VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

COURSE NAME

RISK AND FRAUD ANALYTICS

(MGT3013)

SLOT

D1 + TD1

FACULTY

Dr BHUVANESH C

PROJECT TITLE

CUSTOMER CHURN ANALYSIS IN TELECOM INDUSTRY

TEAM MEMBERS

ALAGARSAMY N | 19MIA1082

NIRANJAN J | 19MIA1003

VIGNESH N | 19MIA1093

ROSHAN SRINIVAAS. S | 19MIA1001

T.S.S. ABINANDHAN KUMAR | 19MIA1062

FALL SEMESTER 2022-2023

INTRODUCTION:

Customer churn analysis and prediction in telecom sector is an issue now a days because it's very important for telecommunication industries to analyse behaviours of various customer to predict which customers are about to leave the from telecom company, So data mining techniques and algorithm plays an important role for companies in today's commercial conditions because gaining a new customer's cost is more than retaining the existing ones. We have planned to focus on various machine learning techniques for predicting customer churn through which we can build the classification models and also compare the performance of these models. Finding the causes of client loss, gauging customer loyalty, and winning back customers have all become crucial topics for many businesses. Instead of recruiting new clients, businesses do a variety of studies and initiatives to keep the ones they already have. The telecommunications industry collects enormous amounts of data as a result of fast developing technology, an increase in subscribers, and value-added services. This industry's unchecked and rapid expansion result in growing losses due to risk, fraud and technological challenges. Therefore, the creation of novel analysis techniques has become essential.

LITERATURE REVIEW:

[1] M.A.H. Farquad, proposed a hybrid approach to overcome the drawbacks of general SVM model which generates a black box model (i.e., it does not reveal the knowledge gained during training in human understandable form).

[2] introduced the hybrid neural networks techniques to predict the customer churners in a CRM dataset provided by American telecom companies. Here, they built two hybrid models by combining two different neural network International Journal of Computer Applications (0975 – 8887).

[3] Wouter Verbeke, proposed the application of Ant-Miner+ and ALBA algorithms on a publicly available churn prediction dataset in order to build accurate as well as comprehensible classification rule-sets churn prediction models.

[4] Ning Lu, proposed the use of boosting algorithms to enhance a customer churn prediction model in which customers are separated into two clusters based on the weight assigned by the

boosting algorithm. As a result, a high risky customer cluster has been found. Logistic regression is used as a basis learner, and a churn prediction model is built on each cluster, respectively. The experimental results showed that boosting algorithm provides a good separation of churn data when compared with a single logistic regression model.

[5] Benlan He, suggested a customer churn prediction methodology based on SVM model, and used random sampling method to improve SVM model by considering the imbalance characteristics of customer data sets. A support vector machine constructs a hyper-plane in a high- or infinite dimensional space, which can be used for classification. Random sampling method can be used to change the distribution of data in order to reduce the imbalance of the dataset. Imbalance in dataset is caused due to the low proportion of churners.

[6] Ssu-Han Chen, used a novel mechanism based on the gamma Cumulative SUM (CUSUM) chart in which the gamma CUSUM chart monitors individual customer's Inter Arrival Time (IAT) by introducing a finite mixture model to design the reference value and decision interval of the chart and used a hierarchical Bayesian model to capture the heterogeneity of customers. Recency, another time interval variable which is complementary to IAT, is combined into the model and tracks the recent status of the login behavior. In addition, benefits from the basic nature of control charts, the graphical interface for each customer is an additional advantage of the proposed method. The results showed that the accuracy rate (ACC) for gamma CUSUM chart is 5.2% higher than exponential CUSUM and the Average Time to Signal (ATS) is about two days longer than required for exponential CUSUM.

[7] Koen W. De Bock [10] proposed two rotation-based ensemble classifiers namely Rotation Forest and Rotboost as modeling techniques for customer churn prediction. An ensemble classifier is a combination of several member classifier models into one aggregated model, including the fusion rule to combine member classifiers outputs. In Rotation Forests, feature extraction is applied to feature subsets in order to turn the input data for training base classifiers, while RotBoost combines Rotation Forest with AdaBoost. Four data sets from

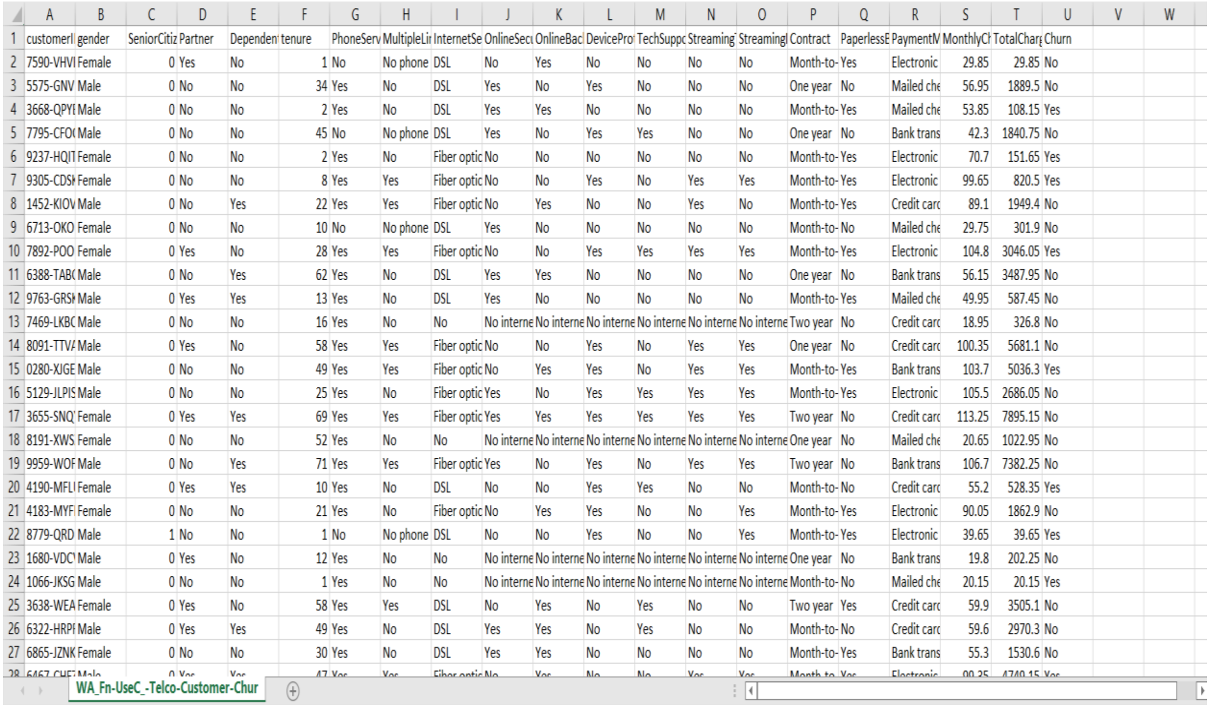
real-life customer churn prediction projects are used here. The results showed that Rotation Forests outperform RotBoost in terms of area under the curve (AUC) and top-decile lift, while RotBoost demonstrates higher accuracy than Rotation Forests. They also compared three alternative feature extraction algorithms namely: Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Sparse Random Projections (SRP) on classification performance of both RotBoost and Rotation Forest.

[8] Ver-braken et al. [14] proposed a new performance measure called the expected maximum profit criterion, which is aligned with the main objectives of the end users. The proposed framework not only assists the companies with selecting the classifier that maximizes the profit, but also provides information about the fraction of the customer base to be included in the retention campaign.

[9] P.C.Pendharkar [15] suggested two Genetic Algorithm(GA) based neural network (NN) models to predict the customer churn. The first GA-based NN model used a cross entropy based criteria to predict customer churn, and the second GA based NN model made some efforts to directly increase the prediction accuracy of customer churn. Using real-world customer dataset and three various sizes of NNs, they compared the two GA-based NN models with a statistical zscore model using model evaluation criterion like prediction accuracy, top 10% docile lift and area under Receiver Operating Characteristics (ROC) curve. The results of experiments indicated that both GA-based NN models outperform the statistical z-score model on all performance criteria.

[10] Y.Xie et al., [16] used an improved balance random forest (IBFR) model which is a combination of balanced random forests and weighted random forests in order to overcome the data distribution problem. The nature of IBRF is that the best features are iteratively learned by altering the class distribution and by putting higher penalties on misclassification of the minority class. The experiments are carried out with Chinese bank dataset which showed that IBRF is better than artificial neural network, decision tree and support vector machines in terms of accuracy.

DATASET DESCRIPTION:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	customerI	gender	SeniorCitiz	Partner	Dependent	tenure	PhoneServ	MultipleLir	InternetSe	OnlineSecu	OnlineBacl	DevicePro	TechSuppc	StreamingT	StreamingI	Contract	PaperlessE	PaymentM	MonthlyCh	TotalCharg	Churn		
2	7590-VHVI	Female	0	Yes	No	1	No	No phone	DSL	No	Yes	No	No	No	No	Month-to-	Yes	Electronic	29.85	29.85	No		
3	5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed che	56.95	1889.5	No		
4	3668-QPYI	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-	Yes	Mailed che	53.85	108.15	Yes		
5	7795-CFO(	Male	0	No	No	45	No	No phone	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank trans	42.3	1840.75	No		
6	9237-HQIT	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-	Yes	Electronic	70.7	151.65	Yes		
7	9305-CDSI	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-	Yes	Electronic	99.65	820.5	Yes		
8	1452-KIOV	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-	Yes	Credit carc	89.1	1949.4	No		
9	6713-OKO	Female	0	No	No	10	No	No phone	DSL	Yes	No	No	No	No	No	Month-to-	No	Mailed che	29.75	301.9	No		
10	7892-POO	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	104.8	3046.05	Yes		
11	6388-TAB(	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank trans	56.15	3487.95	No		
12	9763-GRSI	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-	Yes	Mailed che	49.95	587.45	No		
13	7469-LKBC	Male	0	No	No	16	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Two year	No	Credit carc	18.95	326.8	No		
14	8091-TTVA	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	No	Credit carc	100.35	5681.1	No		
15	0280-XJGE	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to-	Yes	Bank trans	103.7	5036.3	Yes		
16	5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	105.5	2686.05	No		
17	3655-SNQ'	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit carc	113.25	7895.15	No		
18	8191-XWS	Female	0	No	No	52	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Mailed che	20.65	1022.95	No		
19	9959-WOF	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank trans	106.7	7382.25	No		
20	4190-MFLI	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-	No	Credit carc	55.2	528.35	Yes		
21	4183-MYFI	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to-	Yes	Electronic	90.05	1862.9	No		
22	8779-QRD	Male	1	No	No	1	No	No phone	DSL	No	No	Yes	No	No	Yes	Month-to-	Yes	Electronic	39.65	39.65	Yes		
23	1680-VDC'	Male	0	Yes	No	12	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Bank trans	19.8	202.25	No		
24	1066-JKSG	Male	0	No	No	1	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Month-to-	No	Mailed che	20.15	20.15	Yes		
25	3638-WEA	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit carc	59.9	3505.1	No		
26	6322-HRPI	Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes	No	Yes	No	No	Month-to-	No	Credit carc	59.6	2970.3	No		
27	6865-JZNK	Female	0	No	No	30	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-	Yes	Bank trans	55.3	1530.6	No		
28	6467-CHF	Male	0	Yes	Yes	47	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	Yes	Month-to-	Yes	Electronic	99.35	4749.15	Yes		

AIM:

"Predict behaviour to retain customers. You can analyse all relevant customer data and develop focused customer retention programs."

ABOUT DATA:

Each row represents a customer, each column contains customer's attributes described on the column Metadata.

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges

- Demographic info about customers – gender, age range, and if they have partners and dependents

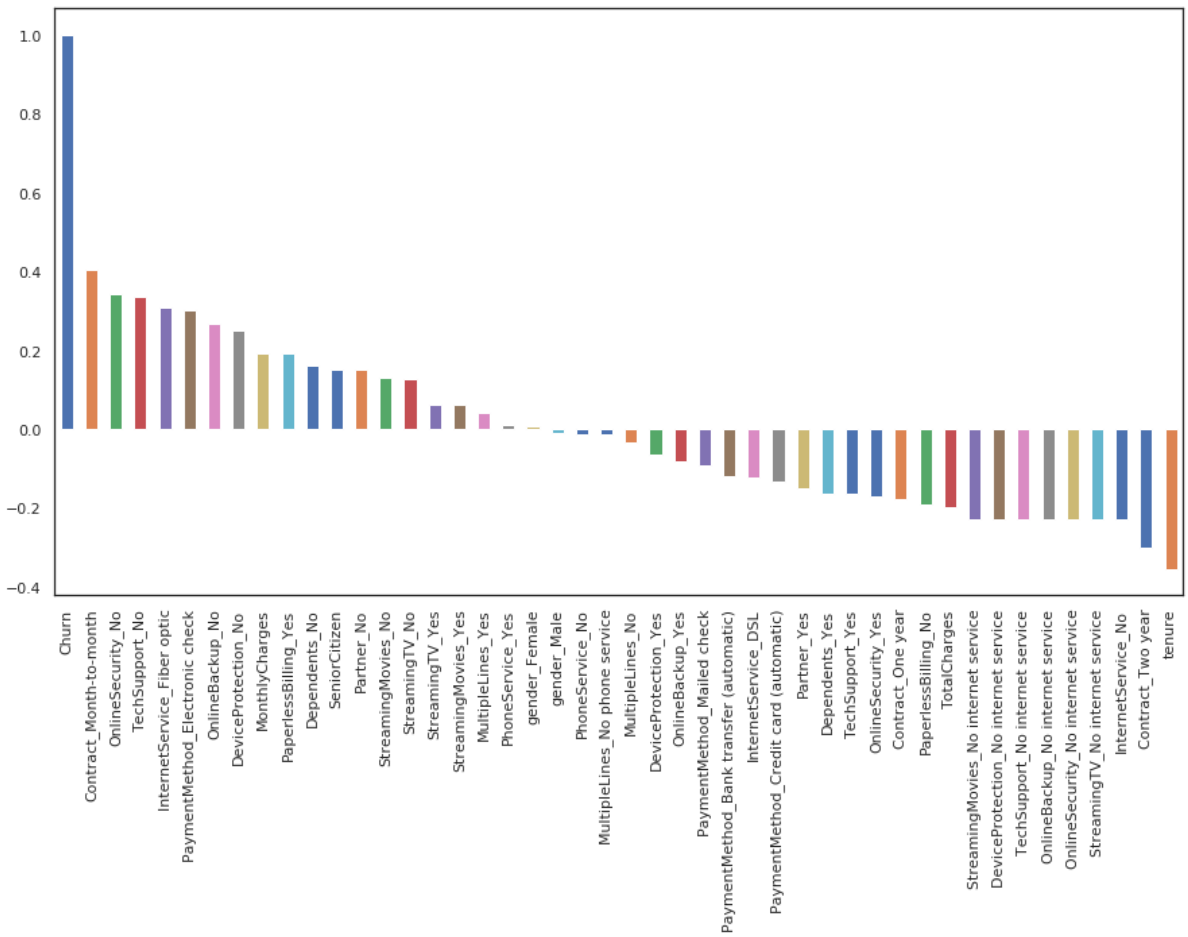
METHODOLOGY:

- Initially we have explored and Visualised various plots on the data to understand the various relations between the predictor and other dependent variables in the dataset.
- To forecast the churn factor, five machine learning techniques were used: logistic regression, XG Boost, Random Forest, SVM, and ADA Boost. From these three models, we discovered that XG Boost's performance is excellent in comparison to other models, increasing the accuracy of the test results.

RESULTS AND FINDINGS:

Initially we performed data cleaning which gives overall productivity and allow for the highest quality information in your decision-making. One of the main issues facing the telecom sector is churn. According to studies, the top 4 wireless carriers in the US have an average monthly churn rate between 1.9% and 2%.

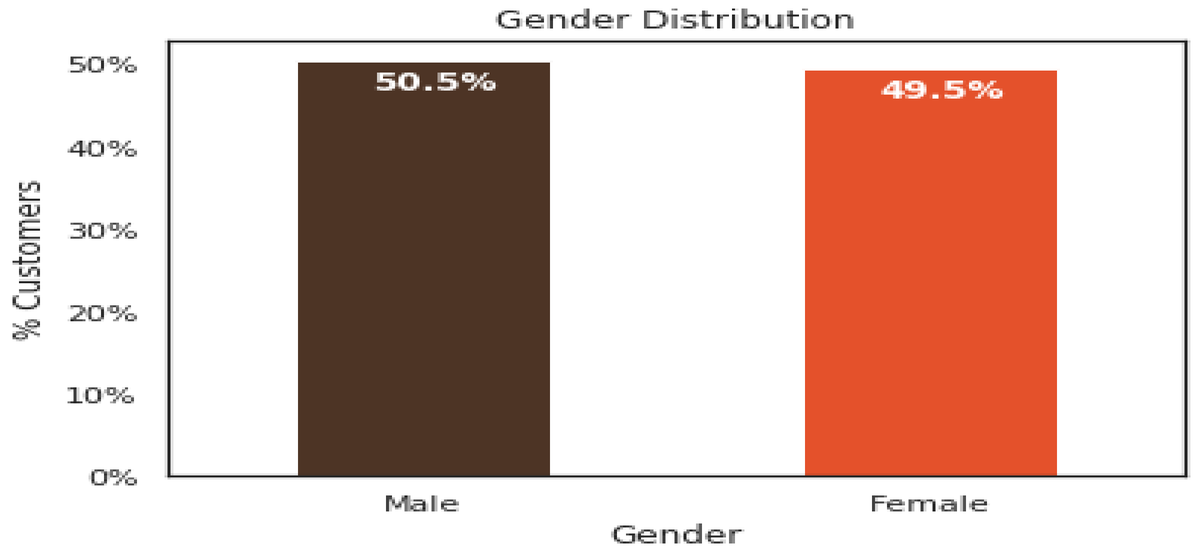
Next we saw the null values in the dataset there are 11 missing values for Total Charges. Let us replace remove these 11 rows from our data set, and found the correlation between predictor variable(churn) and other variables.



Contracts that are month-to-month, a lack of online security, and tech support appear to have a favourable correlation with customer churn. While tenure appears to be positively connected with churn, two-year contracts do not. It's interesting to note that services like tech assistance, streaming TV, online backup, online security, etc. that don't require an internet connection appear to be adversely correlated with turnover. Before we dive into modelling and identifying the key variables, we will first examine the patterns for the aforementioned correlations.

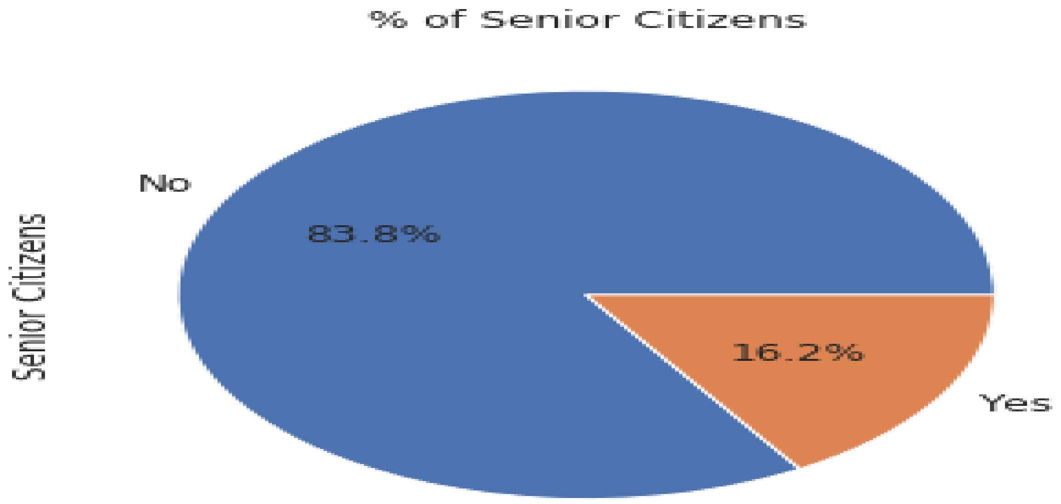
Next, we performed data exploration steps based on the demographics exploration, based on the customer account information like tenure, contract and other services used by the customers.

[1]



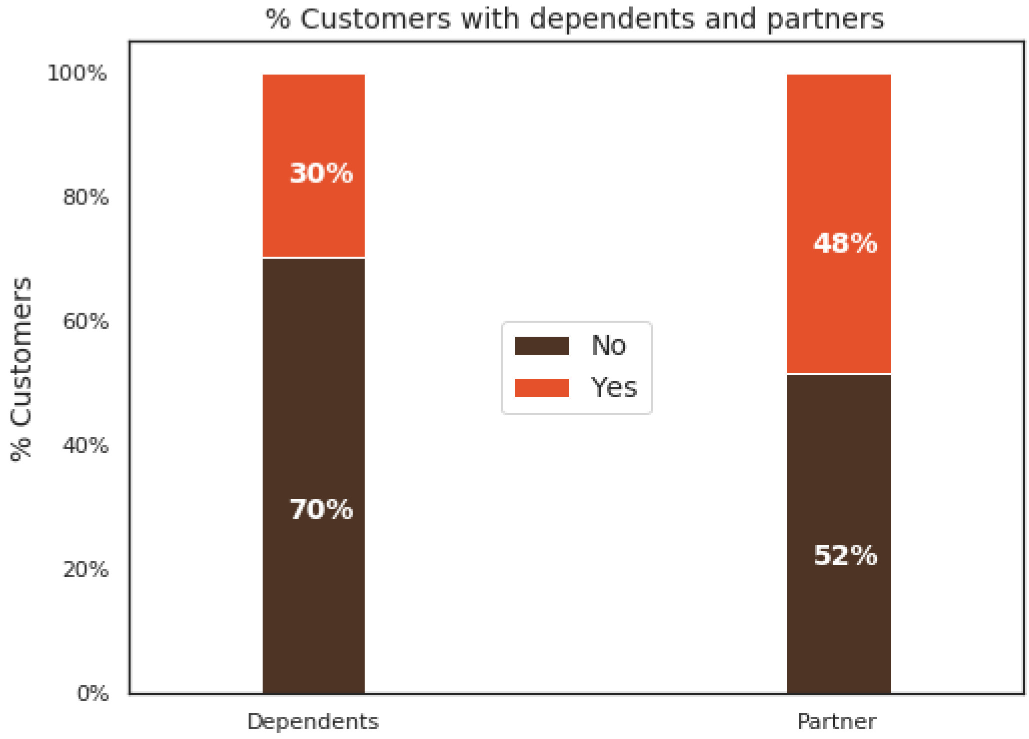
Only 16% of the clients are seniors, according to the statistics. Thus most of our customers in the data are younger people.

[2] Partner and dependent status



About 50% of the customers have a partner, while only 30% of the total customers have dependents.

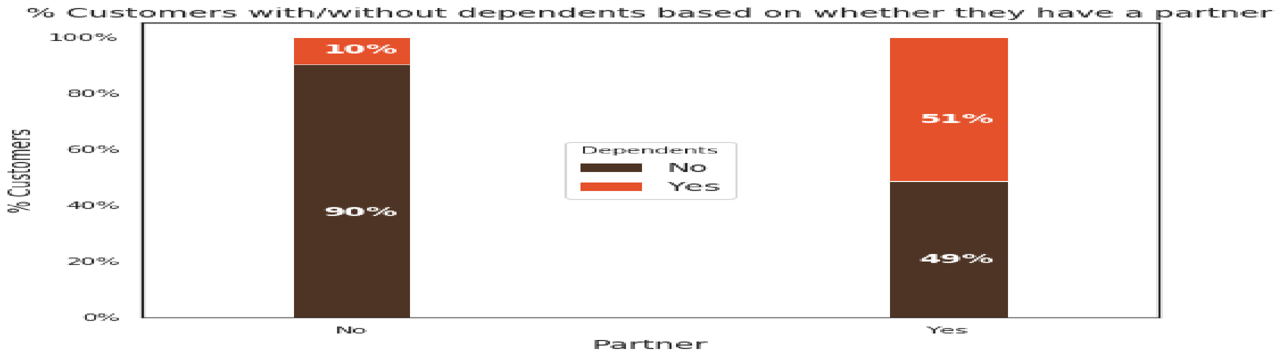
[3]



What would be interesting is to look at the % of customers, who have partners, also have dependents. We will explore this next.

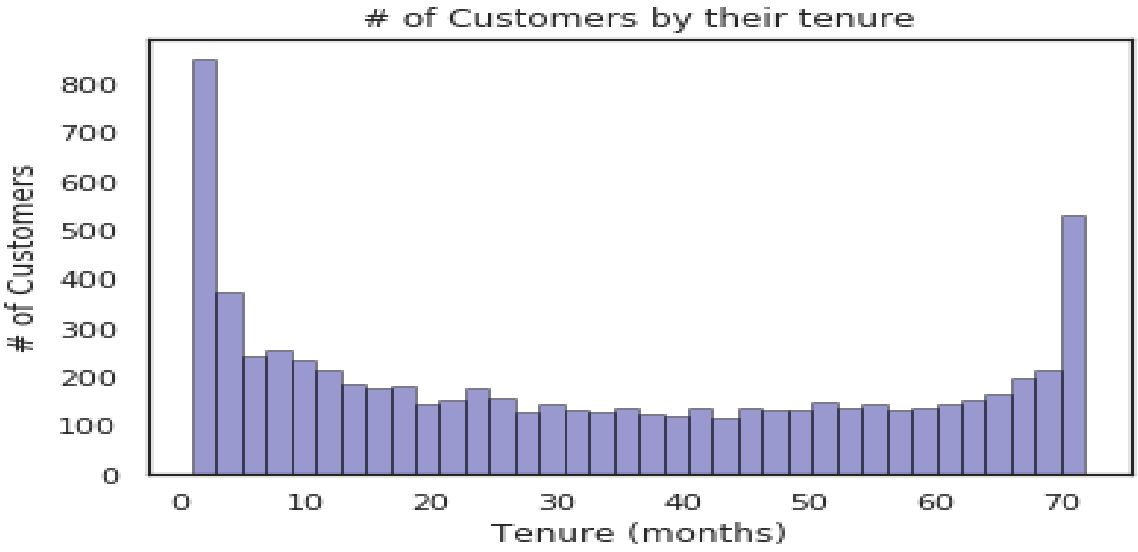
Interestingly, among the customers who have a partner, only about half of them also have a dependent, while other half do not have any independents. Additionally, as expected, among the customers who do not have any partner, a majority (80%) of them do not have any dependents.

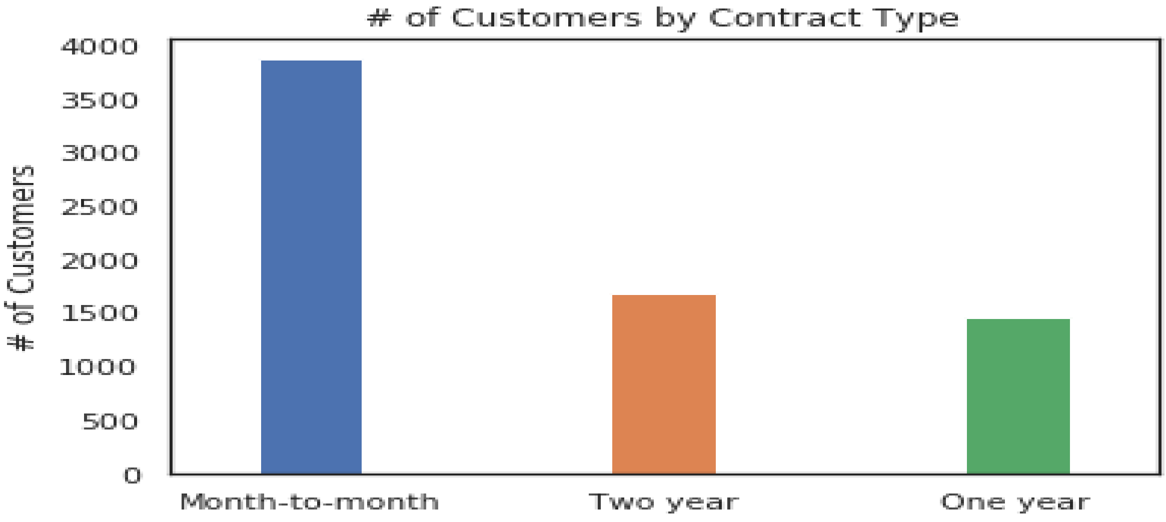
[4]



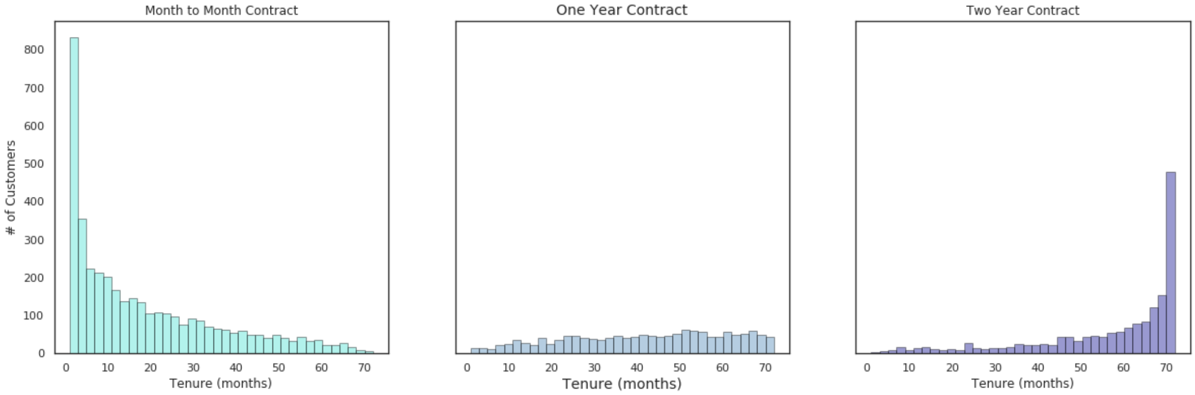
We also examined any gender-based disparities in the percentage of clients having dependents and partners. The distribution of them by gender is the same. Furthermore, there is no distinction in senior citizen status according to gender.

[4]



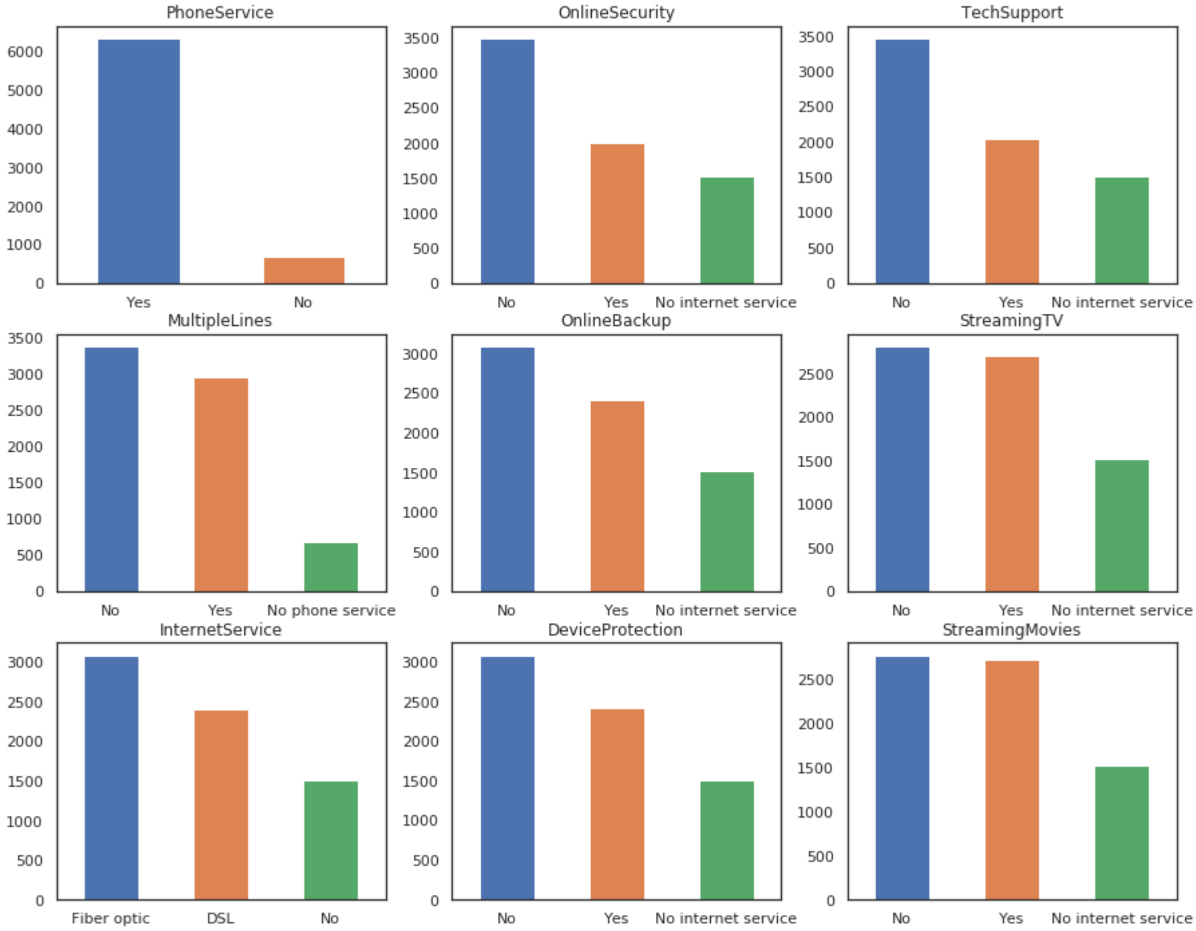


The majority of the consumers are under month-to-month contracts, as this graph shows. The number of customers in the 1 year and 2 year contracts is equal.

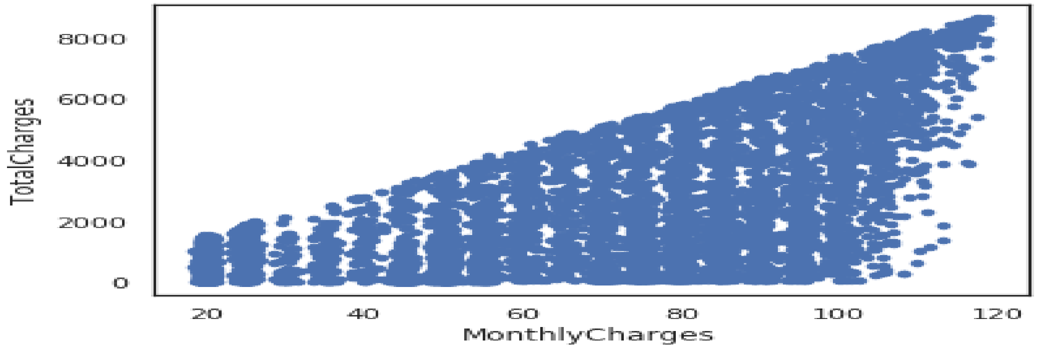


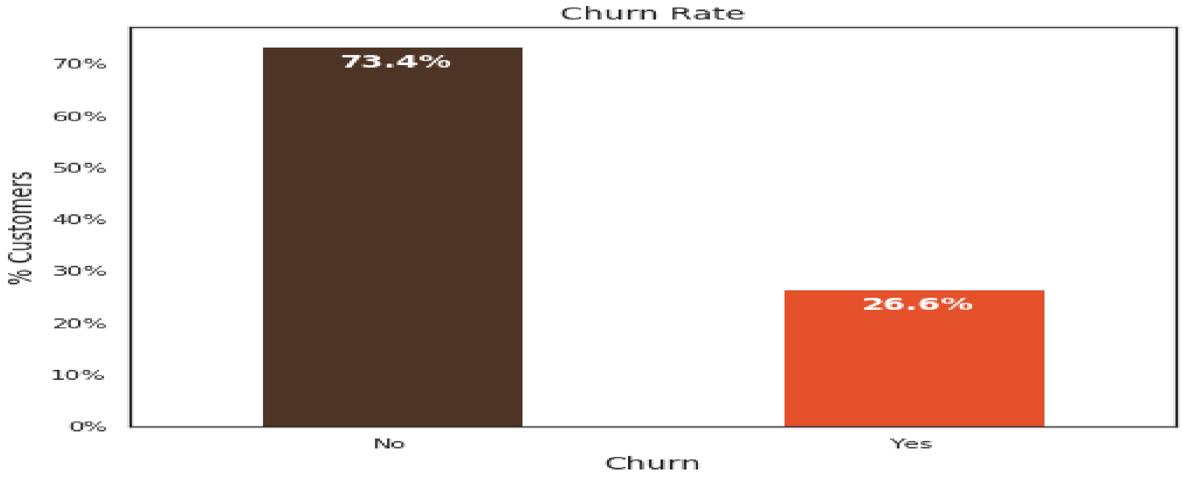
Interestingly, the majority of monthly contracts only last a few weeks to a few months, although two-year contracts often continue for over 70 months. This demonstrates that clients who sign longer contracts are more devoted to the business and have a propensity to stick with it. This is also what the preceding correlation with turnover rate chart showed.

distribution of various services used by customers

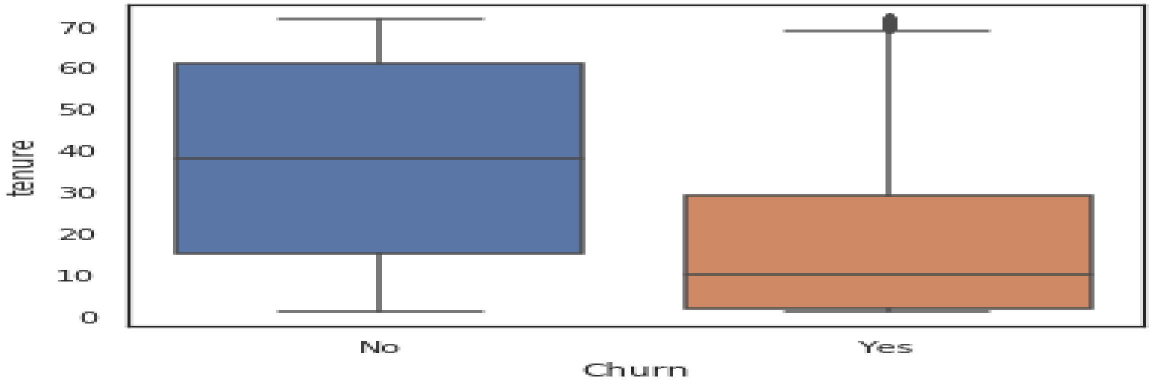


relation between monthly and total charges



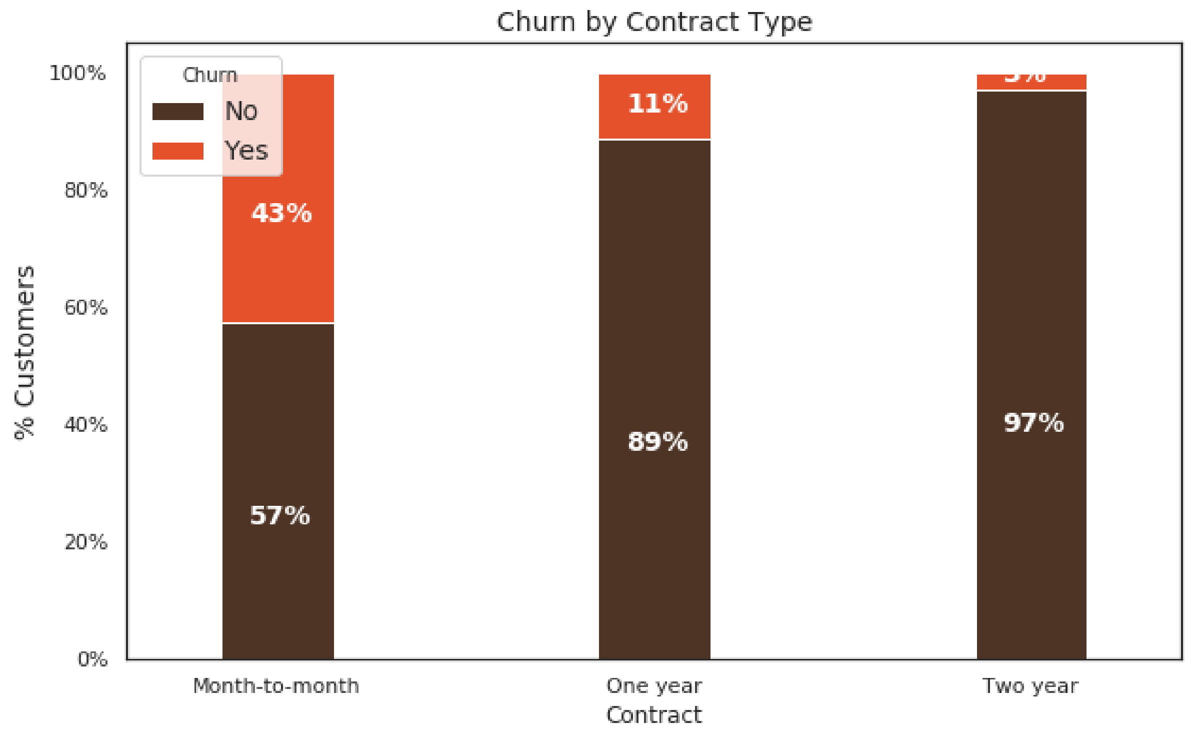


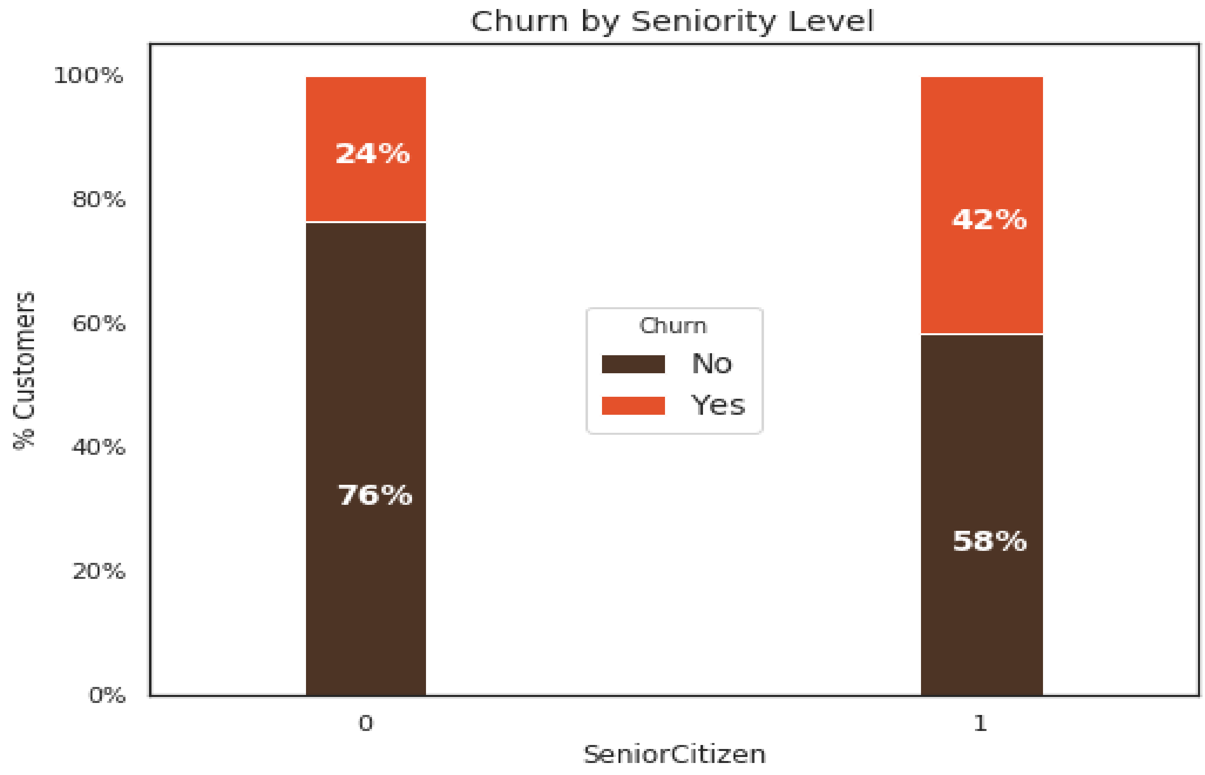
According to our analysis, 74% of clients remain loyal. We would anticipate that a sizable portion of the clients would remain loyal, therefore the data is obviously skewed. This is crucial to remember for our modelling since skewness may result in several false negatives. How to prevent skewness in the data will be covered in the section on modelling.



Churn vs. Tenure: As we can see from the below plot, consumers that do not churn tend to stay with the telecom firm for a longer period of time.

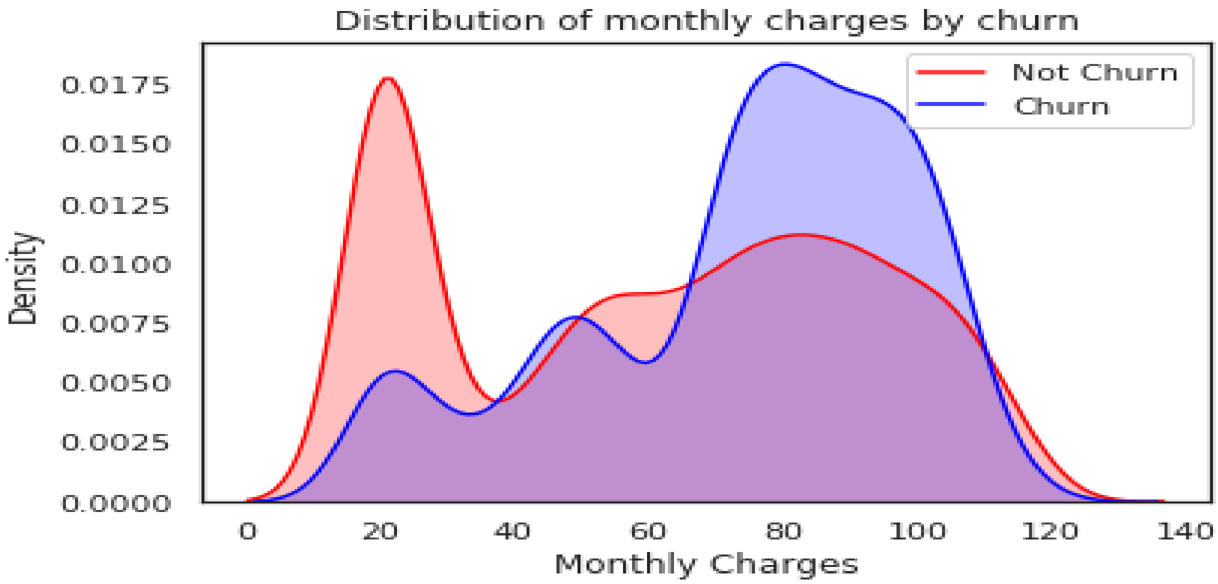
Churn by Contract Type: Similar to what we saw in the correlation plot, the customers who have a month to month contract have a very high churn rate.



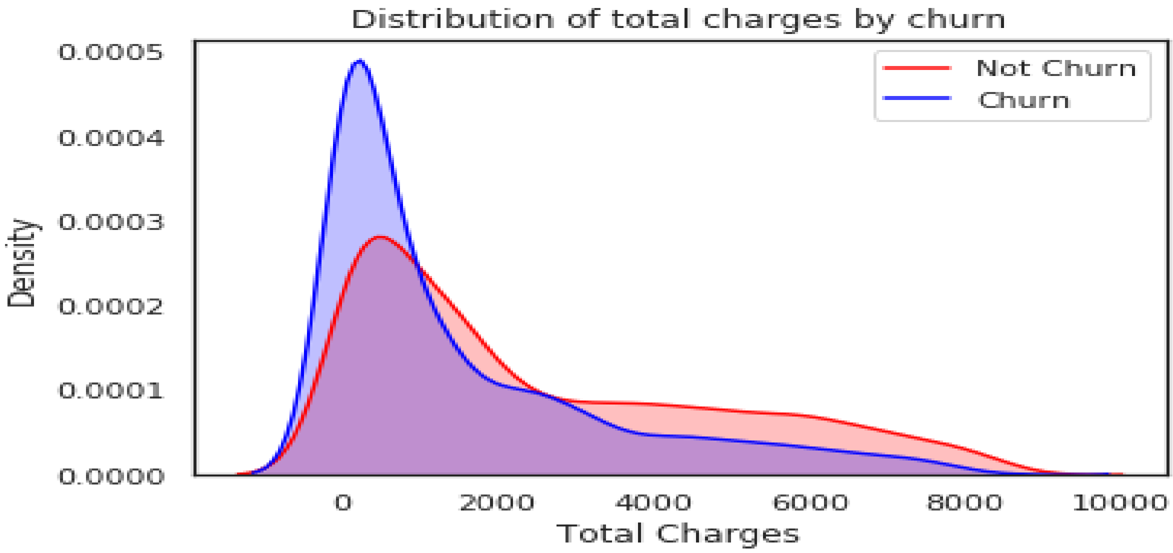


Churn by Seniority: The churn rate for seniors is nearly double that of the general population.

Churn by Monthly Charges: Higher % of customers churn when the monthly charges are high.



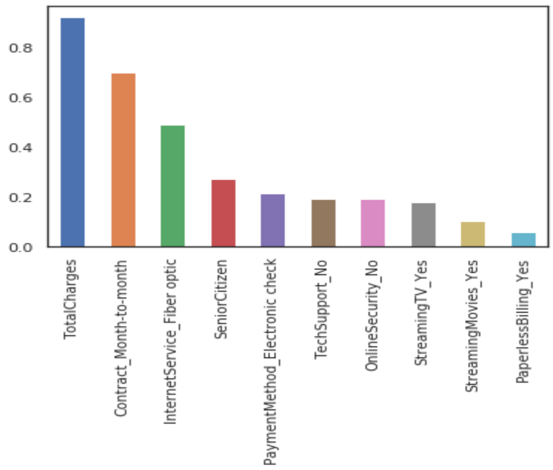
Churn by Total Charges: It seems that there is higer churn when the total charges are lower.

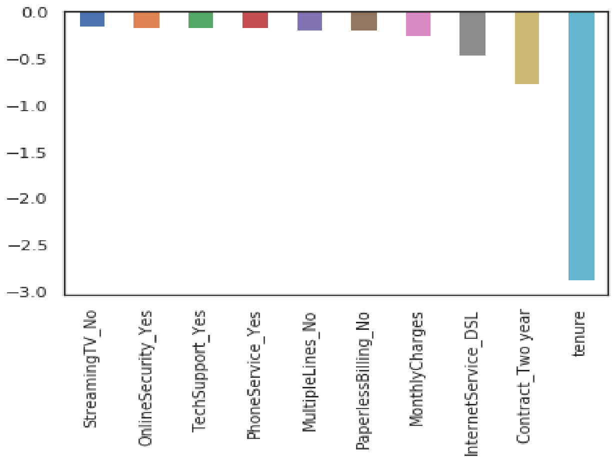


MODELS USED IN OUR PROJECT AND RESPECTIVE RESULTS:

1] LOGISTIC REGRESSION:

WE HAVEscale the variables in logistic regression so that all of them are within a range of 0 to 1. This helped me improve the accuracy from 79.7% to 80.7%. Further, you will notice below that the importance of variables is also aligned with what we are seeing in Random Forest algorithm and the EDA we conducted above.

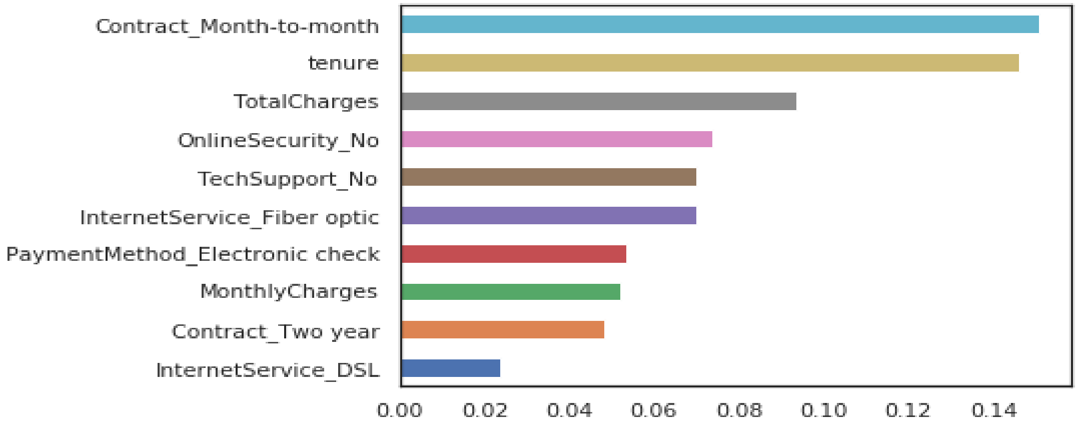




As we can see, certain variables are positively correlated with our predicted variable (Churn), while others are negatively correlated. A negative relationship suggests that as that variable increases, the likelihood of churn reduces. Below, we'll list a few of the noteworthy characteristics:

The likelihood of churn is decreased by a two-month contract, as we observed in our EDA. According to the results of logistic regressions, the 2 month contract and tenure had the most adverse relationships with turnover. Additionally, having DSL internet service lessens the likelihood of churn. Finally, increased churn rates may be caused by total costs, monthly commitments, fibre optic internet services, and seniority. The accuracy of logistic is being estimated as **0.80758**

[2] RANDOM FOREST:

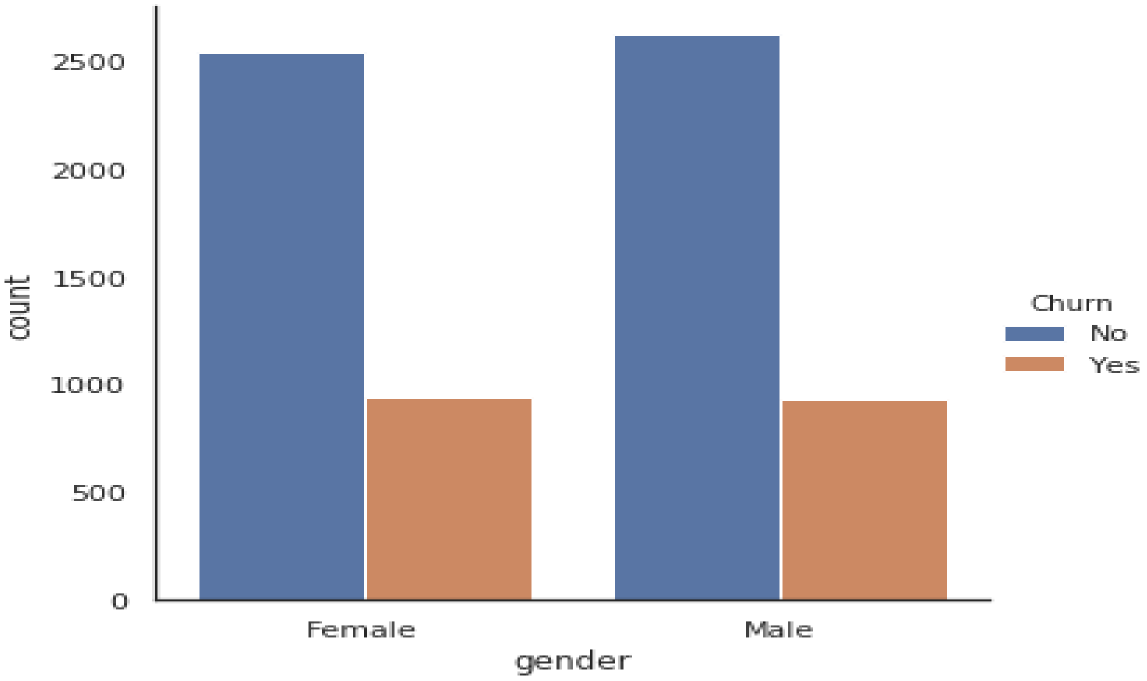


The most significant predictor variables for churn are monthly contract, tenure, and total charges according to the random forest algorithm, and for random forest we got accuracy as 0.8088

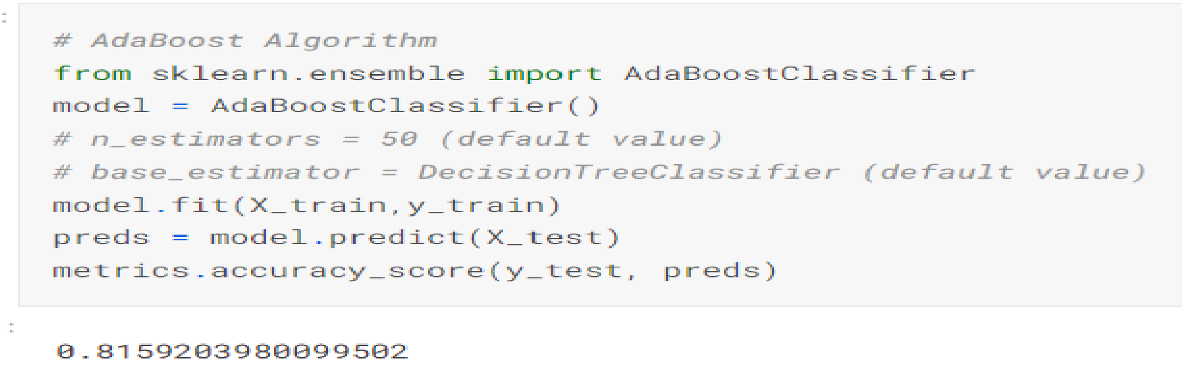
The results from random forest are consistent with our expectations from our EDA and fairly similar to those from logistic regression.

[3] **Support vector machine :**

In SVM we were successful in raising the accuracy to 82% by cross validating the accuracy which boost the result, using SVM we got 953 true positive records, 89 false positive records, 164 false negative records and 201 false positive records, which mean 953 records were found to be genuine in the case.



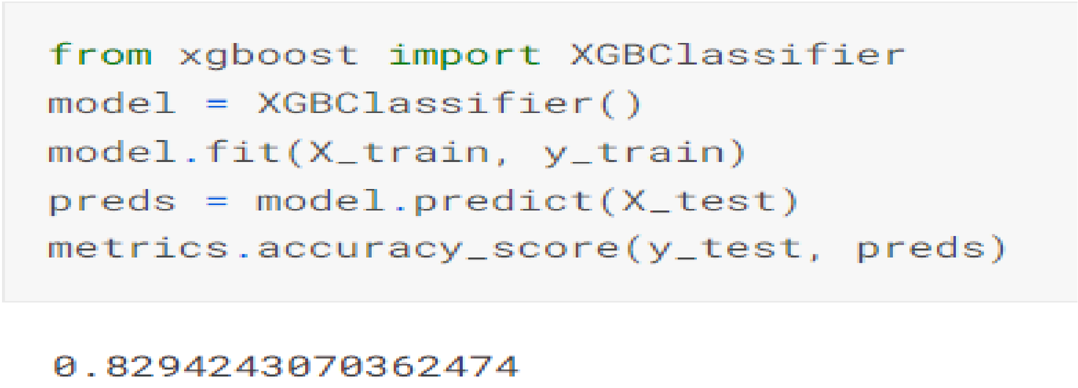
[4] **ADA BOOST:**



```
# AdaBoost Algorithm
from sklearn.ensemble import AdaBoostClassifier
model = AdaBoostClassifier()
# n_estimators = 50 (default value)
# base_estimator = DecisionTreeClassifier (default value)
model.fit(X_train,y_train)
preds = model.predict(X_test)
metrics.accuracy_score(y_test, preds)
```

```
0.8159203980099502
```

[5] **XG BOOST:**



```
from xgboost import XGBClassifier
model = XGBClassifier()
model.fit(X_train, y_train)
preds = model.predict(X_test)
metrics.accuracy_score(y_test, preds)
```

```
0.8294243070362474
```

RESULT:

It's interesting that using XG Boost, we were able to boost test data accuracy to approximately 83%. Without a doubt, XG Boost outperforms all other methods. The slow learning model XG Boost is based on the idea of boosting.

CONCLUSION:

The purpose of this kind of study in the telecom industry is to assist businesses in increasing their profits. It is well known that one of the most significant revenue streams for telecom firms is churn prediction. As a result, the goal of this research was to develop a system that could forecast customer turnover at the telecom business.

Our goal was to identify clients which are likely to churn, so we can do special purpose marketing strategies to avoid the churn event.

- For this we evaluated differently preprocessed datasets and different classifiers. The analysis has shown that the PCA transformation was not found to be useful. Instead, we suggest to use the whole dataset and apply a oversampling technique in order to deal with the unbalanced target variable.
- In the classification chapter we have trained several different classifiers, including a Logistic Regression, a Support-Vector Machine, a ADA BOOST, XG BOOST and a Random Forest.
- It was found that the best performance in accuracy, is achieved by XG boost.

- Concluding, we suggest the Telecom company to use the XG BOOST model to identify potential churn customers and according to the customers life-time value present them special offers and it mainly helps in reducing risk towards the company on behalf of loosing the present customers.

REFERENCE:

[1] AlOmari, D., Hassan, M.M. 2016.PredictingTelecommunication Customer Churn Using Data Mining Techniques. 9th International Conference on Internet and Distributed Computing Systems, 167-178.

[2] Amin, A., Khan, C., Ali, I., Anwar, A. 2014. Customer Churn Prediction in Telecommunication Industry: with and without counter-Example. European Network Intelligence Conference, 134-137

[3] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., Huang, K. 2016. Customer Churn Prediction in Telecommunication Sector using Rough Set Approach. Neurocomputing, http://dx.doi.org/10.1016/j. neucom.2016.12.009, 2016:1-21

[4] Argüden Y., Erşahin B. 2008. Veri Madenciliği: Veriden Bilgiye, Masraftan Değere. ARGE Danışmanlık, ISBN: 978-975- 93641-9-9 1. Basım.

[5] Backiel, A., Verbinnen, Y., Baesens, B., Claeskens, G. 2015. Combining Local and Social Network Classifiers to Improve Churn Prediction. International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 651-658.

[6] Brandusoiu, I., Toderean, G. 2013. Churn Prediction In The Telecommunications Sector Using Support Vector Machines. Annals Of The Oradea Un., Fascicle Manag. and Tech. Eng., 1: 19-22

[7] Brandusoiu, I., Toderean, G., Beleiu, H. 2016. Methods for churn prediction in the pre-paid mobile telecommunications industry. International Conference on Communications (COMM), 97-100.

[8] Coussement, K., Lessmann, S., Verstraeten, G. 2016. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decis. Supp. Sys., 2016, http:// dx.doi.org/10.1016/j.dss.2016.11.007.

[9] Coussement, K., Lessmann, S., Verstraeten, G. 2016. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decis. Supp. Sys., 2016, http:// dx.doi.org/10.1016/j.dss.2016.11.007.

[10] Dalvi, PK., Khandge, SK., Deomore, A. 2016. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. Symposium on Colossal Data Analysis and Networking (CDAN), DOI: 10.1109/ CDAN.2016.7570883, 1-8

[11] Esteves, G., Mendes-Moreira, J. 2016. Churn Perdiction in the Telecom Business, The eleventh International Conference on Digital Information Management (ICDIM 2016), 254-259.

[12] Forhad, N., Hussain, S., Rahman, RM. 2014. Churn Analysis: Predicting Churners. Ninth International Conference on Digital Information Management (ICDIM), 237-241.

[13] Gok, M., Ozyer, T., Jida, J. 2015. A Case Study for the Churn Prediction in Turksat Internet Service Subscription. IEEE/ ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 1220-1224.

****