Churn Prediction

Aatish Kayyath
MS in Machine Learning
Department of Computer Science
Jersey City, USA
akayyath@stevens,edu

Aston Glen Noronha
MS in Data Science
Department of Mathematics
Jersey City, USA
anoronha@stevens.edu

Rohith Sure
MS in Data Science
Department of Mathematics
Jersey City, USA
rsure@stevens.edu

***Abstract*—To build a classification system to predict whether a customer will churn or not based on the Telecom Data from Kaggle. Technically, it is a binary classifier that divides clients into two groups (classes) — those who leave and those who do not. The classifier will be built using Logistic Regression, bagging algorithms like Random Forest, KNN and boosting algorithms like ADABoost, XGBoost.**

I. Introduction

Losing valuable customers is never a pleasant experience for any company. "Churn is defined in business terms as 'when a client cancels a subscription to a service they have been using.'" A common example is people cancelling Spotify/Netflix subscriptions. The purpose of Churn Prediction is to predict which of your clients will cancel a subscription, i.e., leave a business. Obtaining this information is necessary from a company's perspective since acquiring new customers can be arduous and costly. In this way, Churn Prediction enables them to focus more on customers at a high risk of leaving. Telecom churn prediction is the process of predicting customer attrition or customer churn in the telecommunications industry. Customer churn is a critical issue for the telecom industry because it is more expensive to acquire new customers than to retain existing ones. Therefore, it is important to be able to identify customers who are likely to churn and take proactive measures to keep them. Machine learning (ML) algorithms can be used to build churn prediction models. ML algorithms can be used to create models based on customer attributes such as demographics, usage patterns and past behaviours. These models can then be used to accurately predict customer churn.

II. Related Work

[1] Anil Kumar and Ravi used data mining to predict credit card customer churn. They used multilayer perceptron (MLP), logistic regression, DT, random forest, radial basis function, and SVM techniques.
[2] Nie et al. built a customer churn prediction model by using logistic regression and DT-based techniques within the context of the banking industry. In their study, Lin et al. used rough set theory and rule-based decision-making techniques to extract rules related to customer churn in credit card accounts using a flow network graph (a path-dependent approach to deriving decision rules and variables). They further showed how rules and different kinds of churn are related.
[3] Sharma and Panigrahi applied neural networks to predict customer churn from cellular network services. The results indicated that neural networks could predict customer churn with an accuracy of higher than 92%.
[4] Huang et al. presented new-features-based logistic regression (LR), linear classifier (LC), NB, DT, MLP neural networks, and SVM. In their experiments, each technique produced a different output. Data mining by evolutionary learning (DMEL) could show the reason or probability of a churning phenomenon; DT, however, could only show the reason. LR, NB, and MLP could provide probabilities of different customer behaviors. LC and SVM could distinguish between a churner and a non-churner.
[5] Makhtar et al. proposed a model for churn prediction using rough set theory in telecom. As mentioned in this paper Rough Set classification algorithm outperformed the other algorithms like Linear Regression, Decision Tree, and Voted Perception Neural Network.
[6] Burez and Van den Poel studied the problem of unbalanced datasets in churn prediction models and compared performance of Random Sampling, Advanced Under-Sampling, Gradient Boosting Model, and Weighted Random Forests. They used (AUC, Lift) metrics to evaluate the model. The result showed that the undersampling technique outperformed the other tested techniques.
[7] Huang et al. studied the problem of customer churn in the big data platform. The goal of the researchers was to prove that big data greatly enhances the process of predicting the churn depending on the volume, variety, and velocity of the data. Dealing with data from the Operation Support department and Business Support department at China's largest telecommunications company needed a big data platform to engineer the fractures. Random Forest algorithm was used and evaluated using AUC.

III. Our Solution

A. Description of Dataset

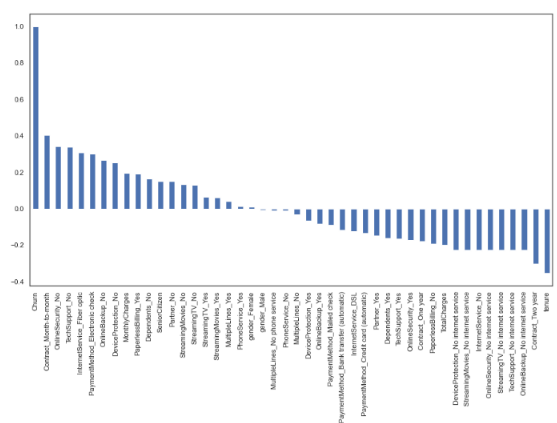
Source of Dataset: https://www.kaggle.com/datasets/aatishizgr8/teleco-curn-data
Data contains 19 independent variables and 1 dependent variable Churn. Here Churn signifies if the customer is going to stay with the provider or not.

19 independent variables are:
1. Gender: Whether the customer is a male or a female
2. Senior Citizen: Whether the customer is a senior citizen or not (1, 0)
3. Partner: Whether the customer has a partner or not (Yes, No)
4. Dependents: Whether the customer has dependents or not (Yes, No)
5. Tenure: Number of months the customer has stayed with

the company
6. Phone Service: Whether the customer has a phone service or not (Yes, No)
7. Multiple Lines: Whether the customer has multiple lines or not (Yes, No, No phone service)
8. Internet Service: Customer's internet service provider (DSL, Fiber optic, No)
9. Online Security: Whether the customer has online security or not (Yes, No, No internet service)
10. Online Backup: Whether the customer has online backup or not (Yes, No, No internet service)
11. Device Protection: Whether the customer has device protection or not (Yes, No, No internet service)
12. Tech Support: Whether the customer has tech support or not (Yes, No, No internet service)
13. Streaming TV: Whether the customer has streaming TV or not (Yes, No, No internet service)
14. Streaming Movies: Whether the customer has streaming movies or not (Yes, No, No internet service)
15. Contract: The contract term of the customer (Month-to-month, One year, Two year)
16. Paperless Billing: Whether the customer has paperless billing or not (Yes, No)
17. Payment Method: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
18. Monthly Charges: The amount charged to the customer monthly
19. Total Charges: The total amount charged to the customer

Exploratory Data Analysis:

First, we convert non-numeric columns to a numeric datatype and categorical variables into dummy variables. Then we check for any missing values. If any, we remove the corresponding rows from the dataset.
Then we find the Correlation between the dependent variable (Churn) and the other independent variables.

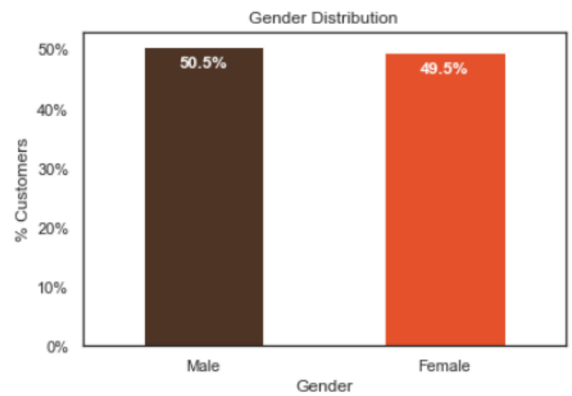


Month-to-month contracts and the absence of online security and tech support seem to be positively correlated with churn. While tenure and two-year contracts seem to be negatively correlated with churn.
Interestingly, services such as Online security, streaming TV, online backup, tech support, etc. without an internet connection seem to be negatively related to churn.
We will explore the patterns for the above correlations below before we delve into modeling and identifying the important variables.

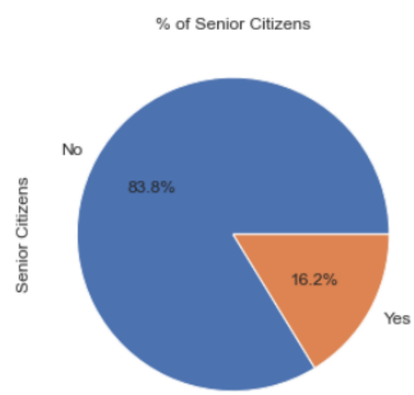
Now, we dive into the dataset to find and understand the patterns in the data which could potentially form some hypothesis. We start with the distribution of individual variables.

1. Demographics

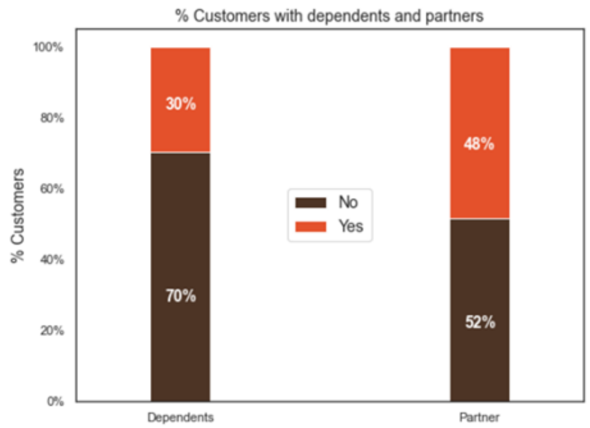
Gender Distribution: About half of the customers in our data set are male while the other half are female



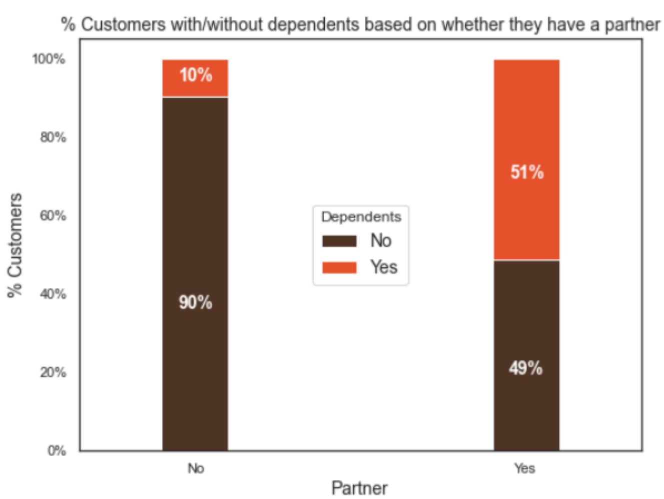
% Senior Citizen: There are only 16% of the customers who are senior citizens. Thus most of our customers in the data are younger people.



Partner and dependent status: About 50% of the customers have a partner, while only 30% of the total customers have dependents.

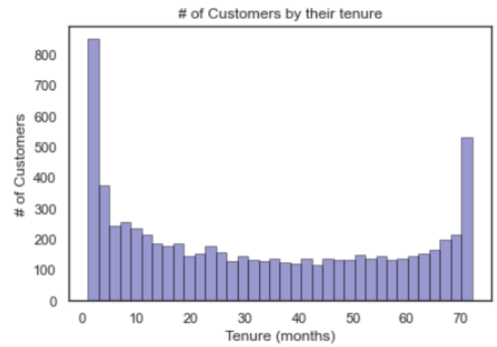


Interestingly, among the customers who have a partner, only about half of them also have a dependent, while the other half do not have any independent. Additionally, as expected, among the customers who do not have any partner, a majority (80%) of them do not have any dependents.

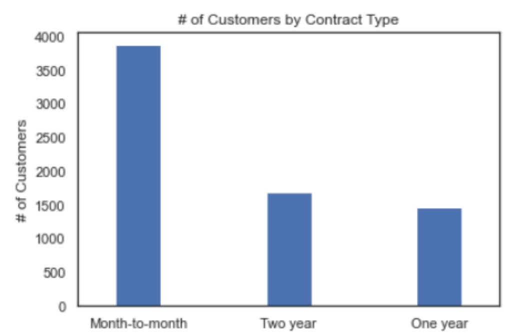


2. Customer Account Information

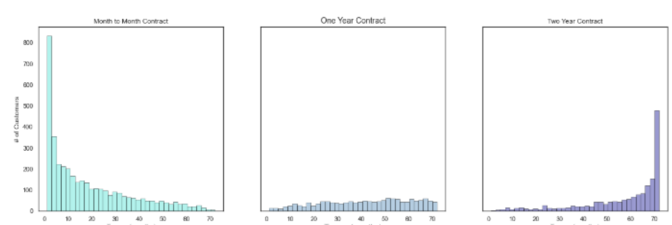
Tenure: After looking at the below histogram we can see that a lot of customers have been with the telecom company for just a month, while quite a many are there for about 72 months. This could be potential because different customers have different contracts. Thus based on the contract they are into it could be more/less easy for the customers to stay/leave the telecom company



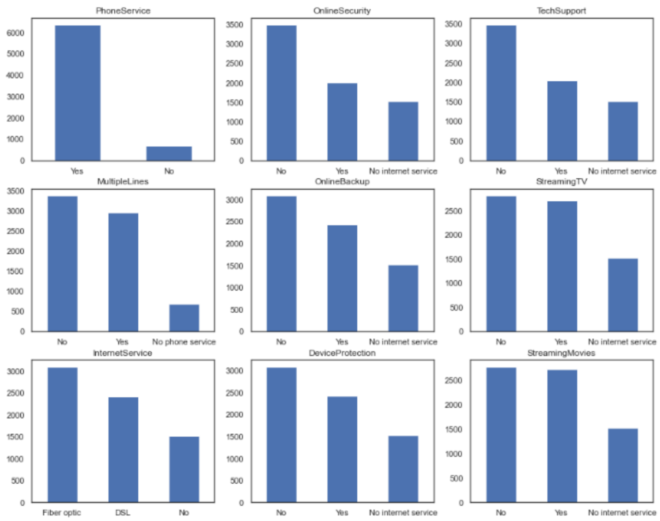
Contract: As we can see from this graph most of the customers are in a month-to-month contract. While there are an equal number of customers in the 1-year and 2-year contracts.



Interestingly most of the monthly contracts last for 1-2 months, while the 2 year contracts tend to last for about 70 months. This shows that the customers taking a longer contract are more loyal to the company and tend to stay with it for a longer period of time.
This is also what we saw in the earlier chart on correlation with the churn rate.

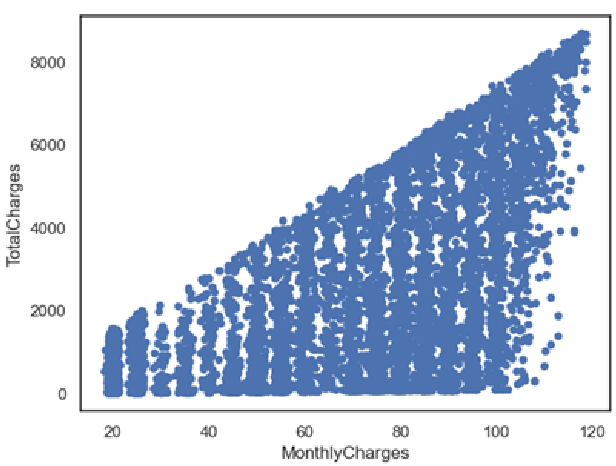


3. Distribution of various services used by customers



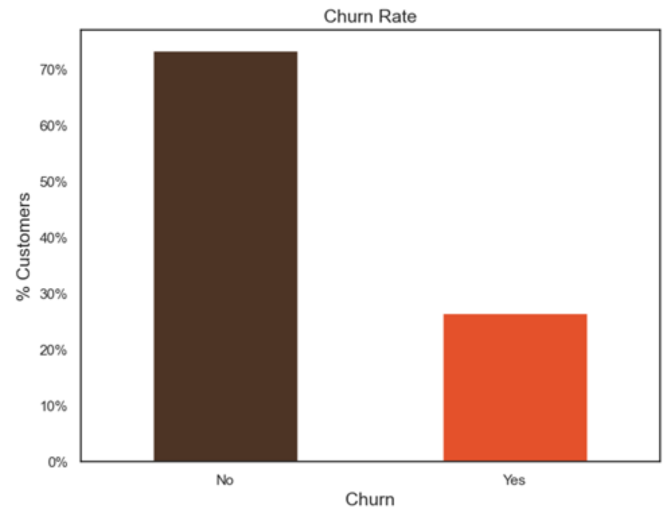
3. Relationship between monthly and total charges

We will observe that the total charges increase as the monthly bill for a customer increases.



4. Finally, let's take a look at our predictor variable (Churn) and understand its interaction with other important variables as was found out in the correlation plot.

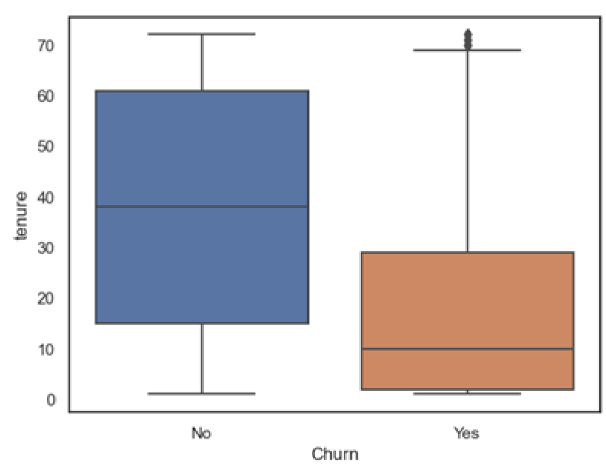
Let's first look at the churn rate in our data



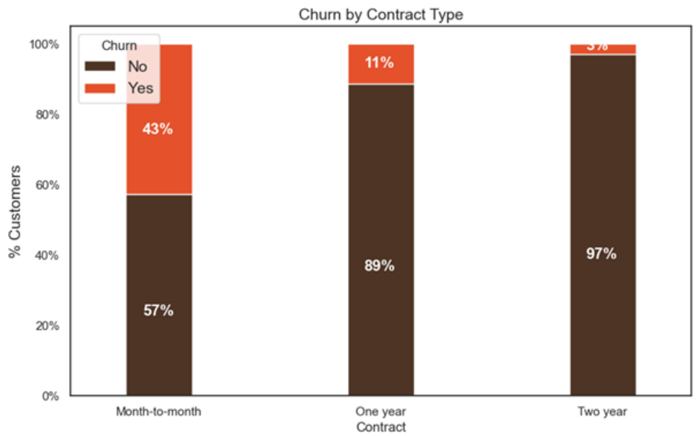
In our data, 74% of the customers do not churn. Clearly, the data is skewed as we would expect a large majority of the customers to not churn. This is important to keep in mind for our modeling as skewness could lead to a lot of false negatives. We will see in the modeling section how to avoid skewness in the data.

Let's explore the churn rate by tenure, seniority, contract type, monthly charges, and total charges to see how it varies by these variables.

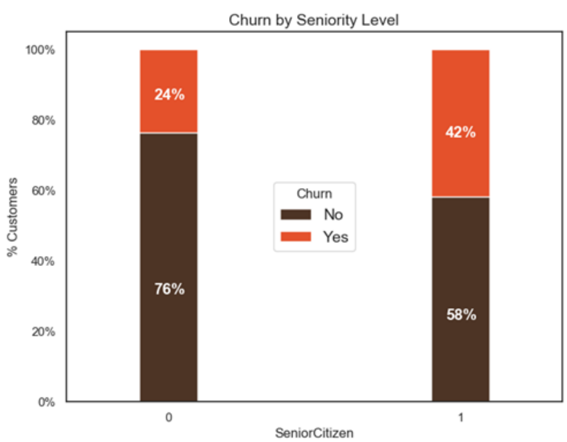
1. Churn vs Tenure: As we can see from the below plot, the customers who do not churn, tend to stay for a longer tenure with the telecom company.



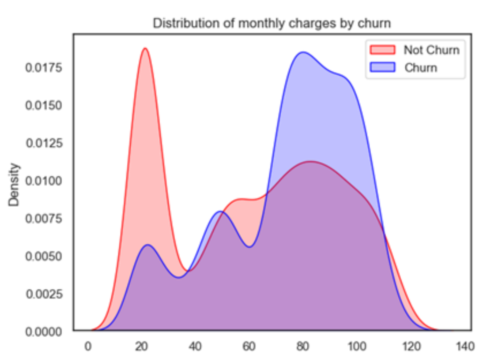
2. Churn by Contract Type: Similar to what we saw in the correlation plot, the customers who have a month-to-month contract have a very high churn rate.



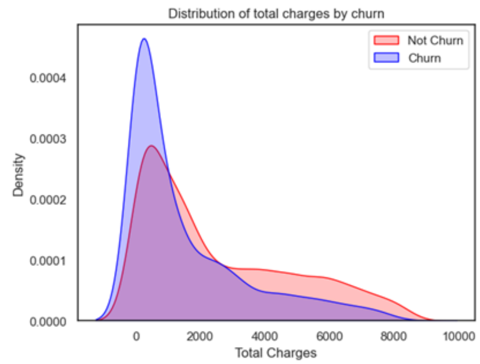
3. Churn by Seniority: Senior Citizens have almost double the churn rate than the younger population.



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



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



B. *Machine Learning Algorithms*

A big part of machine learning is classification — we want to know what class (a.k.a. group) an observation belongs to. The ability to precisely classify observations is extremely valuable for various business applications like predicting whether a particular user will buy a product or forecasting whether a given loan will default or not. The following ML models have been implemented.

Logistic Regression. It is a type of statistical model (also known as the logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

Random forests is a supervised learning algorithm. It can be used both for classification and regression. Random forests create decision trees on randomly selected data samples, get a prediction from each tree, and select the best solution by means of voting. It also provides a pretty good indicator of the feature's importance.

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

Boosting is a special type of Ensemble Learning technique that works by combining several weak learners(predictors with poor accuracy) into a strong learner(a model with strong accuracy). This works by each model paying attention to its predecessor's mistakes.

In Gradient Boosting, each predictor tries to improve on its predecessor by reducing errors. But the fascinating idea behind Gradient Boosting is that instead of fitting a predictor on the data at each iteration, it actually fits a new predictor to the residual errors made by the previous predictor.

AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level which means Decision trees with only 1 split. These trees are also called Decision Stumps. What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a low error is received.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

C. Implementation Details

Firstly, all the above models have been implemented without any parameter tuning. This way we can see if there is any overfitting. And then, hyperparameter tuning like L2 regularization for Logistic Regression, RandomizedSearchCV for Random Forest and Gradient Boosting Classifier and GridSearchCV for ADA Boost and XG Boost Classifier.

IV. Comparison

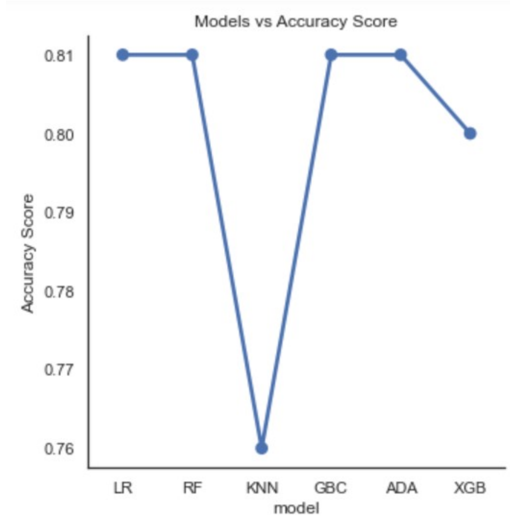
Before tuning:

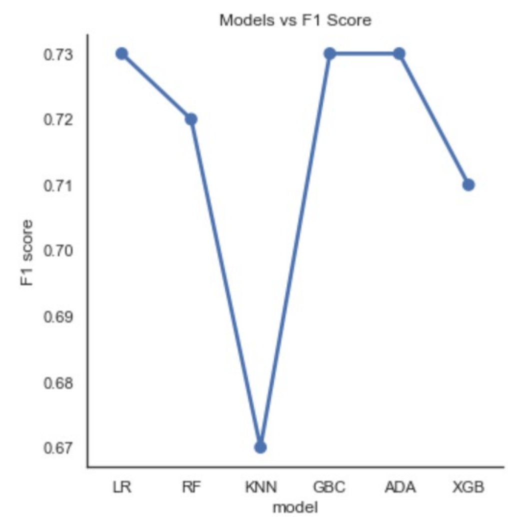
	Classification Model	Accuracy	Precision	Recall	F1-score
1	Logistic Regression	0.81	0.75	0.72	0.73
2	Random Forest	0.81	0.75	0.71	0.72
3	KNN	0.76	0.68	0.67	0.67
4	Gradient Boost	0.81	0.75	0.71	0.73
5	ADA Boost	0.81	0.75	0.72	0.73
6	XG Boost	0.80	0.73	0.70	0.71

After tuning:

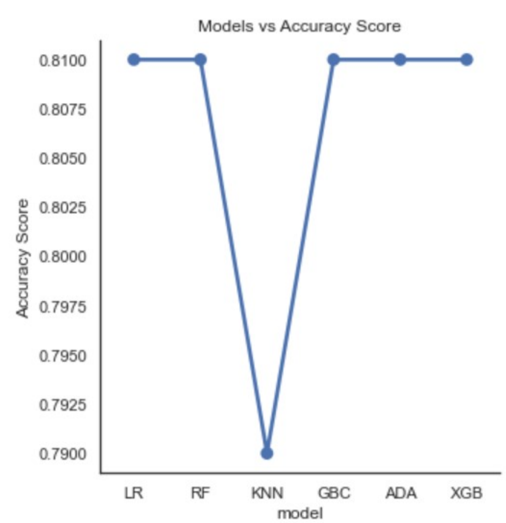
	Classification Model	Accuracy	Precision	Recall	F1-score
1	Logistic Regression	0.81	0.75	0.72	0.73
2	Random Forest	0.81	0.76	0.71	0.73
3	KNN	0.79	0.72	0.72	0.72
4	Gradient Boost	0.81	0.75	0.72	0.73
5	ADA Boost	0.81	0.75	0.72	0.73
6	XG Boost	0.81	0.75	0.72	0.73

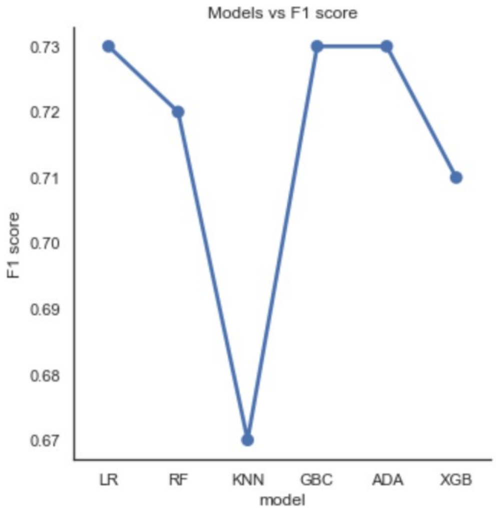
Before tuning:





After tuning:





From the above graphs, when we compare the results before tuning and after tuning, we can say that the hyperparameter tuning has only slightly improved for most of the models, but in the case of KNN, there is a significant improvement.

V. Future Directions

Utilizing Machine Learning algorithms to create more accurate predictive models: By using advanced Deep Learning algorithms such as RNN, CNN, xDeepFM etc telecom companies can create more accurate predictive models that can predict customer churn with greater accuracy.

Incorporating more data sources: Telecom companies should incorporate more data sources such as customer support tickets, customer feedback, customer surveys, and social media data to get a better understanding of customer behavior. This will enable them to make better decisions about customer retention, and ultimately reduce churn.

VI. Conclusion

After the data preprocessing and exploratory data analysis, insights on a lot of things from the data were drawn and our models have predicted them successfully. All the models except for KNN have produced good results. But the most preferred one would have to be Random Forest because it is fast and is based on decision trees. ADA Boost and XG Boost could also be used. For most reasonable cases, ADA Boost and XG boost will be significantly slower than a properly parallelized random forest.

References

[1] D. Anil Kumar and V. Ravi, "Predicting credit card customer churn in banks using data mining," *International Journal of Data Analysis Techniques and Strategies*, vol. 1, no. 1, pp. 4–28, 2008.

[2] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," *Expert systems with applications*, vol. 38, no. 12, pp. 15 273–15 285, 2011.

[3] A. Sharma and D. P. K. Panigrahi, "A neural network based approach for predicting customer churn in cellular network services," *arXiv.org*, 2013.

[4] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert systems with applications*, vol. 39, no. 1, pp. 1414–1425, 2012.

[5] M. Makhtar, S. Nafis, M. Mohamed, M. Awang, M. Rahman, and M. Deris, "Churn classification model for local telecommunication company based on rough set theory," *Journal of fundamental and applied sciences*, vol. 9, no. 6S, pp. 854–, 2018.

[6] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert systems with applications*, vol. 36, no. 3, pp. 4626–4636, 2009.

[7] X. Zhu, X. Liu, A. Huang, J. Yuan, and W. Deng, "Impacts of sst configuration on monthly prediction of western north pacific summer monsoon in coupled and uncoupled models," *Climate dynamics*, vol. 59, no. 5-6, pp. 1687–1702, 2022.