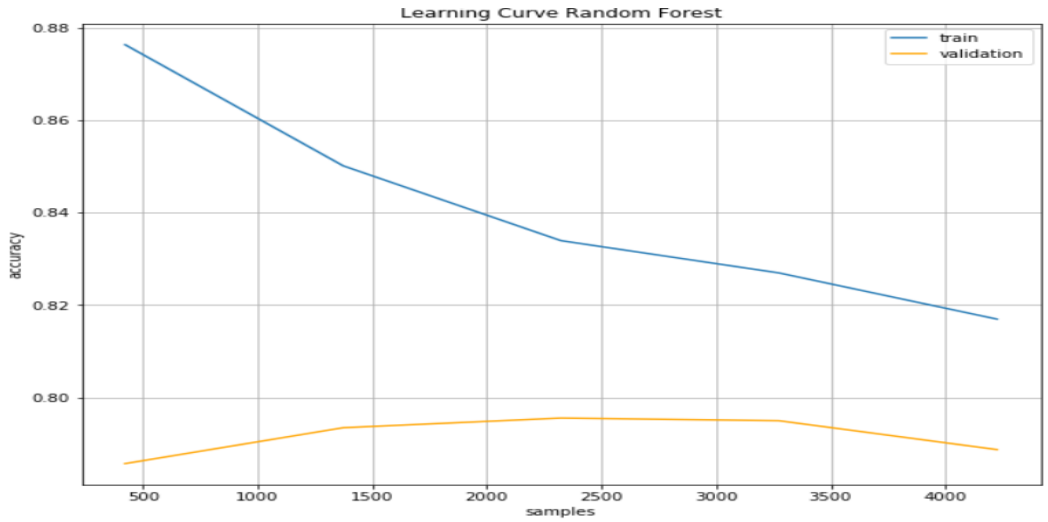
11.1 Random Forest Initial Predictions (Chosen) Model

A random forest is an ensemble of bagging trees trained on independent subsampled datasets and whose individual predictions are averaged. This reduces variance often for noisy data.

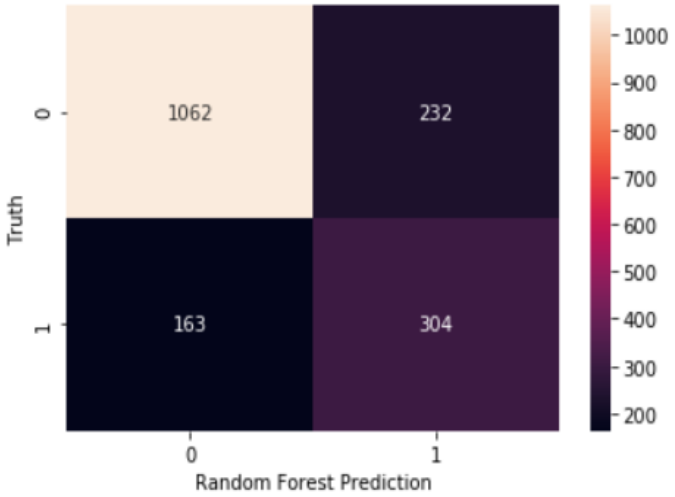
	fit_time	score_time	test_score	train_score
0	0.076137	0.111032	0.790918	0.816805
1	0.089761	0.106454	0.781457	0.815858
2	0.074482	0.111408	0.794508	0.815902
3	0.085234	0.111419	0.794508	0.814955
4	0.082812	0.106489	0.781250	0.820398

```
print('Mean Test Score for Random Forest Classifier: ', scores_rf['test_score'].mean())
print('Mean Train Score for Random Forest Classifier: ', scores_rf['train_score'].mean())
```

```
Mean Test Score for Random Forest Classifier:  0.7885279593474958
Mean Train Score for Random Forest Classifier:  0.81678337258504
```



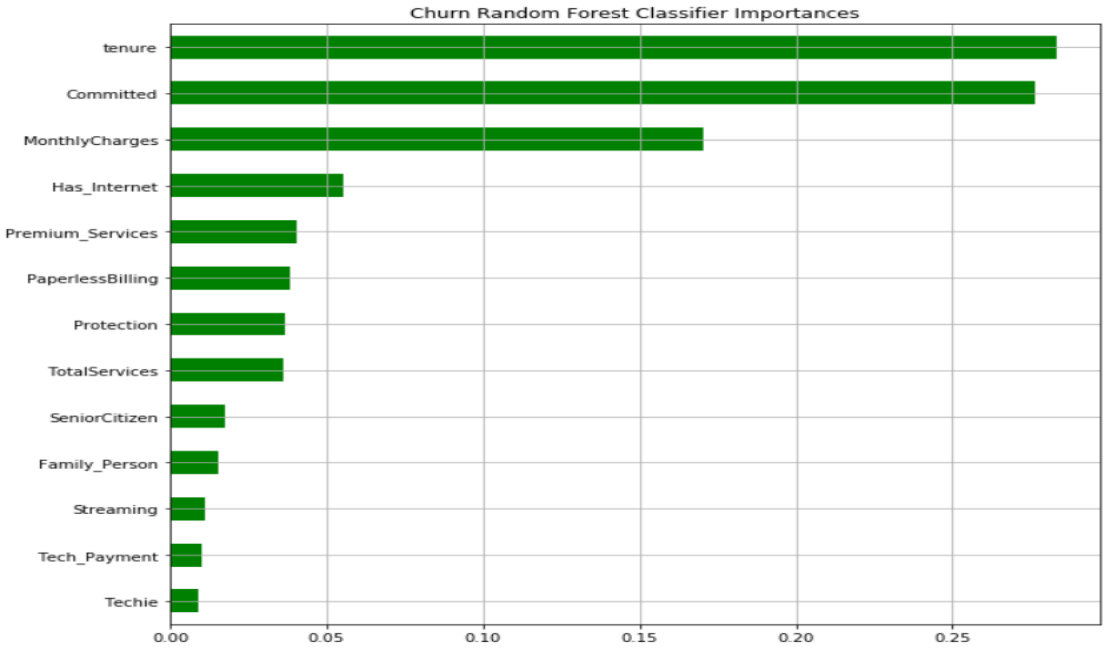
The train and validation score learning curves are converging, indicating that the Random Forest Classifier is learning well. 30 estimators with a depth of 8 nodes are used. A ratio of assigned weights to classes 0:1 is 1:1.8, which means model is being instructed to pay greater attention to churn category.



	precision	recall	f1-score	support
No	0.87	0.82	0.84	1294
Yes	0.57	0.65	0.61	467
accuracy			0.78	1761
macro avg	0.72	0.74	0.72	1761
weighted avg	0.79	0.78	0.78	1761

Confusion matrix shows the frequency distribution of true positives, true negatives, false positives and false negatives.

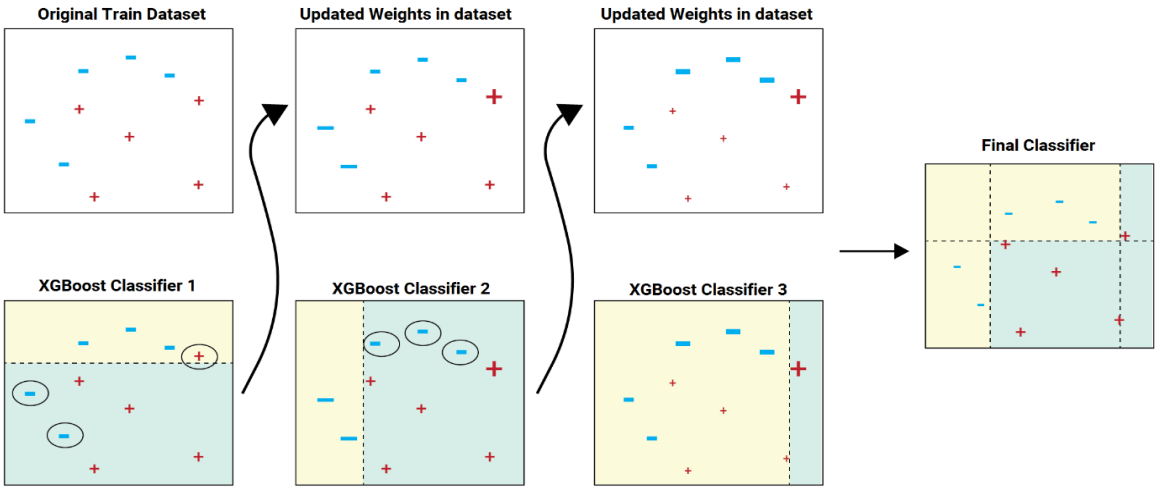
232 transactions marked as churn but actually non-churn (false positives). 163 tranactions marked as non-churn but actually are churn customers (false negatives). As for classification report, f1-score and accuracy score were considered in choosing of model apart from learning curve that determines if a model overfits or not.



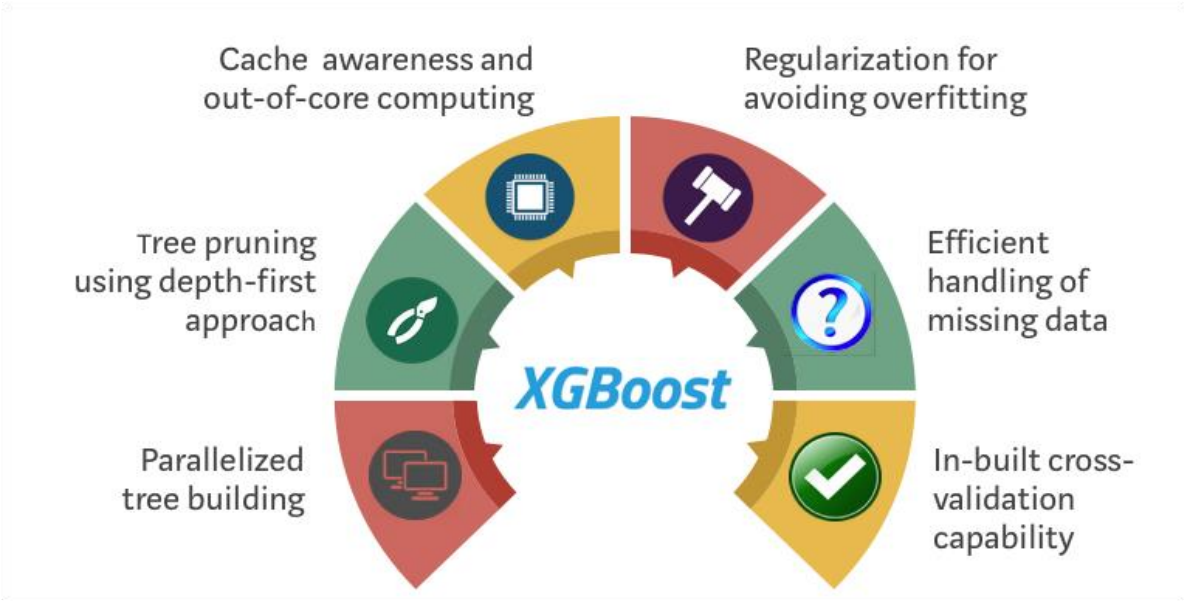
The random forest places high emphasis on factors related to tenure and costs when feature importance score are calculated.

11.2 Extreme Gradient Boosting Classifier Predictions

An extreme gradient boosting algorithm is an ensemble of sequential decision trees that learns from error of previous learners. Previous errors given higher weights subsequently depending on learning rate.



XGBoost algorithm is extremely popular among boosting algorithms. It is fast, able to handle missing values and known for parallelized computations and in-built cross validation capabilities. You can play with **depth** and **number of estimators** like random forests too. **Learning rate** responsible for size of weights of errors that determines how fast model learns. Regularization functions (**reg_alpha** & **reg_lambda**) that control coefficients of features to avoid overfitting.

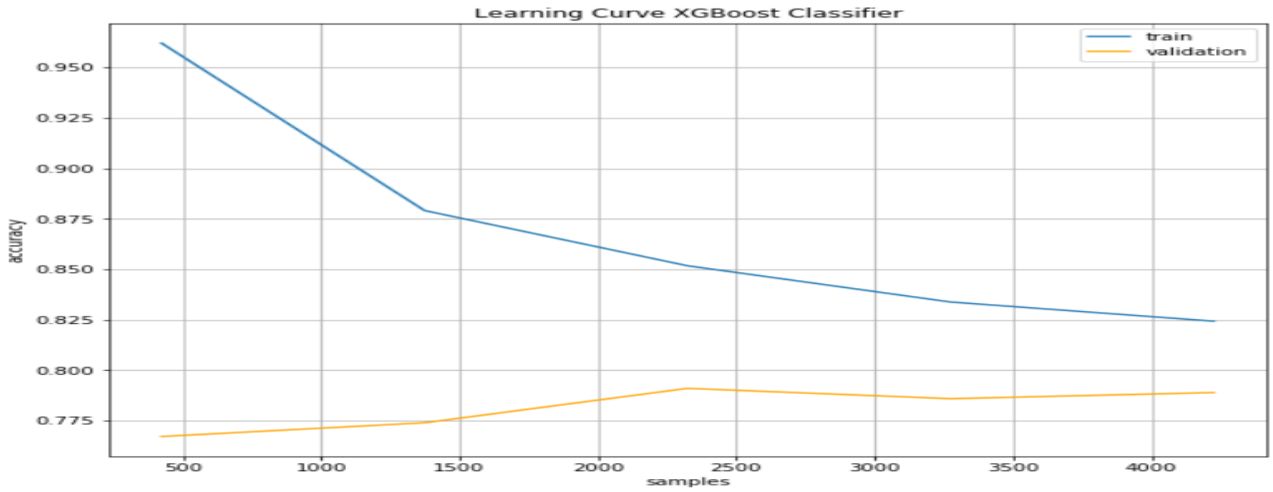


```
# cross validate
scores_xgb = cross_validate(xgb, x_train, y_train, cv=5, return_train_score=True,retu
scores_xgb = pd.DataFrame(scores_xgb)
scores_xgb
```

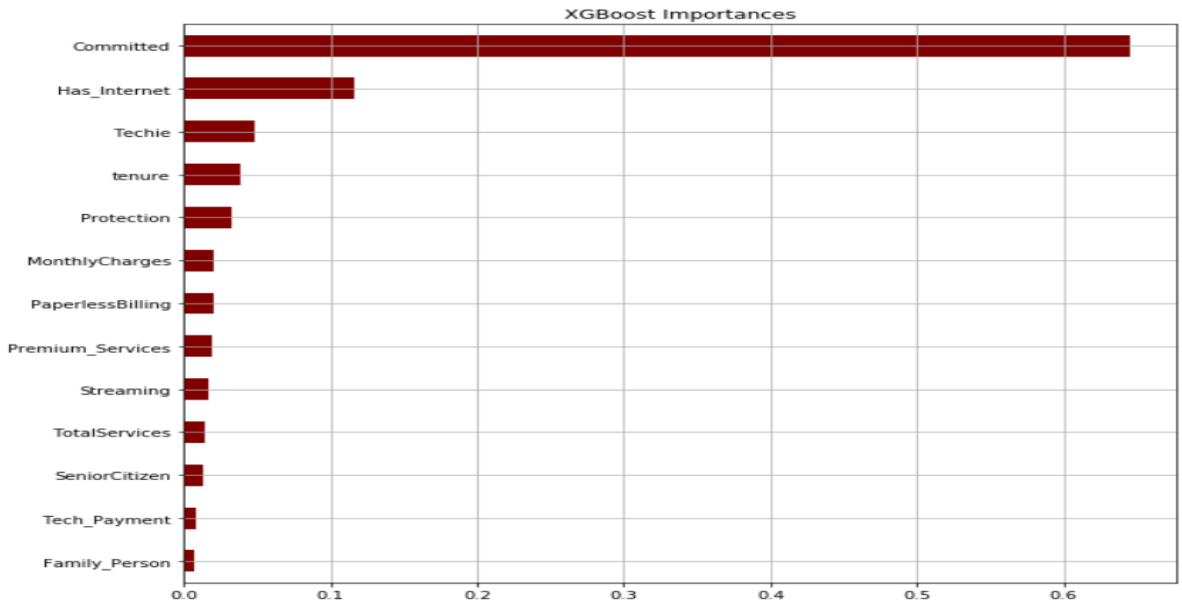
	fit_time	score_time	test_score	train_score
0	0.027925	0.004986	0.799432	0.825799
1	0.025931	0.004987	0.786187	0.821538
2	0.027925	0.004987	0.785985	0.825367
3	0.024933	0.004987	0.801136	0.824420
4	0.028923	0.004987	0.775568	0.824894

```
print('Mean Test Score for XGBoost Classifier: ', scores_xgb['test_score'].mean())
print('Mean Train Score for XGBoost Classifier: ', scores_xgb['train_score'].mean())
```

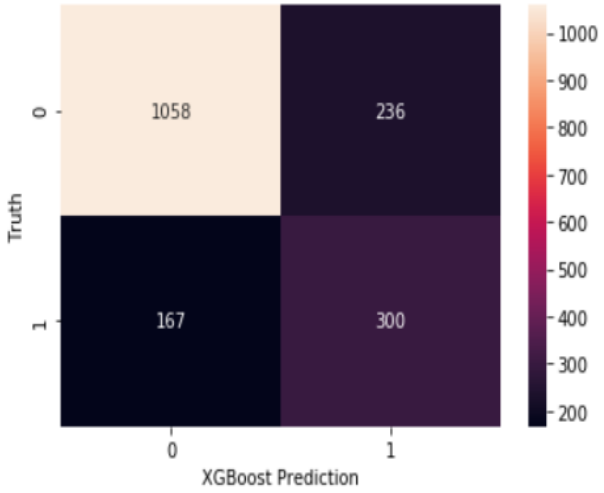
```
Mean Test Score for XGBoost Classifier:  0.7896618144548608
Mean Train Score for XGBoost Classifier:  0.8244035654177997
```



XGBoost has a higher tendency to overfit and harder to finetune but in this case it shows same performance as random forest during cross validation.



XGBoost places sole importance on commitment status and presence of internet to some extent.

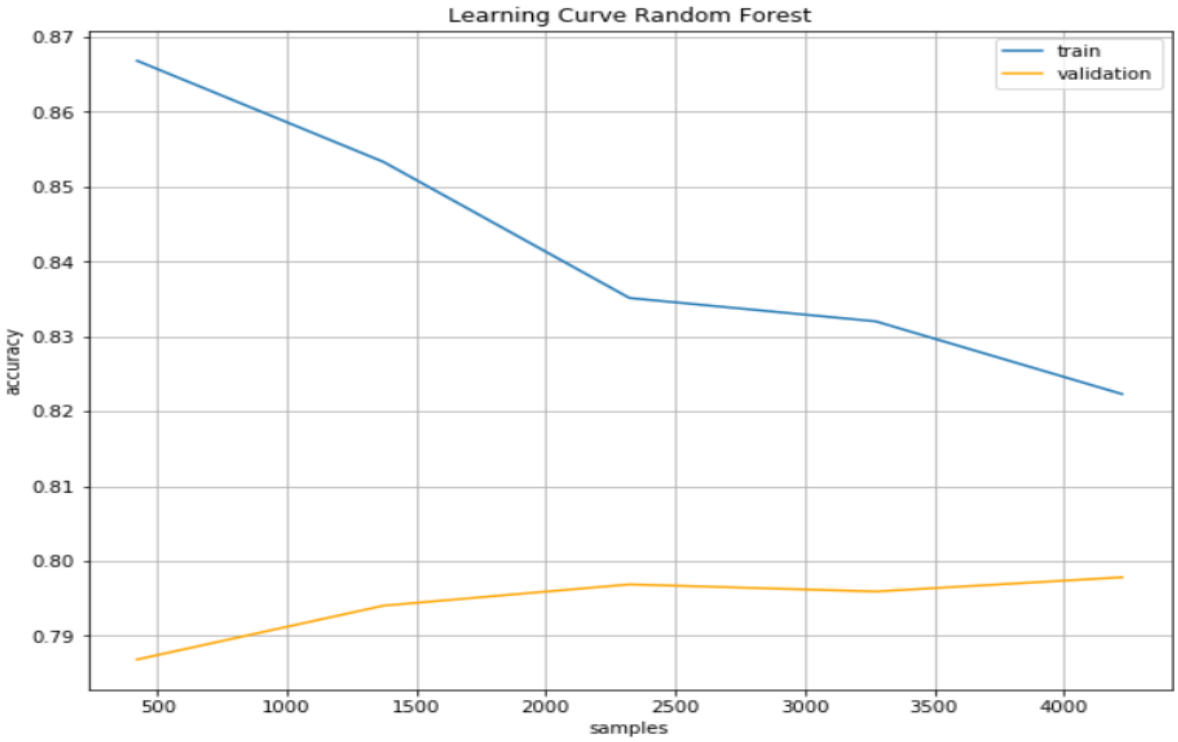


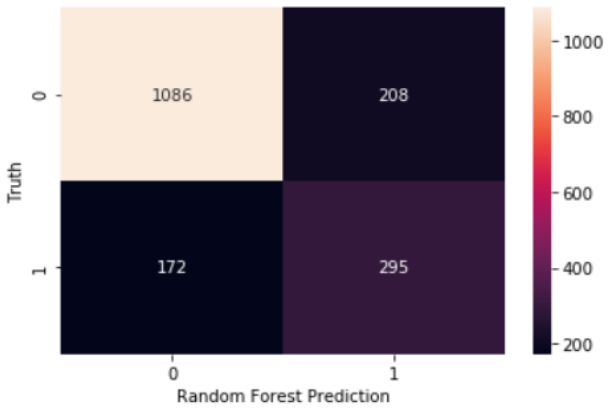
	precision	recall	f1-score	support
No	0.86	0.82	0.84	1294
Yes	0.56	0.64	0.60	467
accuracy			0.77	1761
macro avg	0.71	0.73	0.72	1761
weighted avg	0.78	0.77	0.78	1761

XGBoost shows higher number of false positives and negatives than random forest. It also has a lower F1-score and accuracy. Trade-off between false positives and false negatives. High false positives result in wastage of money spent on marketing campaigns. False negatives also pose a problem: customer leaving result in loss of revenue. F1-score considers mean harmonics of precision and recall (FP and FN both considered), useful during uneven class distribution.

11.3 Finetuned Random Forest Classifier Predictions (Final)

We now look at performance of finalized Random Forest Classifier whose few hyperparameters were optimized using random search algorithm to be discussed in next section.





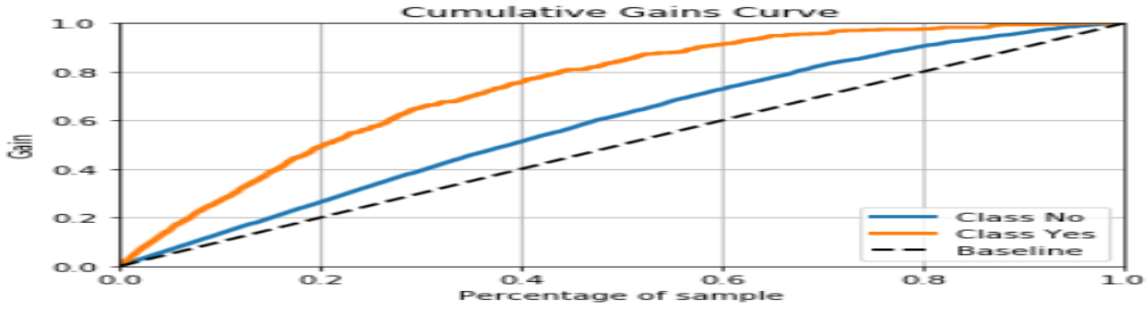
	precision	recall	f1-score	support
No	0.86	0.84	0.85	1294
Yes	0.59	0.63	0.61	467
accuracy			0.78	1761
macro avg	0.72	0.74	0.73	1761
weighted avg	0.79	0.78	0.79	1761

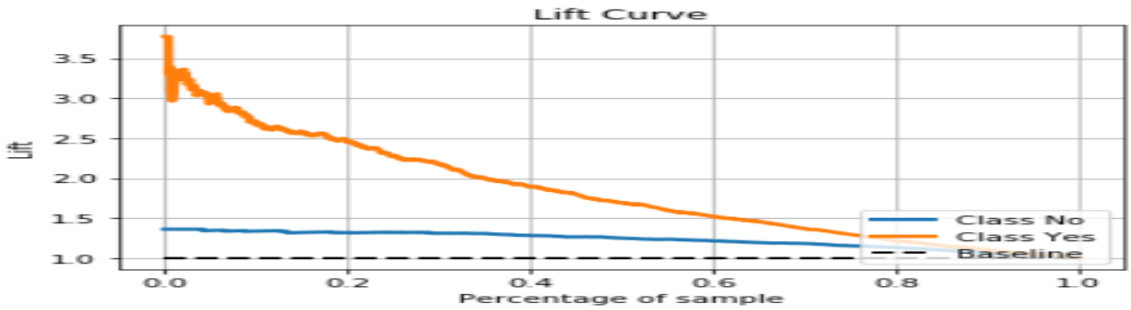
Random Forest finetuned is fitting better than original one as the 2 curves are converging faster.F1-score and accuracy of final random forest model is same as original one. However, number false positives has decreased significantly and that offsets increase in false positives.

Lift and cumulative charts for Random Forest:

1) **Lift** is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.
2) **Cumulative gains** and lift charts are visual aids for measuring model performance.
3) Both charts consist of a curve and a baseline.

If we observe carefully, we can reach out to over 80% of churn customers with this random forest model if marketing budget targets 50% of its customers according to **cumulative gain curve**. With a random pick in absence of this model, we will be reaching out to 50% of churn customers. **Lift** is calculated as the ratio of Cumulative Gains from classification and random models. If the average incidence of targets is 20%, so the lift is 2.5. Thus, the model allows addressing two-and-half times more targets for this group, compared with addressing without the model (randomly).





12. Random Search Optimization

Random search optimization technique was used to find optimal parameters of the Random Forest model to maximize it's predictive performance.

Begin by selecting certain hyperparameter search space. Randomness or probability (typically in the form of a pseudorandom number generator) in its methodology. The random element may be introduced through sampling specifications of the algorithm, or through noise in the function observation. Random search works by iteratively moving to better positions in the search-space, which are sampled from a hypersphere surrounding the current position.

Define f(x) as a fitness / cost function which must be minimized. Let x be a position or candidate solution in the search-space. The basic Random Search algorithm can then be described as:

1. Initialize x with a random position in the search-space.
2. Until a termination criterion is met (e.g. number of iterations performed, or adequate fitness reached), repeat the following:
    1. Sample a new position **y** from the hypersphere of a given radius surrounding the current position **x**
    2. If *f(x) > f(y)* then move to the new position by setting **x** = **y**

Random Forest Finetuning Strategies:

1. Choose a relatively **low number of estimators** somewhere between 20-40. Random Forests perform slowly with larger datasets.
2. **Tune tree-specific parameters** like max_depth (5-10), max_features (0.3-0.5) and class_weights (0: 1, 1: 1.5-1.8) for deciding depth of learning, what fraction of total features used for splitting nodes at a time and paying greater importance to underrepresented churn class. Take note of ranges specified.
3. Set **bootstrap=True** and **OOB score = True** (out-of-bound score) to ensure that model is trained on randomly sampled subset of entire dataset and we get training score for remaining part of dataset.

13. Business Threshold & Evaluation

Average industry thresholds for churn:

- Recall - 80%
- Precision – 70%

Current deficiencies – lack of data representing churn categories

Business Evaluation

- Identifying the customer segments and their behaviors that lead to churn you can design recommendations for your team to increase customer retention. Designing, implementing, and inferring results from smart experimentation and marketing tactics is a topic of debate too.
- For a telecom company experiencing churn with low income demographics who are using more texts than actual telephone calls, it may be about creating a niche 'plan' targeted to that segment to prevent the users from switching to the next provider.
- Customers with a low probability of churning can be removed from re-targeting lists, this could lead to cost saving in marketing.

- There are customers who are sensitive to price, especially low-income ones with dependents and need monitoring

14. Solutions To Reduce Churn Rate

Acquiring a customer is far more costly than keeping a customer. Any company that wants to retain its customers should find some value in analysing and lowering down the churn rate. Even emerging markets, which witnessed high growth in the past, are now looking to consolidate their customer base and differentiate themselves from their peers to reduce churn rates.

Telecom players use a variety of different metrics to determine when customers are about to leave. It is profitable for companies to explore the reasons why customers are leaving, and then target at risk customers with enticing offers. There are several different tactics companies use to maintain their customer bases [1]

- One of the most important is simply providing **efficient customer service**. Providing clients with an easy way to get questions answered and issues handled is the key to maintaining cellular clients.
- **Value- added services** serve as a subscriber retention tool, especially for established players. While for newer entrants, it will become a part of the marketing strategy to attract customers. If VAS providers leverage the opportunities to tie up with operators, there could be a major increase in the uptake of their services.
- **Offer upgrades** on the client's existing account. Expanding on services offered and giving better rates or discounts to the client often improves customer retention rates.
- Another tactic is offering free access or **reduced rates** on smartphone applications. The increasing regular use by customers of cell-phone applications makes free access to such applications an enticing bonus for many customers.
- Competing cellular providers aggressively market **special deals** to churn customers away from their current provider. Common practices include offering free phones and buying out any existing service contract. The cellular service business is highly competitive and will likely remain so; therefore, churn rates will continue to be an important focus for cellular providers.
- Fighting wireless churn with trendy smartphones and **fast data network**
- **One-on-one marketing campaigns** is one of the best tactics to reduce churn rate. Make sure that customers are communicated the new services offering based on their usage analysis and trends and should be given proactive information on the plans which will benefit the customer.

15.0 Conclusion

Our team had wonderful time during this project journey. We had lot of brainstorming sessions to discuss and confirm the end-to-end application development activities like project scope, project requirement, application architecture & development approach. Also, we had picked up required domain expertise and knowledge of various machine learning algorithms to build this system.

The purpose of this type of analysis using machine learning module is to increase the customer base and consequentially improve the revenue.

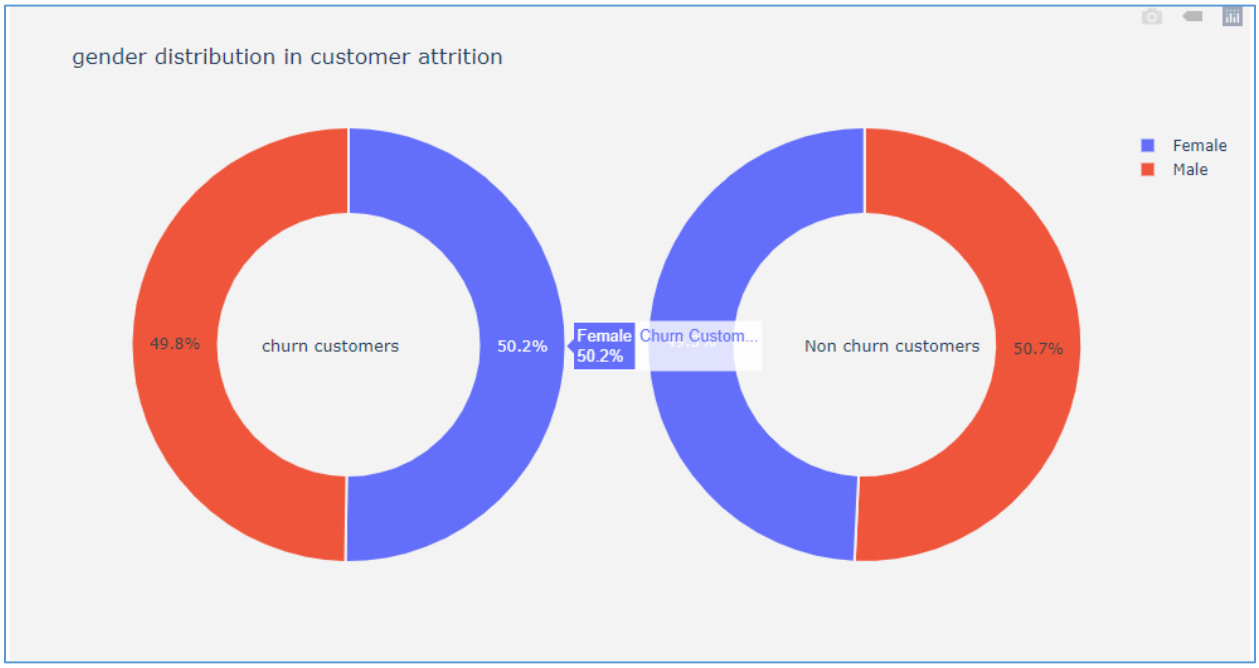
16.0 Reference:

1. 12 ways to stop churn immediately: https://www.superoffice.com/blog/reduce-customer-churn/

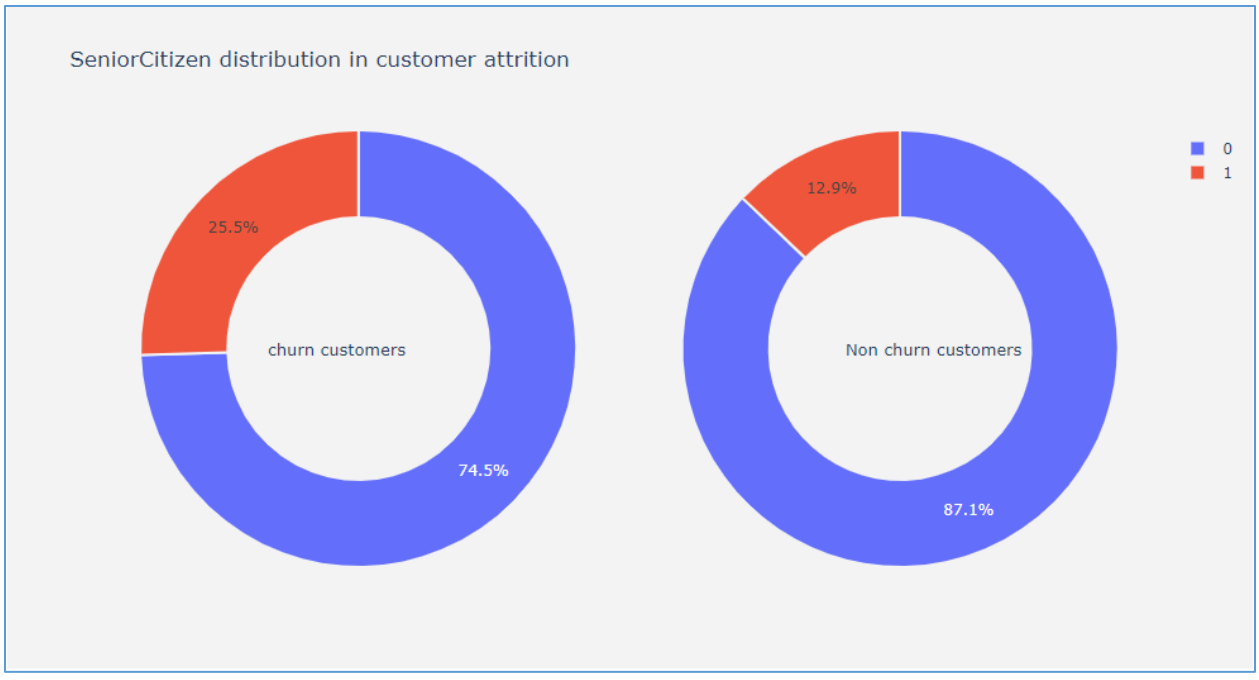
2. What is customer churn: https://www.bonjoro.com/blog/post/what-is-churn-why-does-it-matter

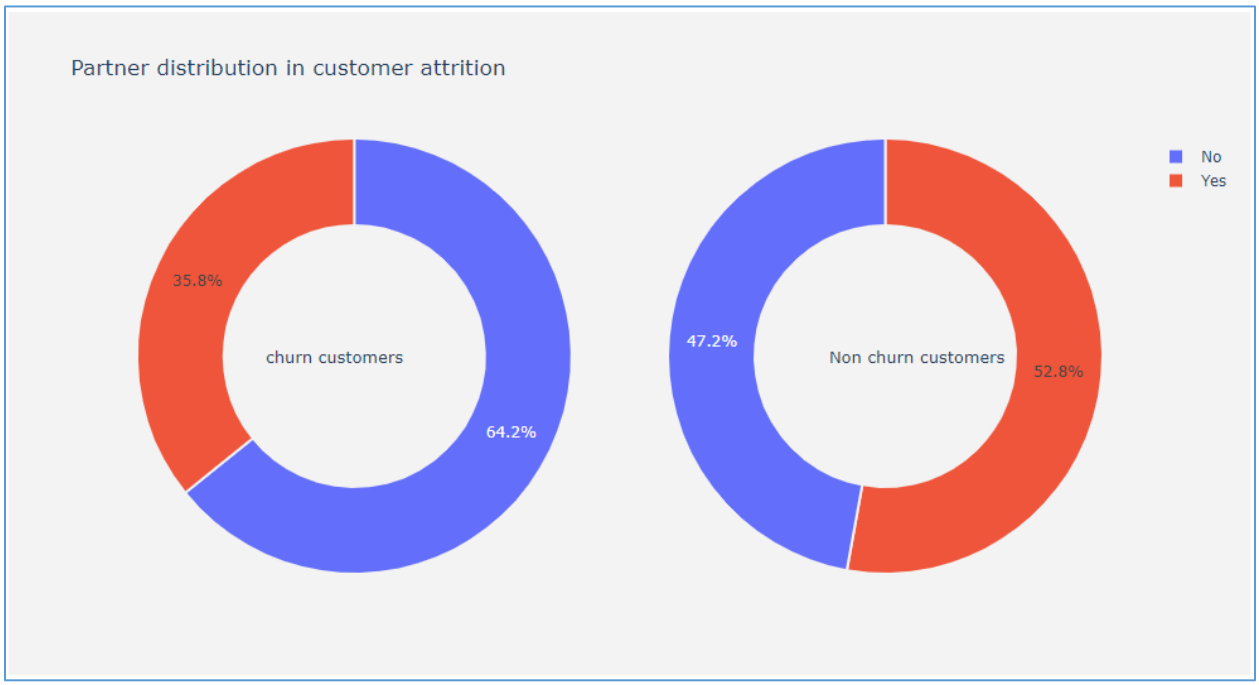
17. Appendix A Exploratory Data Analysis

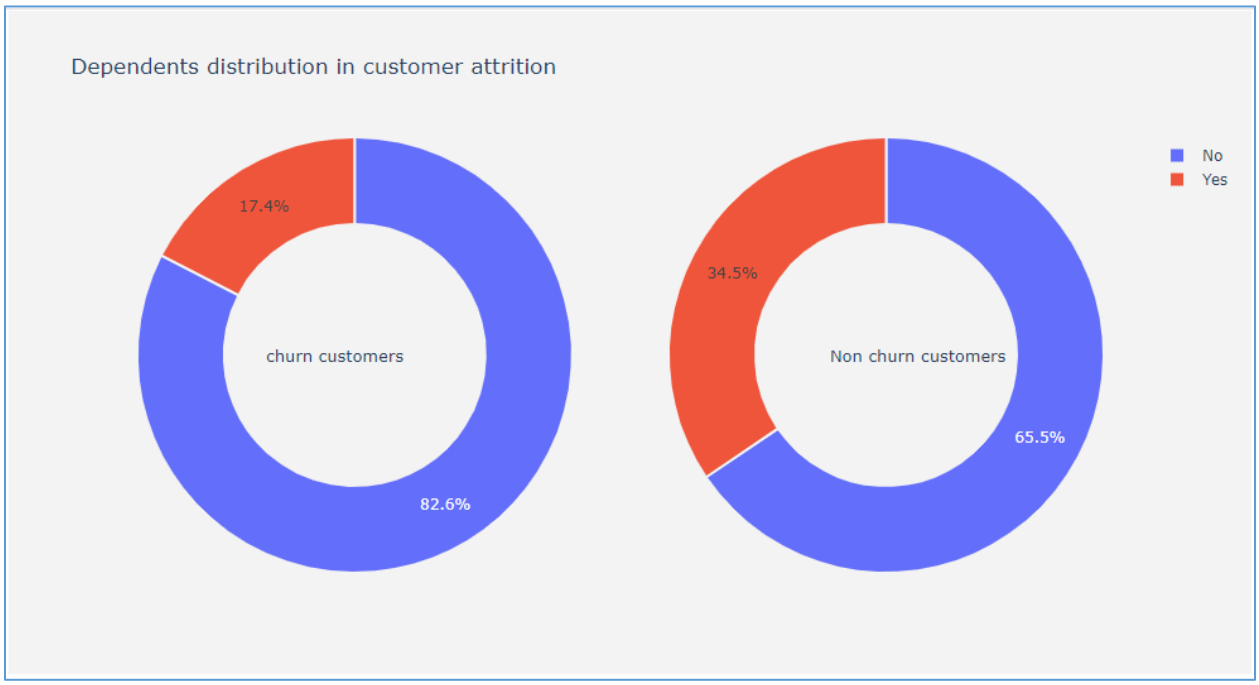
Customer Attrition based on categorical influencers

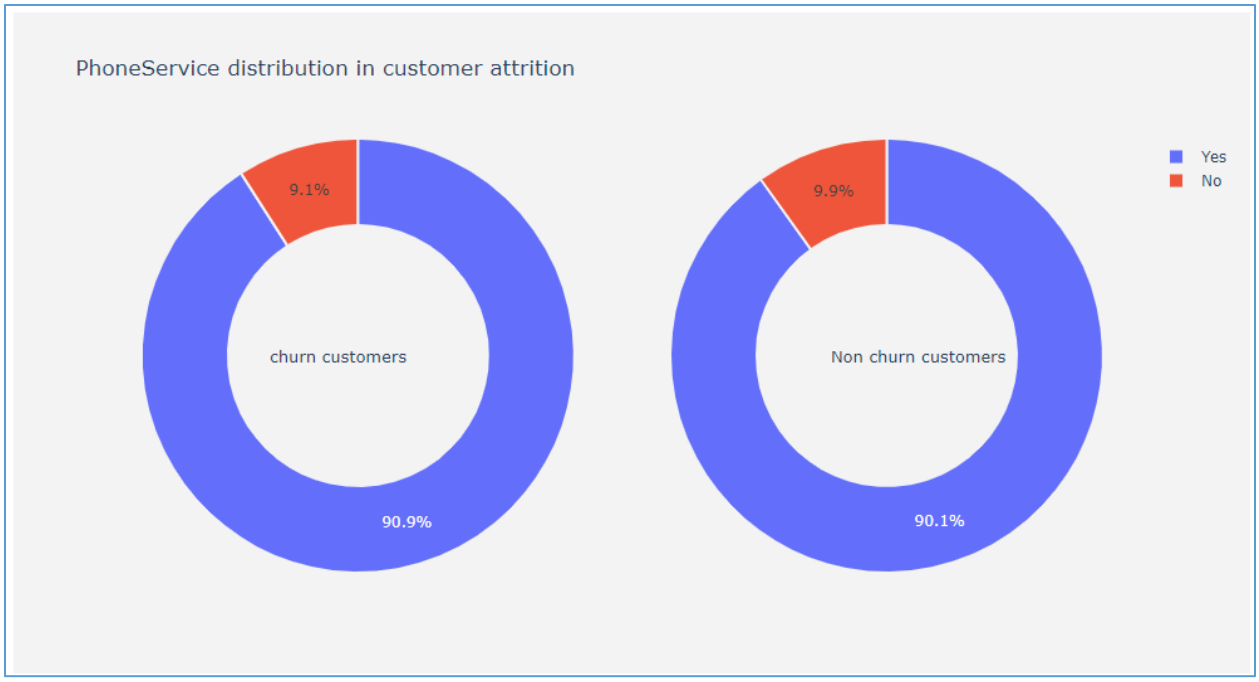


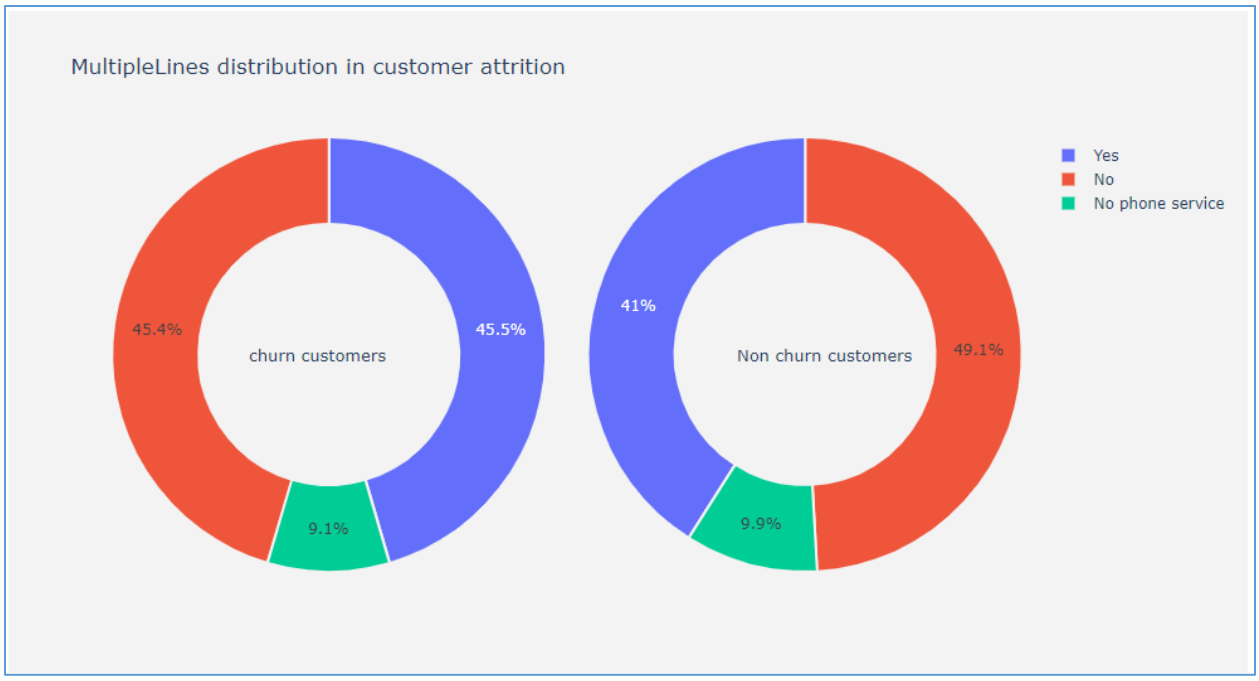
50-50 percent chance of a male or female each appearing for both churn and non-churn customer categories.

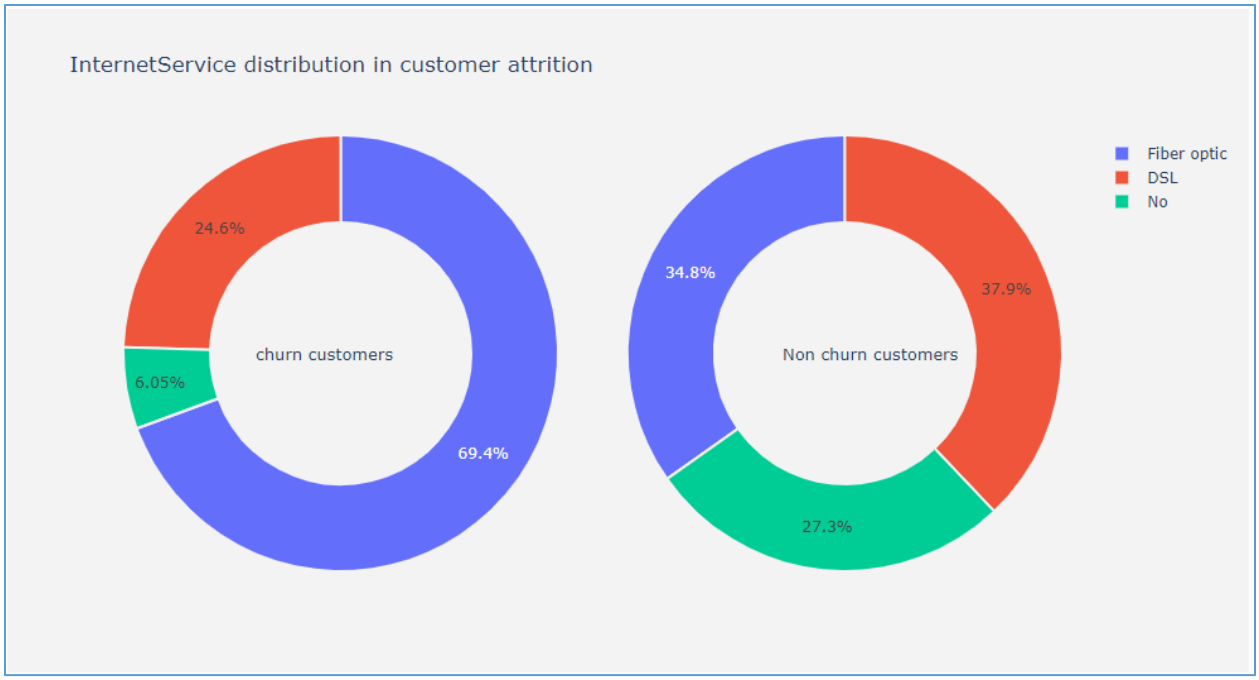


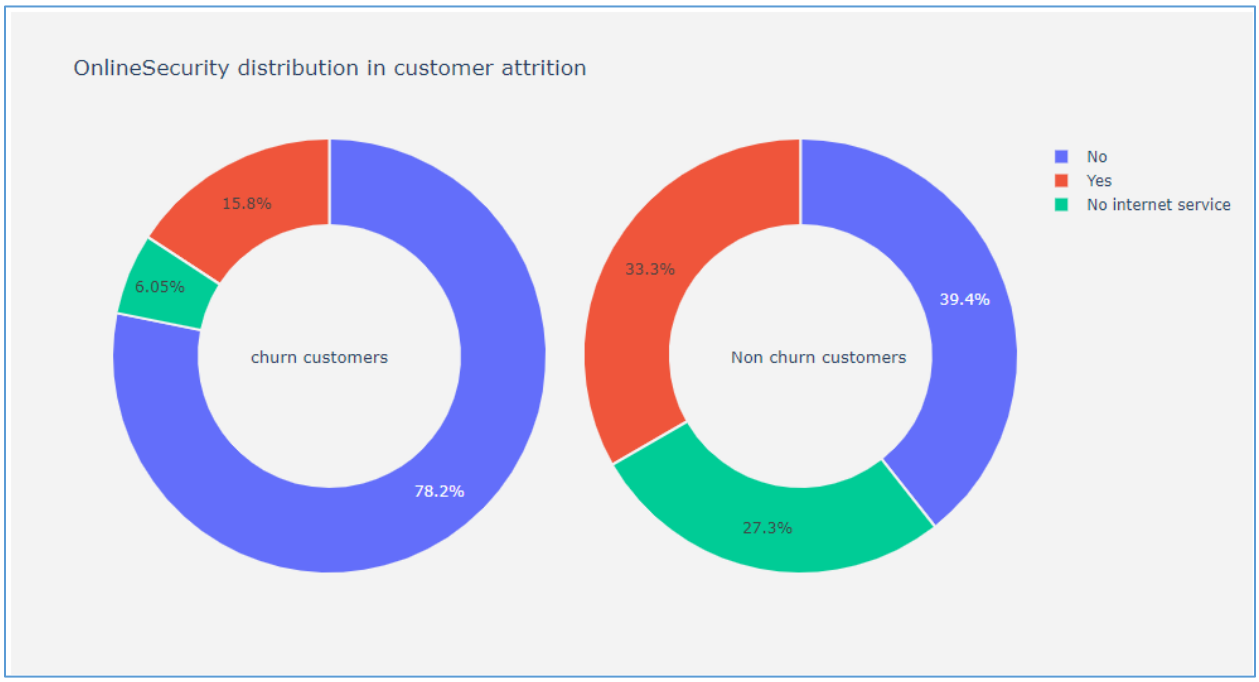

Partner distribution in customer attrition
35.8%
churn customers
64.2%
47.2%
Non churn customers
52.8%
No
Yes

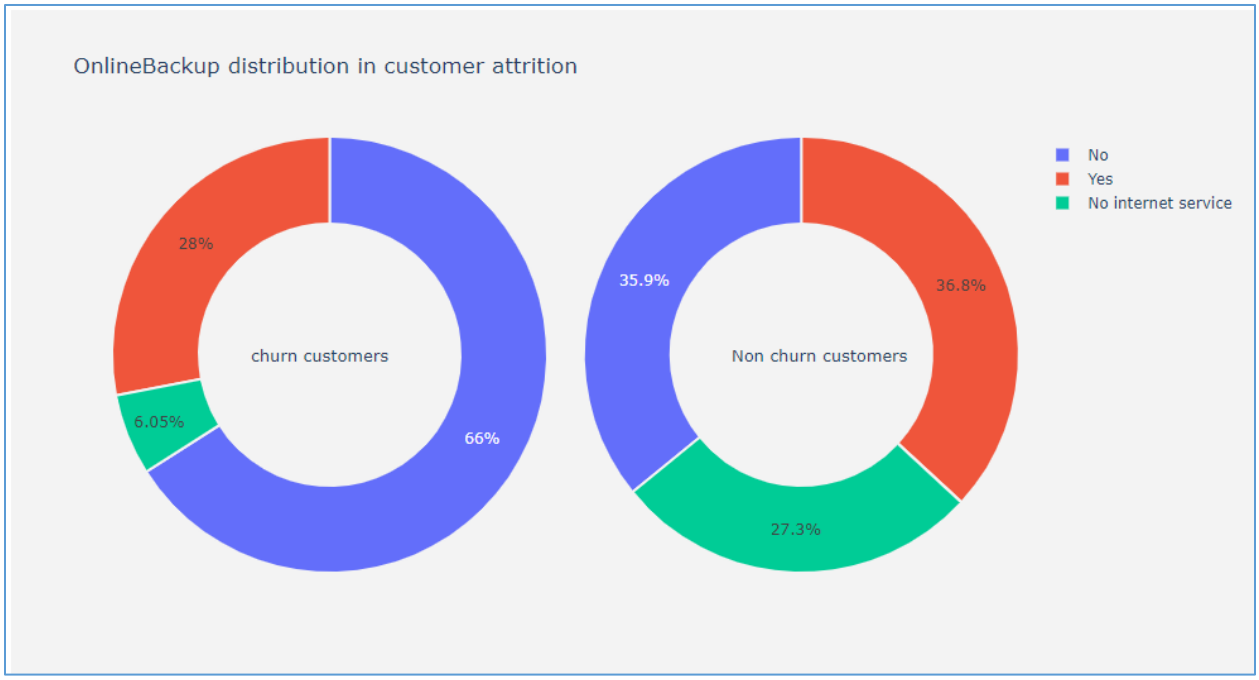

Dependents distribution in customer attrition
17.4%
churn customers
82.6%
34.5%
Non churn customers
65.5%
No
Yes

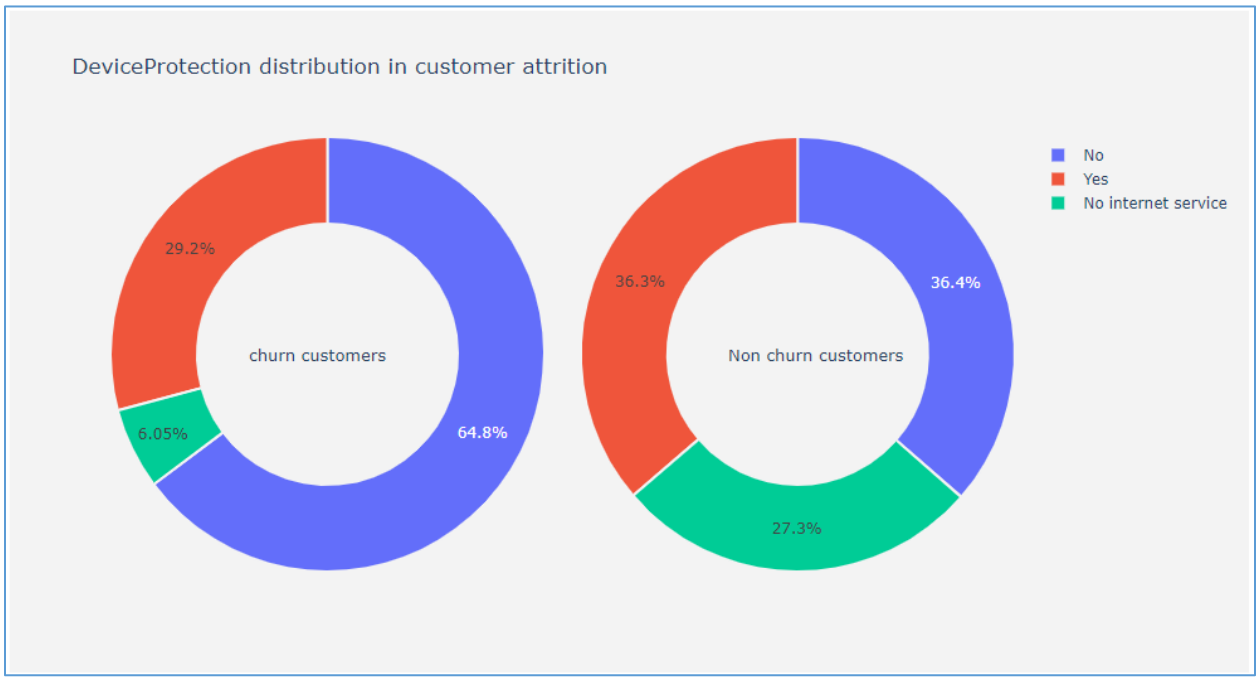

PhoneService distribution in customer attrition
9.1%
churn customers
90.9%
9.9%
Non churn customers
90.1%
Yes
No

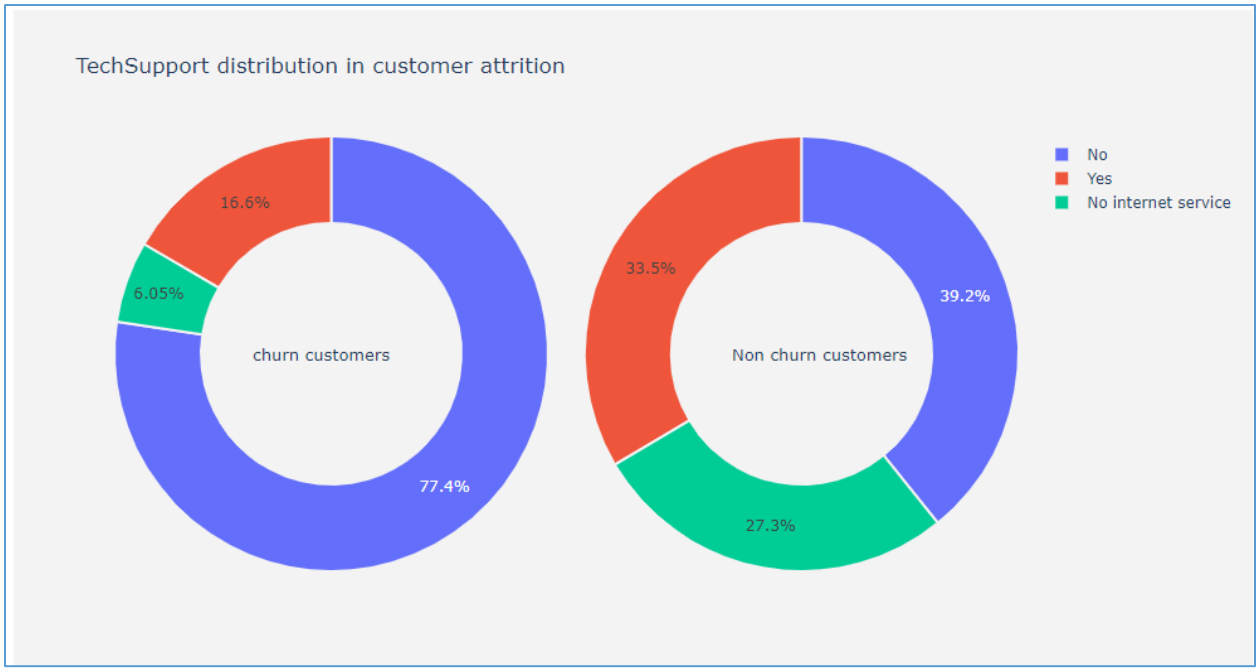

MultipleLines distribution in customer attrition
45.4%
45.5%
churn customers
9.1%
41%
49.1%
Non churn customers
9.9%
Yes
No
No phone service


InternetService distribution in customer attrition
24.6%
churn customers
6.05%
69.4%
34.8%
37.9%
Non churn customers
27.3%
Fiber optic
DSL
No

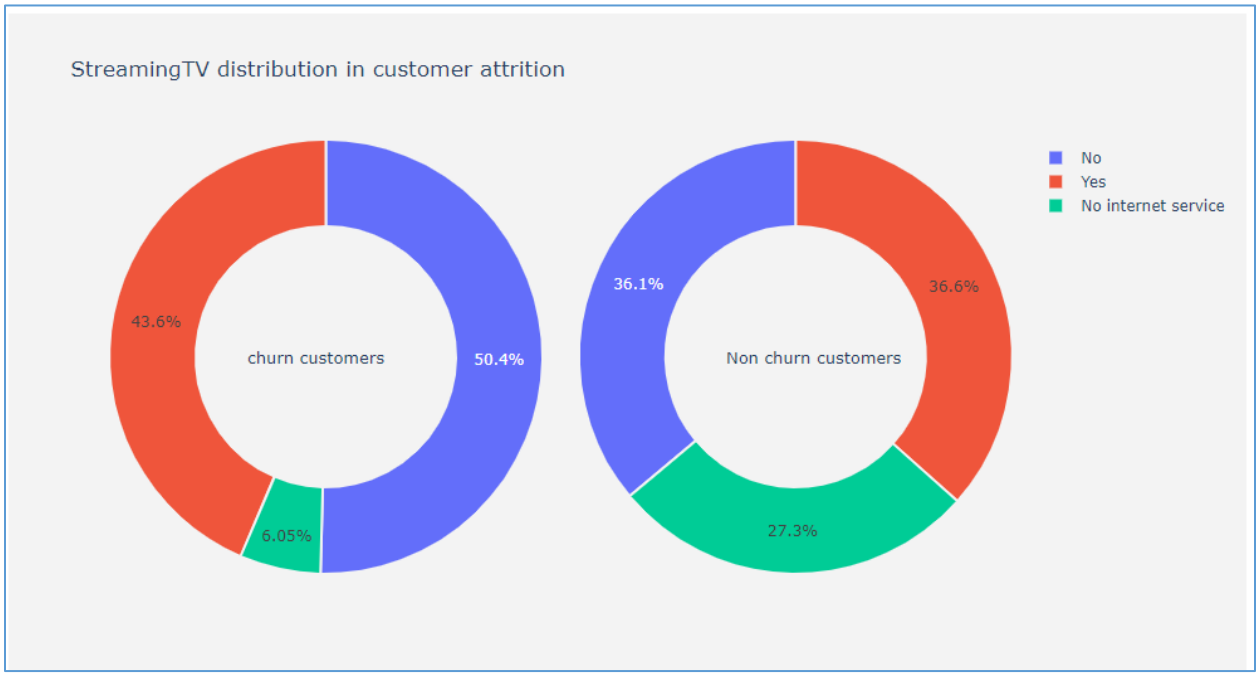

OnlineSecurity distribution in customer attrition
15.8%
6.05%
churn customers
78.2%
33.3%
39.4%
Non churn customers
27.3%
No
Yes
No internet service

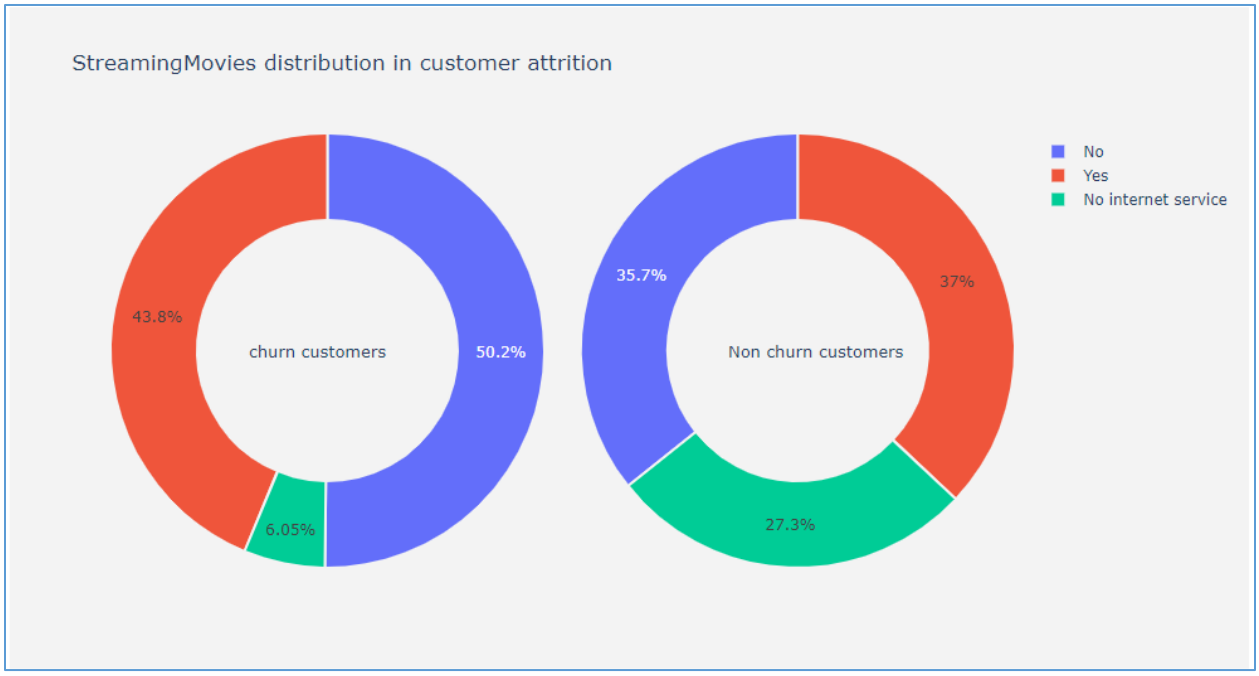

OnlineBackup distribution in customer attrition
28%
6.05%
66%
churn customers
35.9%
36.8%
27.3%
Non churn customers
No
Yes
No internet service

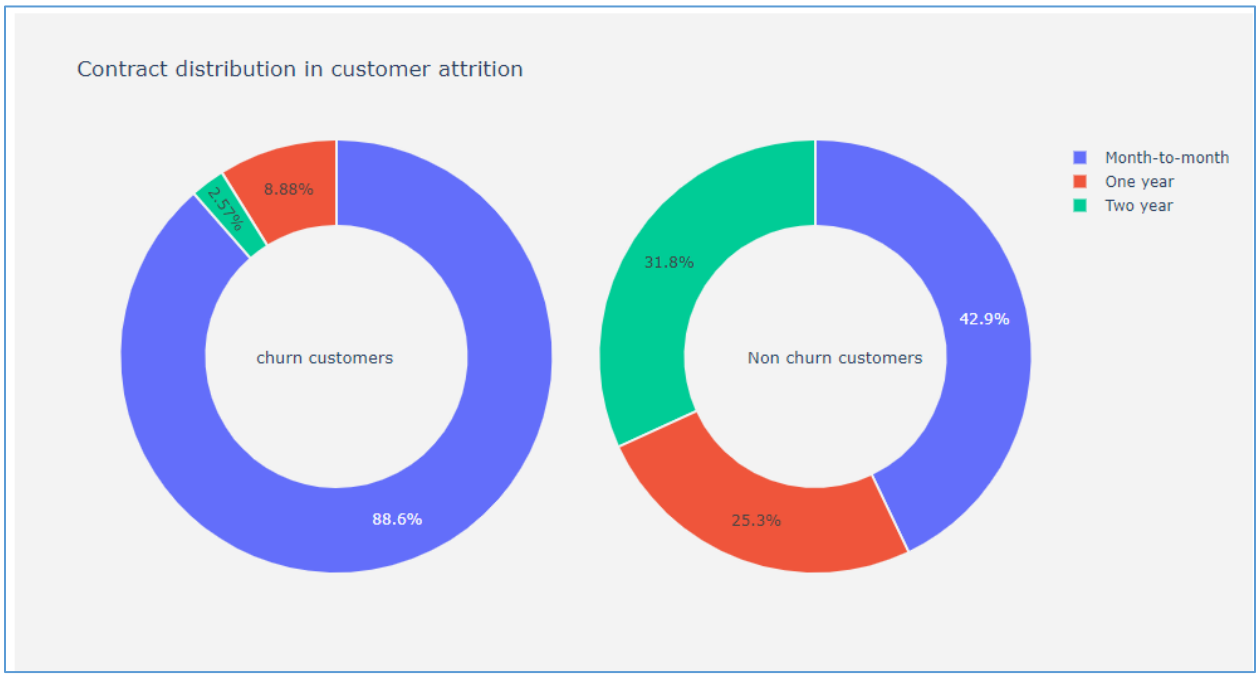

DeviceProtection distribution in customer attrition
29.2%
6.05%
64.8%
churn customers
36.3%
36.4%
27.3%
Non churn customers
No
Yes
No internet service

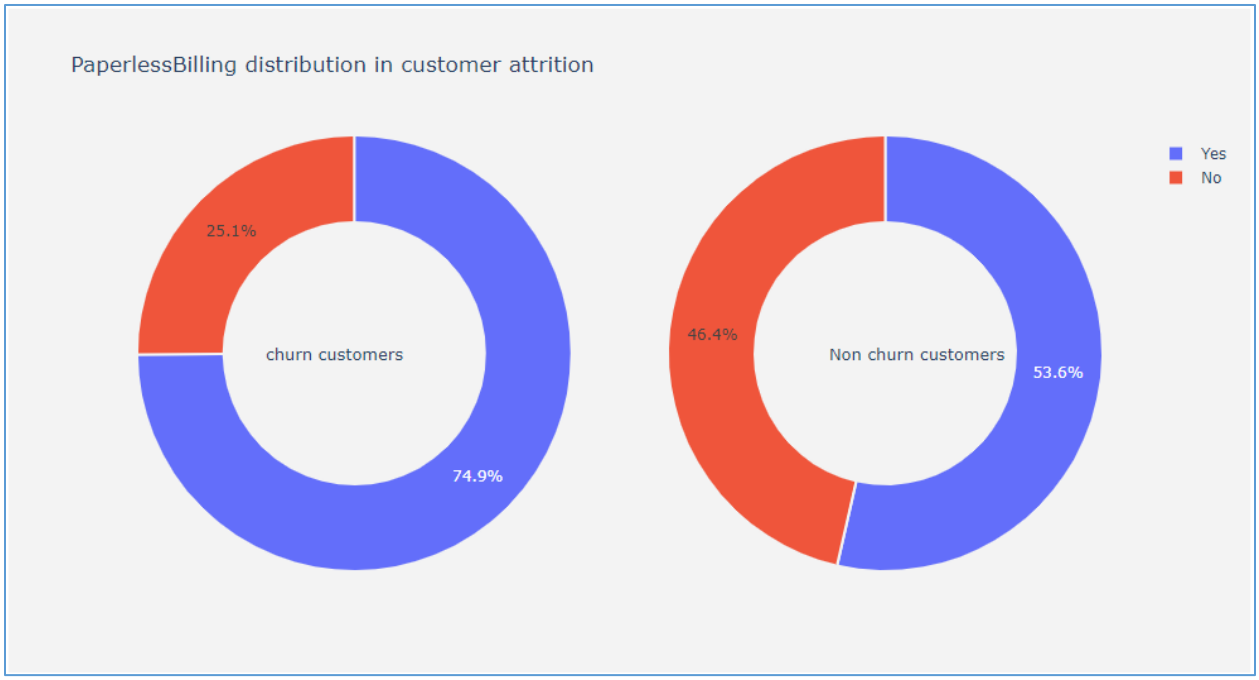


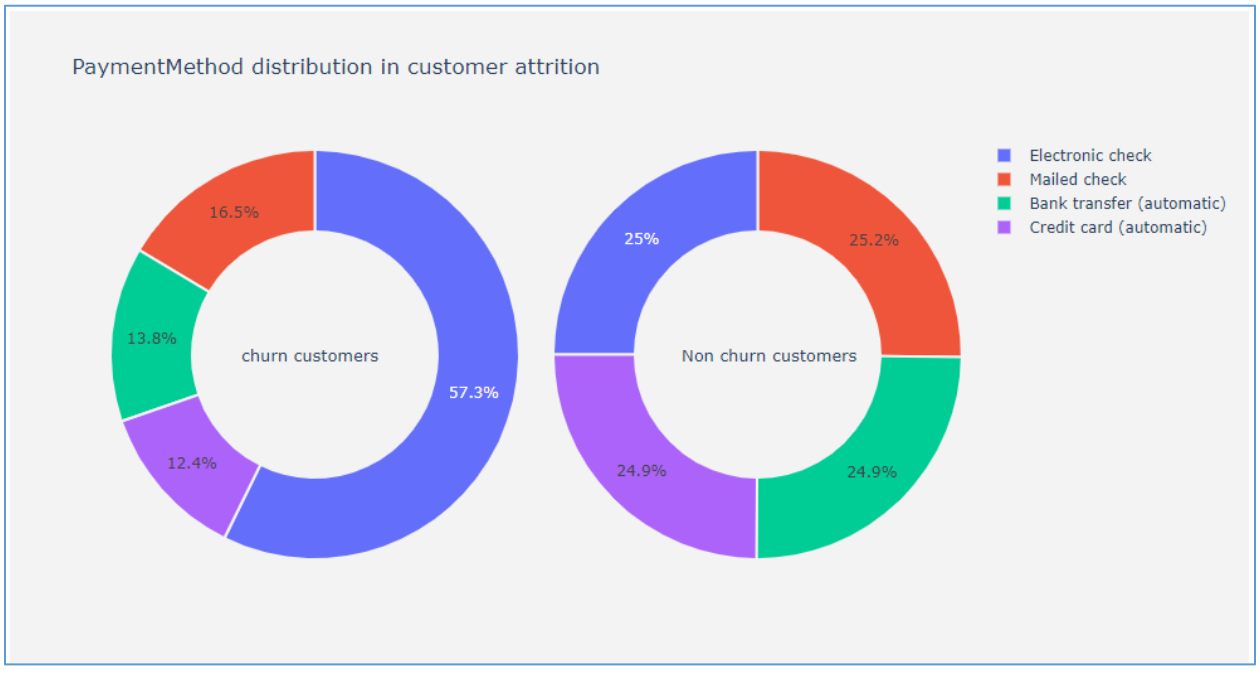
Availability of internet required to enjoy online related services.

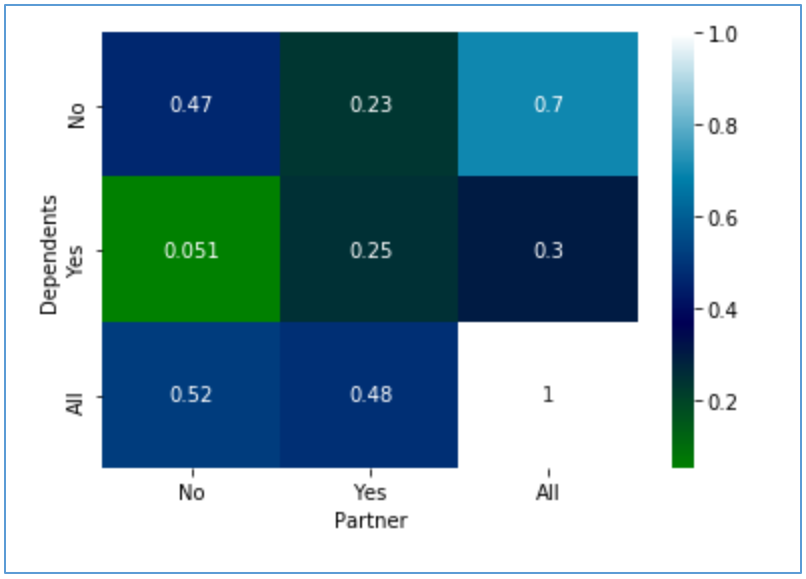



StreamingMovies distribution in customer attrition
43.8%
churn customers
50.2%
6.05%
35.7%
Non churn customers
37%
27.3%
No
Yes
No internet service

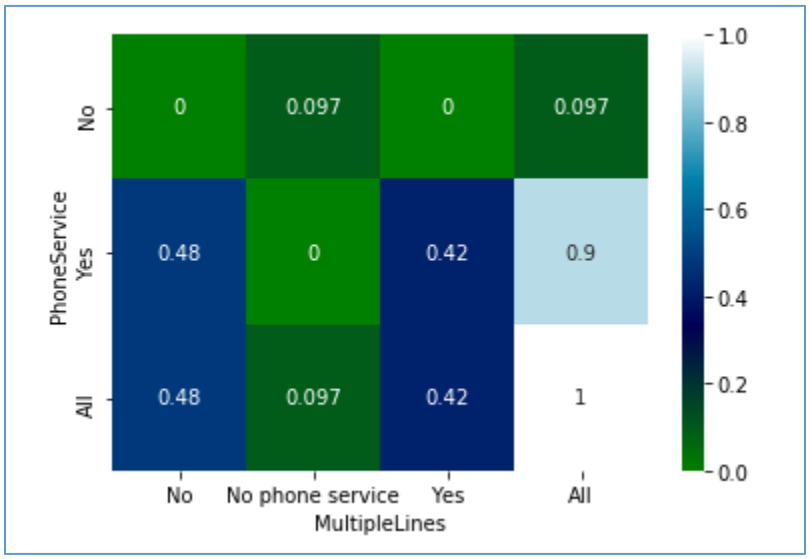

Contract distribution in customer attrition
8.88%
2.57%
churn customers
88.6%
31.8%
Non churn customers
42.9%
25.3%
Month-to-month
One year
Two year


PaperlessBilling distribution in customer attrition
25.1%
churn customers
74.9%
46.4%
Non churn customers
53.6%
Yes
No

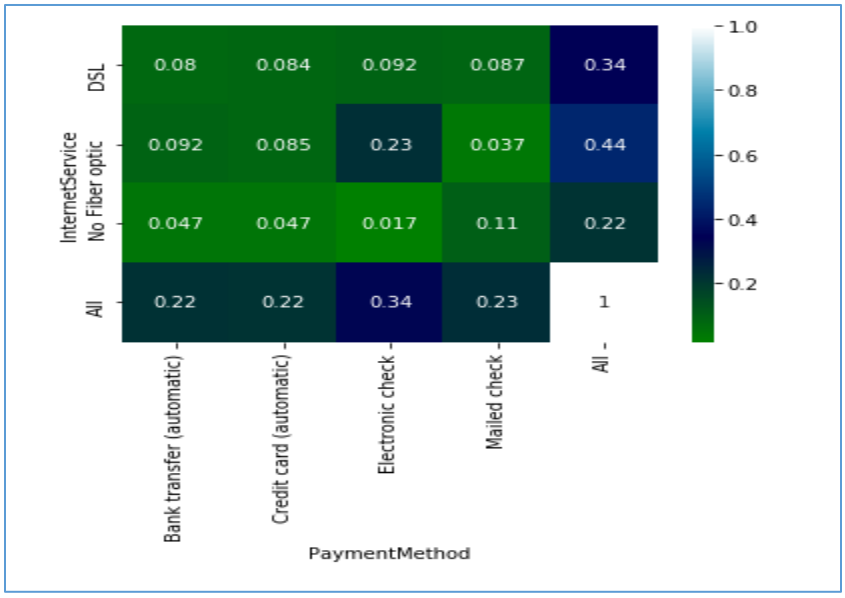

PaymentMethod distribution in customer attrition
16.5%
13.8%
churn customers
57.3%
12.4%
25%
25.2%
Non churn customers
24.9%
24.9%
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)

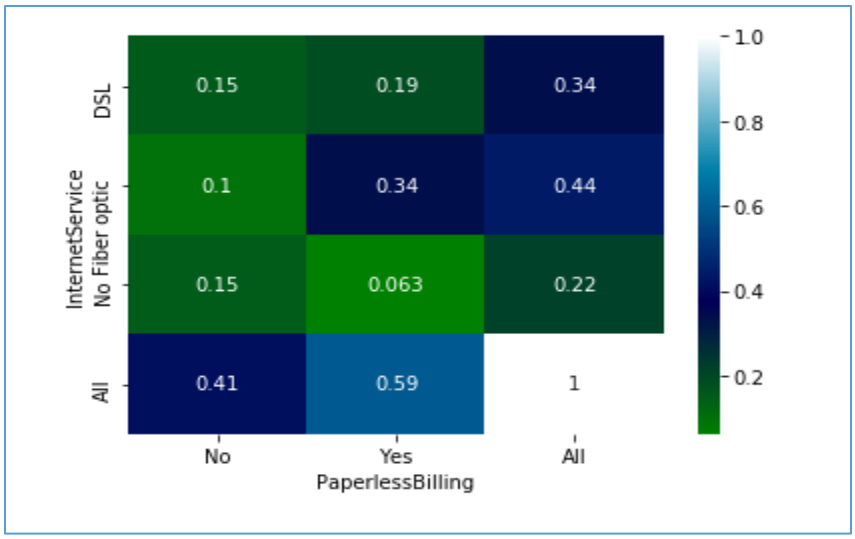


Those with partners usually have higher chance of having dependents. Those without partners usually don't have dependents.

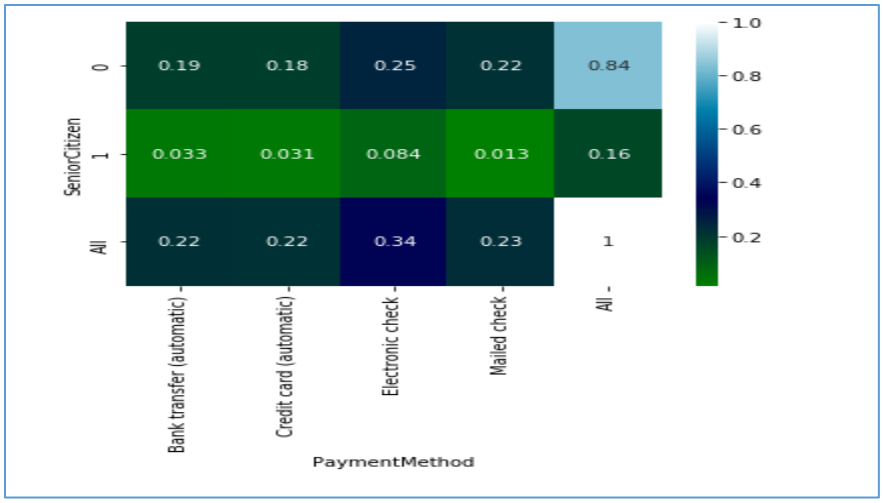


- Those with phone services have equal probability of having multiple phone lines.
- Multiple lines is not actually a strong predictor of churn but it can raise monthly costs that in turn affects probability

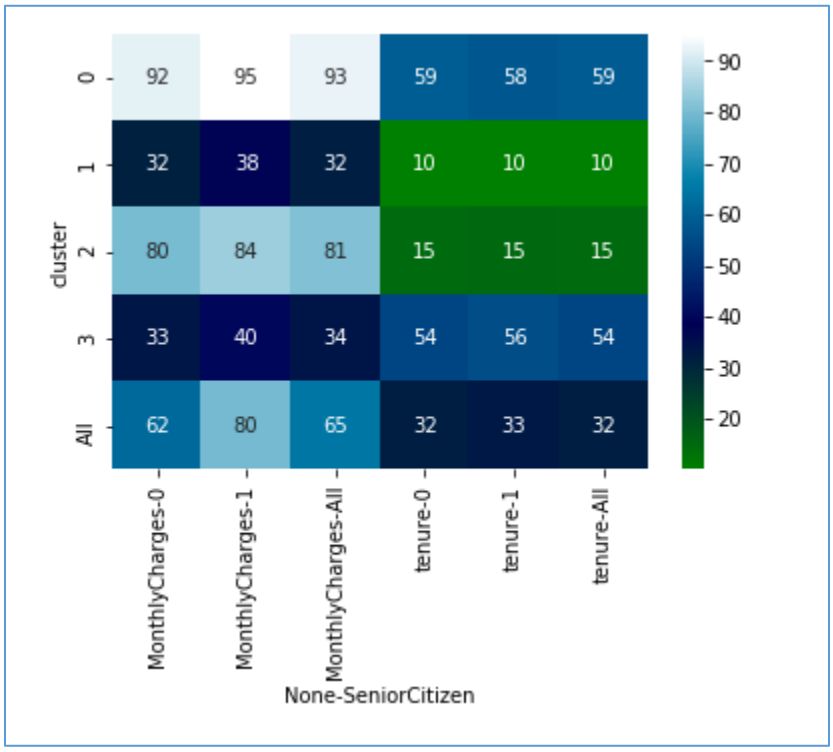


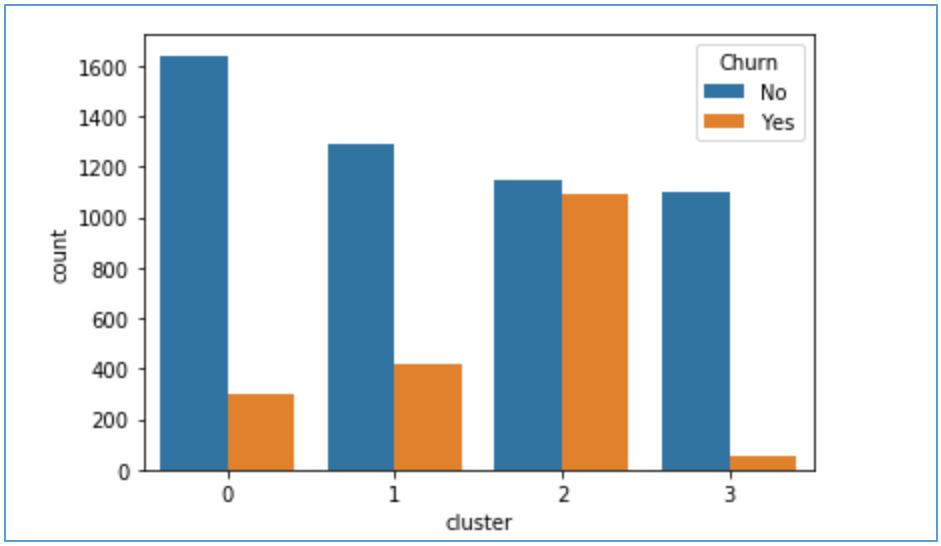


- People who opt for paperless billing tend to utilize internet service.
- Those with internet service have a clear-cut preference for automatic transfers, especially fiber optic subscribers. Those without internet services tend to use mailed check mostly.
- Among non-manual payment methods, electronic check most popular
- Those with fiber optic services utilize automatic transfer methods and these people are well versed with modern technologies
- Those with direct phone line internet services have equal probability of choosing automatic and check payment methods.



Cluster Analysis Based On Monthly Charges and Tenure



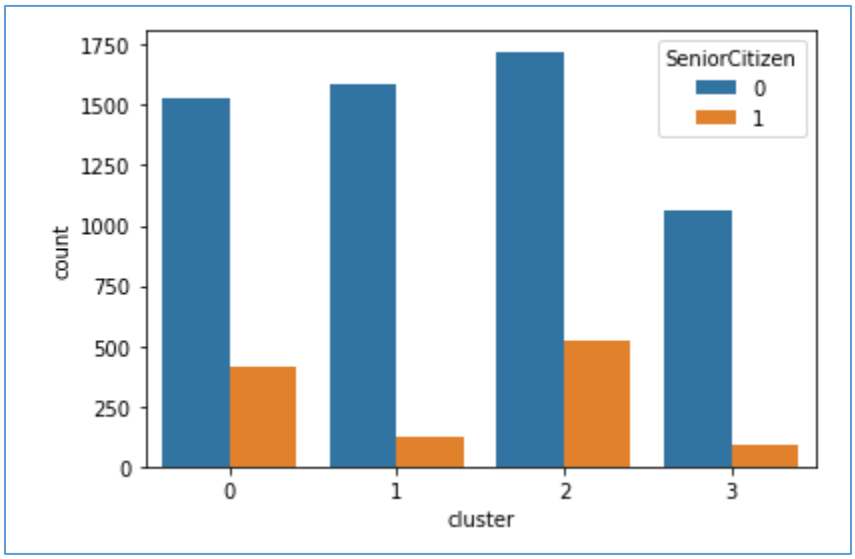


Clusters with descending order of churning probability: 2, 1, 0, 3

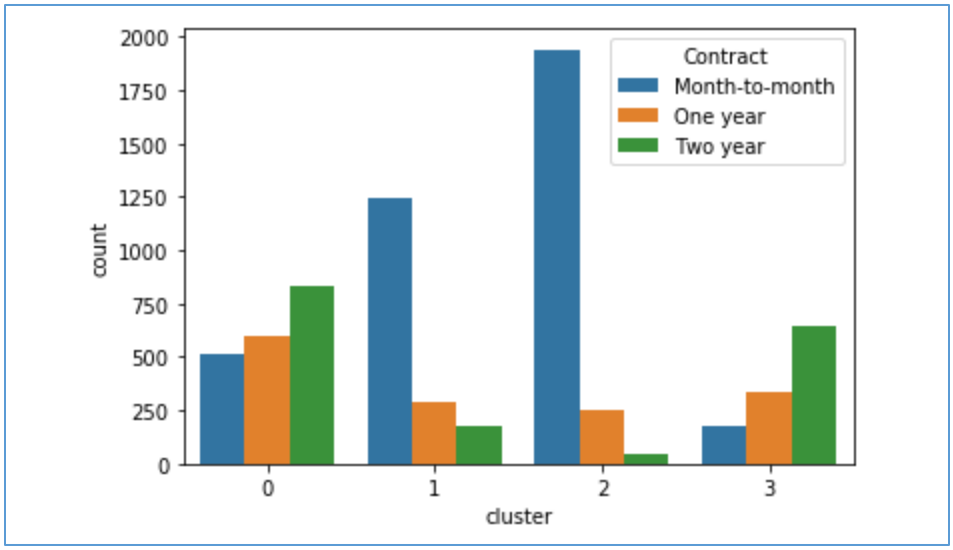
Cluster 3 is defined by high tenure and low, monthly charges, ideal for retaining customers.

Customers in category 2 (low tenure and high monthly charges), have highest probability of churning.

Cluster 0 customers have high tenure but high monthly charge. This shows monthly charge also an important predictor.

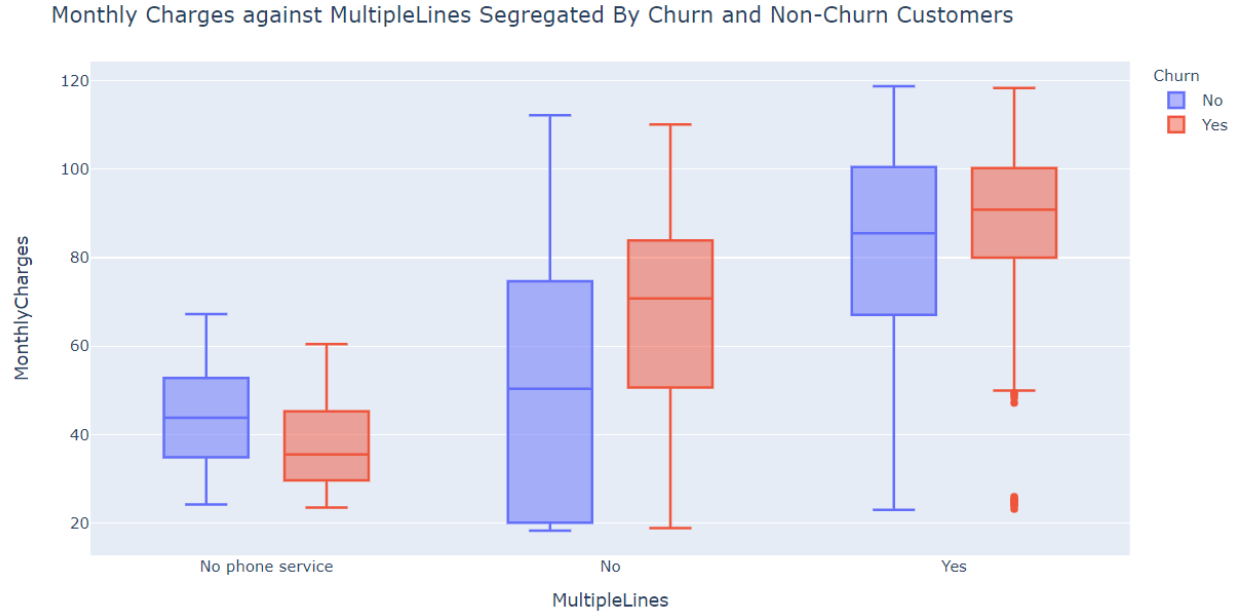


Senior citizens fall under clusters 0 and 2, clusters with high monthly charge. This means high monthly charge is problematic for senior citizens.

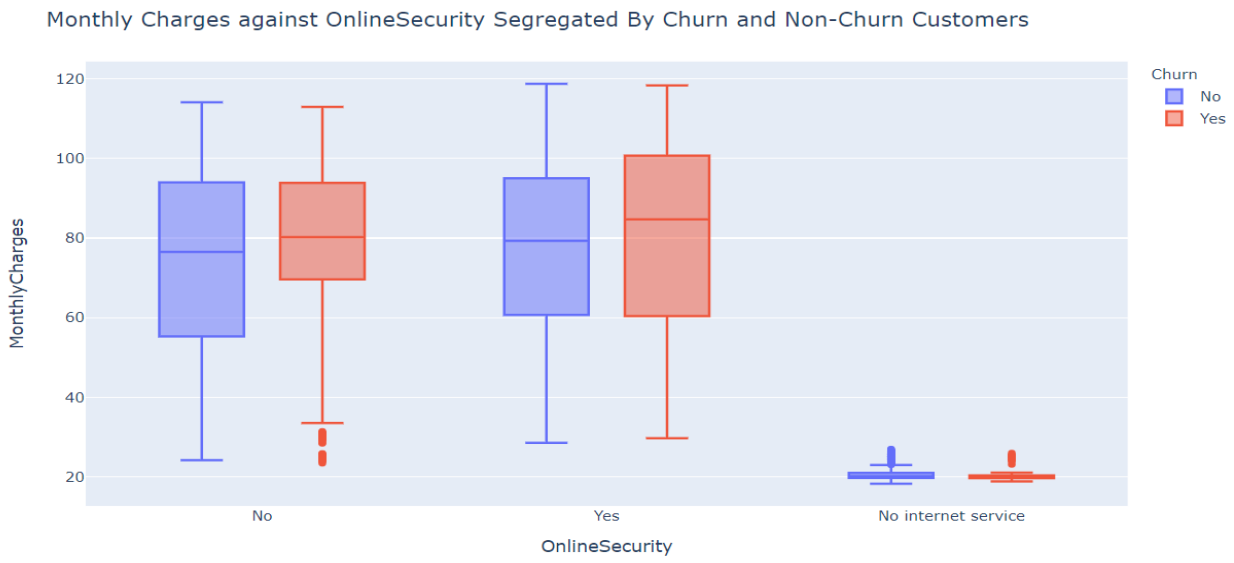


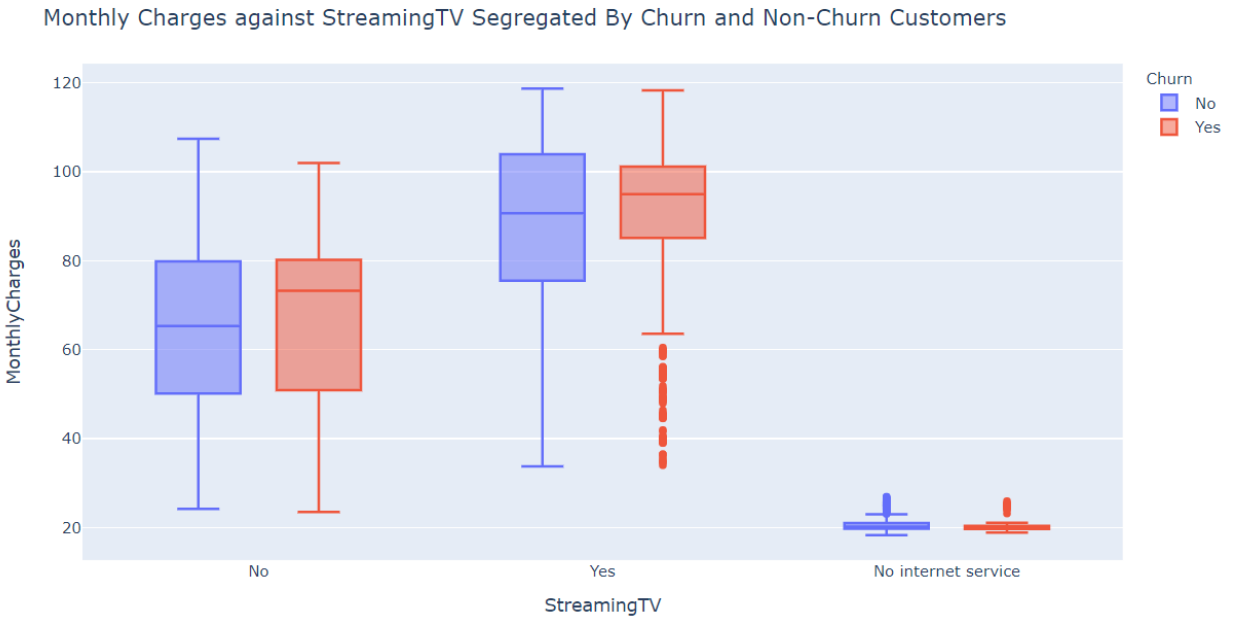
Customers of over two-year contracts are found in clusters with high tenure. Customers of month-to-month contract are found in clusters with low tenure and they cause churn significantly.

Boxplots Measuring Monthly Cost Against Amenities and Securities

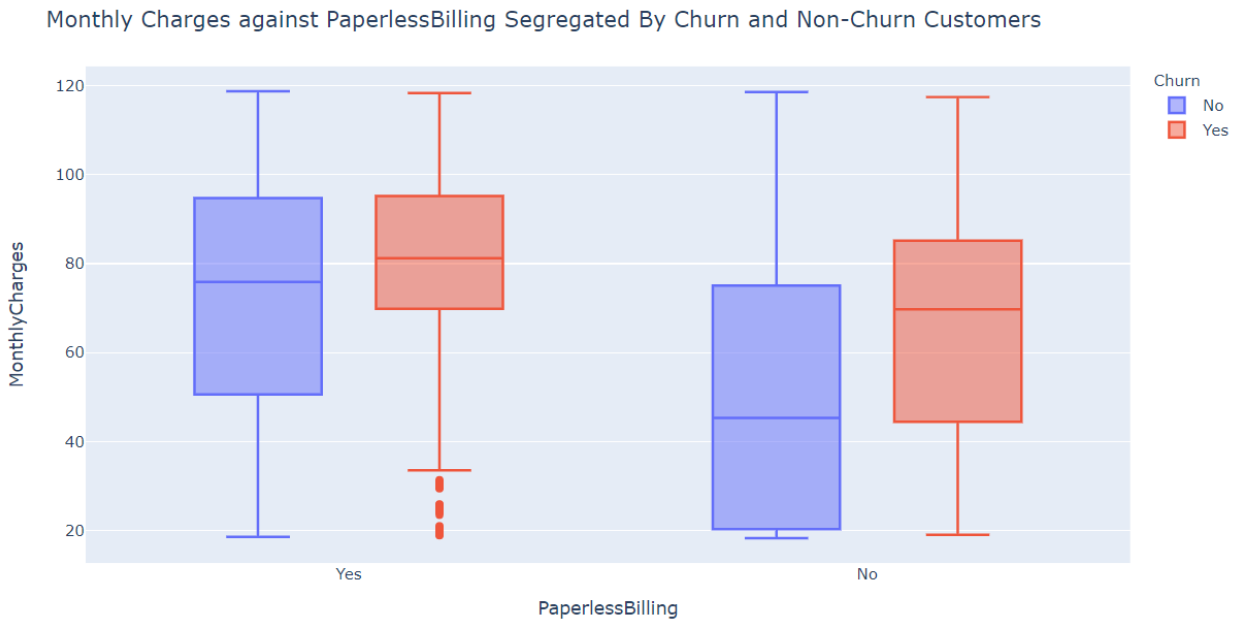


Presence of additional phonelines simply increase costs even though it may itself not influence churn probability according to pie charts.

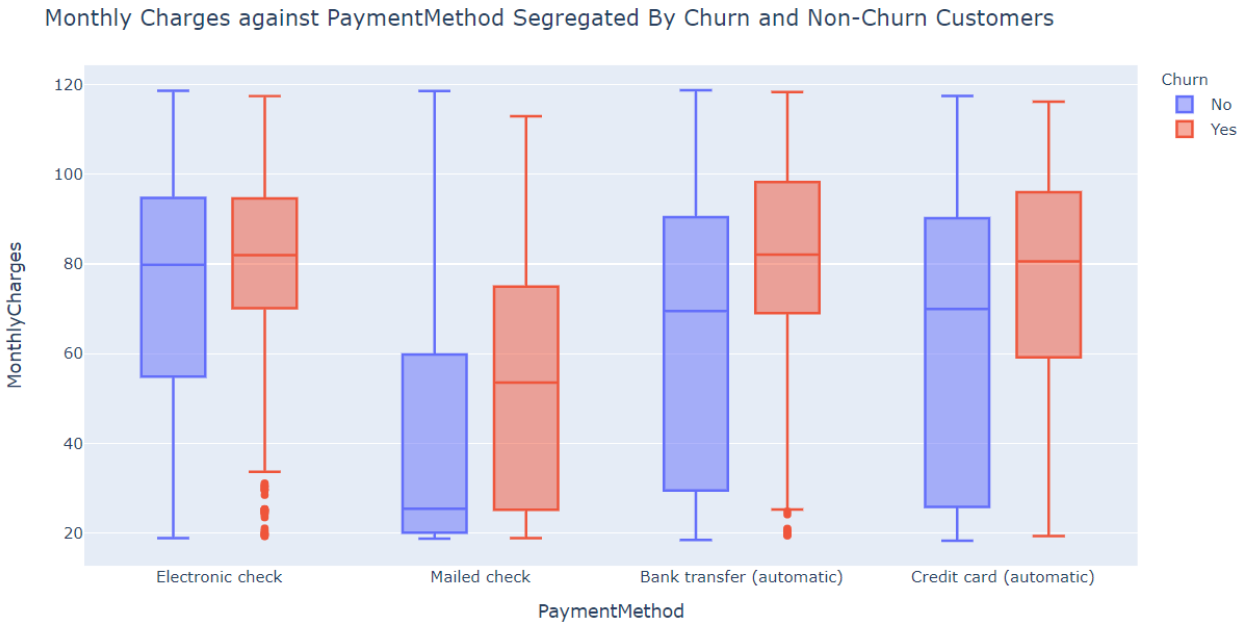




Without internet service, premium streaming and online services cannot be enjoyed and costs are lower too. Their presence drives up monthly cost and churn probability too. Lower interquartile range implies narrow spread and thus upward pressure on monthly price.

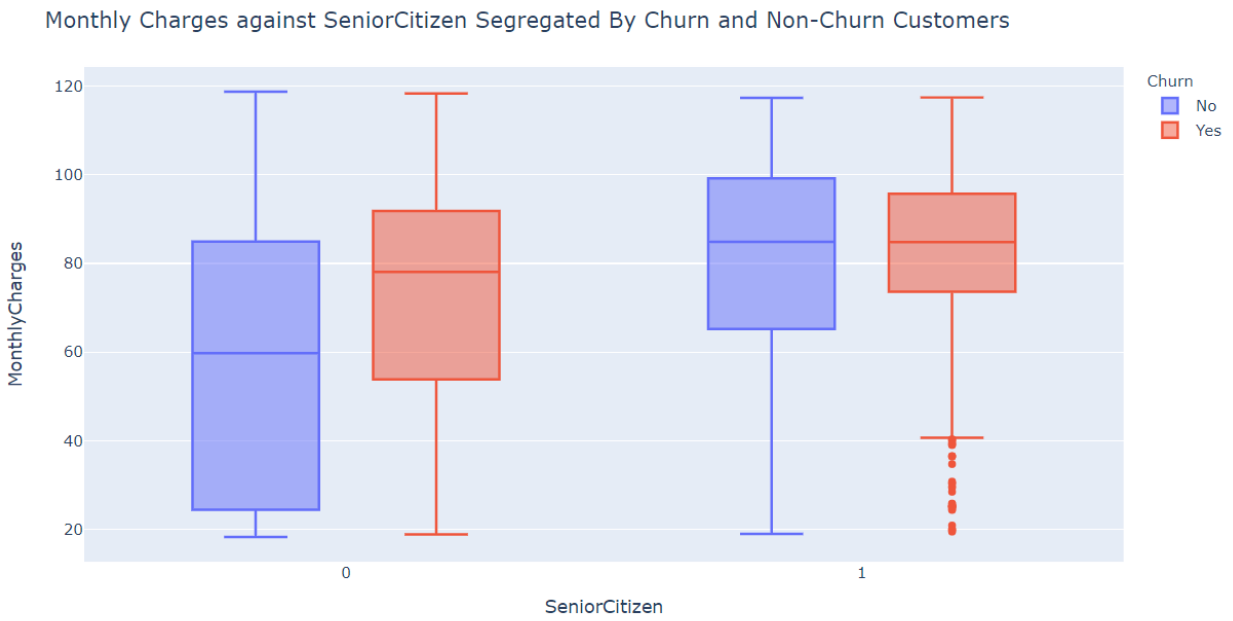


People who use paperless billing have higher mean but smaller range, not considering outliers. Spread is lower.



For non-churn customers, same proportion of users for all kinds of payment methods. Their mean monthly differs because of correlation/confounding effect with other variables like presence of multiple phone lines or streaming facilities.

Higher proportion of churn customers using electronic check than mailed check so they have higher monthly mean charge. Most of the people using automatic transfer methods are young people so the monthly mean used by these 2 groups are high too.



Senior citizens higher likelihood of churning than younger people, since their churn customers have higher monthly mean charges. The presence of outliers make the range of monthly charge similar to that of young people.

18. Appendix B Mapping Functions of System Architecture

Let's look at the skills/techniques utilized for this project.

- CRIPS - DM Process (Data Mining) → **Reasoning System (Data Mining)**
- K-Means Clustering Based On Monthly Charges and Tenure → **Machine Learning (Unsupervised)**
- Cluster Analysis Visualizations → **Reasoning System (Data Discovery)**
- Boxplots Measuring Monthly Cost Against Amenities and Securities → **Reasoning System (Data Discovery)**
- Feature Engineering Using If-Else Rules inferred from Visualizations → **Machine Reasoning (Forward Inference)**
- Supervised Machine Learning Models Training & Evaluation → **Machine Learning (Supervised)**
- Random Forest Initial Predictions (Chosen) Model → **Machine Learning (Supervised)**
- Extreme Gradient Boosting Classifier Predictions → **Machine Learning (Supervised)**
- Finetuned Random Forest Classifier Predictions (Final) → **Machine Learning (Supervised)**
- Random Search Optimization → **Reasoning System (Optimisation)**
- Annex A Exploratory Data Analysis & Knowledge Discovery → **Reasoning System (Data Discovery)**

19. Appendix C: GRADUATE CERTIFICATE - Intelligent Reasoning Systems (IRS) Project Proposal

Date of proposal: 10 April 2020

Project Title: Churn Fortuneteller

Sponsor/Client: *(Name, Address, Telephone No. and Contact Name)*

Academic Self Sponsored Project.

Background/Aims/Objectives:

Customer churn is one of the main factors that will affect the Telecom industry player's market share & the revenue. The Telecommunication industry mostly depends on subscription-based services. The profitability of the Organization mainly depends on its market share or its customer base. The customer acquisition and retention are two important factors that will directly impact the Organization's profitability. so, the customer churn is a major problem and one of the serious concerns for large companies in the Telecommunication industry.

Our aim is to create a system, Churn Fortuneteller. Which is fully focused to develop a model Machine Learning/Machine Reasoning/Optimization Techniques (Supervised Machine Learning models - Random Forest Classifier & XGBoost Classifier). The Churn Fortuneteller system would predict those customers about to leave in near future. Based on this prediction, the management will take appropriate actions to retain the customer base and improve the market share and revenue.

Using powerful analytic capabilities to uncover the insights that allow the management to take wise business or strategic decisions.

Requirements Overview:

To avoid / reduce the churn rating - develop prediction model based on machine learning - machine reasoning - random search optimization to improve model predictive performance – Python programming

Resource Requirements (please list Hardware, Software and any other resources)

Hardware proposed for consideration: NIL

Software proposed for consideration:
Python Web Development using Flask framework with Libraries

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Sklearn
- XGBoost Classifier
- Flask deployment for model prediction user-interface

Number of Learner Interns required: (Please specify their tasks if possible)

A team of four project members required to architect and implement this system.

Methods and Standards:

Procedures	Objective	Key Activities
Requirement Gathering and Analysis	The team should meet with ISS to scope the details of project and ensure the achievement of business objectives.	1. Gather & Analyze Requirements 2. Define internal and External Design 3. Prioritize & Consolidate Requirements 4. Establish Functional Baseline
Technical Construction	· To develop the source code in accordance to the design. · To perform unit testing to ensure the quality before the components are integrated as a whole project	1. Setup Development Environment 2. Understand the System Context, Design 3. Perform Coding 4. Conduct Unit Testing
Integration Testing and acceptance testing	To ensure interface compatibility and confirm that the integrated system hardware and system software meets requirements and is ready for acceptance testing.	1. Prepare System Test Specifications 2. Prepare for Test Execution 3. Conduct System Integration Testing 4. Evaluate Testing 5. Establish Product Baseline
Acceptance Testing	To obtain ISS user acceptance that the system meets the requirements.	1. Plan for Acceptance Testing 2. Conduct Training for Acceptance Testing 3. Prepare for Acceptance Test Execution 4. ISS Evaluate Testing 5. Obtain Customer Acceptance Sign-off
Delivery	o deploy the system into production (ISS standalone server) environment.	1. Software must be packed by following ISS's standard 2. Deployment guideline must be provided in ISS production (ISS standalone server) format 3. Production (ISS standalone server) support and troubleshooting process must be defined.

IRS Group 1 – Churn Fortune Teller

Team Name: GROUP 1
Project Title (repeated): Churn Fortuneteller.
System Name (if decided):
Team Member 1 Name: Anirban Kar Chaudhuri
Team Member 1 Matriculation Number: A0108517H
Team Member 1 Contact (Mobile/Email): Mobile: 86118180 Email: anirban.karchaudhuri@gmail.com
Team Member 2 Name: MARADANA VIJAYAKRISHNA
Team Member 2 Matriculation Number: A0178453W
Team Member 2 Contact (Mobile/Email): Mobile: 93896379 Email: mvskrishna@yahoo.com
Team Member 3 Name: Putrevu Manoj Niyogi
Team Member 3 Matriculation Number: A0213557E
Team Member 3 Contact (Mobile/Email): Mobile: 94575890 Email: manojniyogi@yahoo.com
Team Member 4 Name: Sivasankaran Balakrishnan
Team Member 4 Matriculation Number: A0065970X
Team Member 4 Contact (Mobile/Email): Mobile: 97379441 Email: bsivaa@gmail.com

Team Formation & Registration

<table>
<tr><th colspan="3">For ISS Use Only</th></tr>
<tr><td>Programme Name:</td><td>Project No:</td><td>Learner Batch:</td></tr>
<tr><td colspan="3">Accepted/Rejected/KIV:</td></tr>
<tr><td colspan="3">Learners Assigned:</td></tr>
<tr><td colspan="3">Advisor Assigned:

Contact: Mr. GU ZHAN / Lecturer & Consultant

Telephone No.: 65-6516 8021

Email: zhan.gu@nus.edu.sg</td></tr>
</table>

Appendix D : Individual Project Report

Project Title: Churn Fortuneteller

Team Member 4 Name: Kar Chaudhuri Anirban

Team Member 4 Matriculation Number: A0108517H

Journey in Churn Fortuneteller

It is a great pleasure and wonderful opportunity to form a team from diversified backgrounds with the right, balanced mixture of talents to kick start the Churn Fortuneteller project as part of our semester project. Every team member helped one another in sharing their expertise, knowledge, and experience to sprout the thoughts on planning and execution of the project.

Due to the current circuit breaker we could meet twice face to face and then decided to continue our journey using zoom calls for sharing ideas and follow up on our individual work with each other throughout the project development time. We four of our project members discussed and given well thought and come to a consensus right from choosing the domain of the project and formulating the project plan of analysis, design, development.

I helped by coordinating research, having candid discussions, and giving own inputs of ideas which enables us to successfully complete the project in time. I architected the machine reasoning rules and machine learning models too in python. I had important share in documenting main report, for both data analysis visualizations and their interpretations, explaining interpretability of machine learning models as well.

Acquire Domain Knowledge and Machine Learning Algorithms Principles:

Even though we initially did not have experience in telecom customer management domain we patiently spent significant portion of time allocated to this project in understanding this business by picking the required business knowledge to understand analyze the data and make necessary decisions and argument to justify in selecting the churn variables.

I was reading about various services offered by different competitors in the market related to premium services, marketing strategies and price sensitivities of different segments of customers. I explored metrics/key-performance-indicators for measuring customer churn and possible solutions to reduce customer churn Some of us having a decent understanding on customer and their churn patterns from different industry which helped us correlate to the telecom industry as on plus point.

We proceeded onto reading about various machine learning models, especially tree -based models like decision trees, random forest and XGBoost. I enabled other team members to understand the principles by using simple visual aids and flow diagrams of bagging and boosting algorithms, as well as brute force vs random search optimization. Lastly, real understanding took place during simulations by running and finetuning python codes.

Learn new technologies:

i) I learned and became familiar in web development using html and css

ii) Learnt deployment of micro web services using Python Flask library

iii) Explored and discovered ways for writing efficient codes in Python as well as testing runtime

My Involvement in Churn Fortunetelling Project

I am involved in the application development of Churn Fortuneteller project from end-to-end with following stages of application development lifecycle.

- Application scope & design
- Requirement Analysis
- Data gathering, analysis and visualizations
- Developing, testing, and optimizing machine learning model
- Application Development
- Code review
- Documentation
  - Project Proposal
  - Churn Fortuneteller Project Report
  - Prepared the Churn Fortuneteller presentation slides
- Verification activities – documentation and coding

Project Title: Churn Fortuneteller

Team Member 3 Name: Putrevu Manoj Niyogi

Team Member 3 Matriculation Number: A0213557E

Journey in churn fortuneteller

It is a great pleasure and wonderful opportunity to form a team from diversified backgrounds with the right, balanced mixture of talents to kick start the Churn Fortuneteller project as part of our Semester project. It was fun working with the team and learning at the same time. Each of the project member helped others in sharing their expertise, knowledge and experience to sprout the thoughts on planning and execution of the project.

We four of our project members discussed and given well thought and come to a consensus right from choosing the domain of the project and formulating the project plan of analysis, design, development. We helped each other during the project phase by having candid discussions and exchanging ideas which enables us to successfully complete the project in time.

Due to the current circuit breaker we could meet twice face to face and then decided to continue our journey using Zoom calls for sharing ideas and follow up on our individual work with each other throughout the project development time.

Acquire Domain Knowledge:

Even though team had no experience to telecom domain each individual spent a good amount of time in understanding business using online media and acquired fair amount of knowledge about different models and services offered by different competitors in the market. Some of us having a decent understanding on Customer and their Churn patterns from different industry which helped us correlate to the telecom industry as on plus point.

At the end we picked up the required business knowledge to understand analyze the data and make necessary decisions and argument to justify in selecting the Churn variables.

Learn new technologies:

iv) I learned and became familiar in development using Python and flask frame work.

v) Acquired the working knowledge of Python libraries applicable for Machine Learning projects.

vi) Learned and explored the data analysis / data discovery tools (Orange, Panda – Python Library).

vii) Upon acquiring the above knowledge, looking forward to work in machine learning / data analysis related projects.

My Involvement in Churn Fortunetelling Project

I am involved in the application development of Churn Fortunetelling project from end-to-end with following stages of application development lifecycle.

- Application Scope
- Application Design
- Requirement Analysis
- Data Gathering
- Data Analysis
- Application Development
- Code review
- Verification Activities
- Unit & final Testing
- Documentation
    - Project Proposal
    - Churn Fortuneteller Project Report
    - Prepared the Churn Fortuneteller presentation slides

Project Title: Churn Fortuneteller

Team Member Group 2 Name: MARADANA VIJAYA KRISHNA

Team Member Group 2 Matriculation Number: A0178453W

Journey in churn fortuneteller

We used Zoom video conference call platform to host our meetings once a week and as well whenever required to discuss regarding project. We had daily communication between ream members and shared research materials, case studies for overall improvement of the project. We helped each other and we all had clear goal to complete the Churn Fortuneteller project successfully on time.

Acquire Domain Knowledge:

Though we are new to telecom sector, the basic functionality of churn we used in our previous long experience has helped us a lot in deriving the business salability of the project scope.

Now we are better equipped with the telecommunication business knowledge and analyze their massive data and bring out valuable insights that will help the company to take better decisions.

Technologies learned & applied:

viii) I had picked up and familiar with web development using Python and its flask framework.

ix) Acquired the working knowledge of Python libraries applicable for Machine Learning projects.

x) Learned and explored the data analysis / data discovery tools (Orange, Panda – Python Library).

xi) Also learned and impressed with working and implementation of ASR, speech recognition and natural language processing technologies.

xii) Upon acquiring the above knowledge, looking forward to work in machine learning / data analysis related projects.

My Involvement in Churn Fortunetelling Project :

I am involved in the application development of Churn Fortunetelling project from end-to-end with following stages of application development lifecycle.

- Requirement Analysis
- Identification & Analysis of Data
- Scope definition, Design & Development of the Application
- Code verification & review
- Unit & final Testing
- Documentation
    - Project Proposal
    - Churn Fortuneteller Project Report
    - Prepared the Churn Fortuneteller presentation slides

Project Title: Churn Fortuneteller

Team Member 1 Name: Sivasankaran Balakrishnan

Team Member 1 Matriculation Number: A0065970X

Journey in churn fortuneteller

It is quite fun and really enjoyed working as a team. We all had good understanding and shared our knowledge and experience. We helped each other and we all had clear goal to complete the Churn Fortuneteller project successfully on time.

We had numerous face-to-face / Zoom / tele conversation sessions to discuss and confirm the project scope, selecting the technologies, application development tools and development methodology and run through the progress.

Acquire Domain Knowledge:

Our team is new to the telecommunication sector, so we had spent lot of time and effort to understand the business model and their service offerings to their customers.

Now we are better equipped with the telecommunication business knowledge and analyze their massive data and bring out valuable insights that will help the company to take better decisions.

Learn new technologies:

xiii) I had picked up and familiar with web development using Python and its flask framework.

xiv) Acquired the working knowledge of Python libraries applicable for Machine Learning projects.

xv) Learned and explored the data analysis / data discovery tools (Orange, Panda – Python Library).

xvi) Also learned and impressed with working and implementation of ASR, speech recognition and natural language processing technologies.

xvii) Upon acquiring the above knowledge, looking forward to work in machine learning / data analysis related projects.

My Involvement in Churn Fortunetelling Project

I am involved in the application development of Churn Fortunetelling project from end-to-end with following stages of application development lifecycle.

- Requirement Analysis
- Data Identification
- Data Analysis
- Application Scope
- Application Design
- Application Development
- Code review
- Verification Activities
- Unit & final Testing
- Documentation
    - Project Proposal
    - Churn Fortuneteller Project Report
    - Prepared the Churn Fortuneteller presentation slides