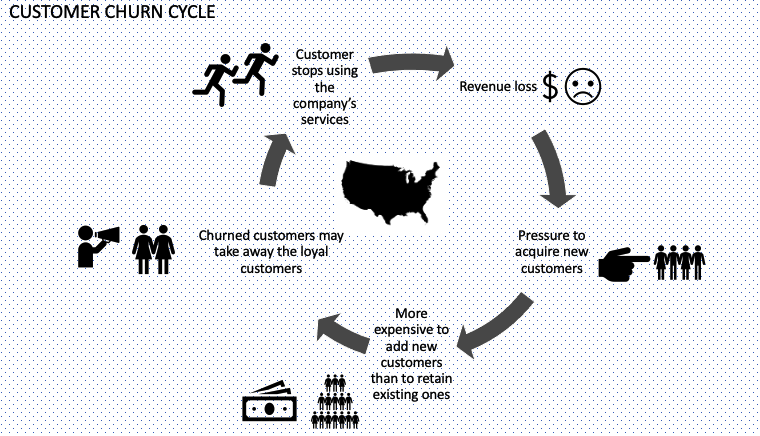
**Capstone Project**

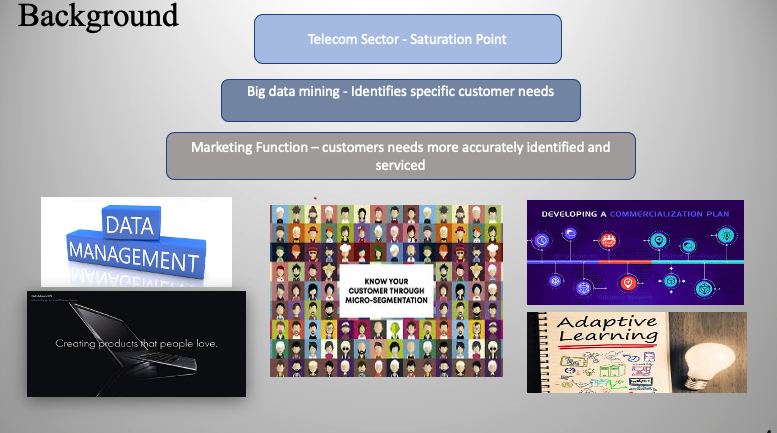
**Customer Churn Prediction Using Machine Learning**

**Problem Statement**  
An American telecom provider wants to explore the possibilities of machine learning models with highest accuracy in predicting their customers in advance who would be more likely to quit the subscription. And also, to determine the factors influencing them to churn.

Basically, in order to increase profits, a business needs to retain its existing customers and increase new customers. In spite of more initiative of reducing tariffs and promoting more offers, the churn rate is well above due to its saturated market. The customers are migrating to our competitors.

Desired state of the project is to make a model with the highest accuracy predicting on the existing customers in advance who are more likely to quit the subscription. And to determine the factors which are influencing to churn.



****

**The telecom sector has reached saturation point** in many emerging markets, with market penetration sometimes higher than 100 percent of the population. Companies are slowly transitioning towards identifying specific customer needs by employing analytical marketing, mining the Big Data to respond continually to the behaviour of individuals and market micro- segments.

Telecom operators are building five key capabilities within the marketing function: data management, micro- segmentation, commercialization, product design, and adaptive learning. Deploying these capabilities will ensure that customers’ needs are more accurately identified and serviced. Customers will be more likely to respond positively to marketing approaches.

Customer churn prediction using machine learning can be applied in various domain such as pay TV, Insurance companies, bank, Alarm Security Monitoring Companies.

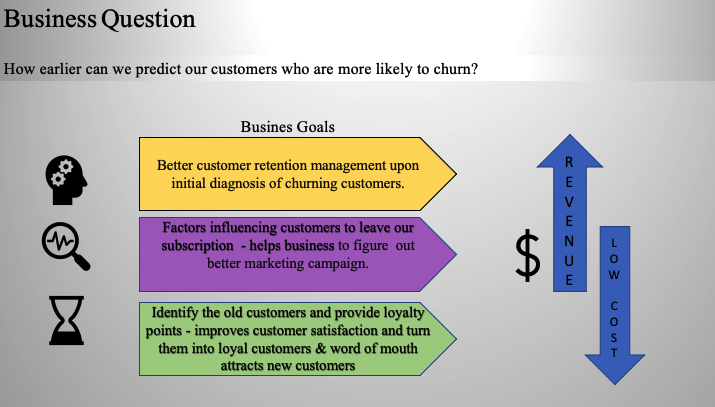
**Stakeholders:**

Internal stakeholders’- employees, managers, the board of directors and investors.

External stakeholders - consumers, regulators, investors and suppliers.

### **Business question**

How earlier can we predict our customers who are more likely to churn?



The business wants to explore the possibilities of machine learning by creating a model which predicts the customers in advance who will quit the subscription.

This enables the business to improve the customer retention management.

And through this model it can also be identified the factors which is influencing the customers to churn and helps the business to figure out better marketing campaign.

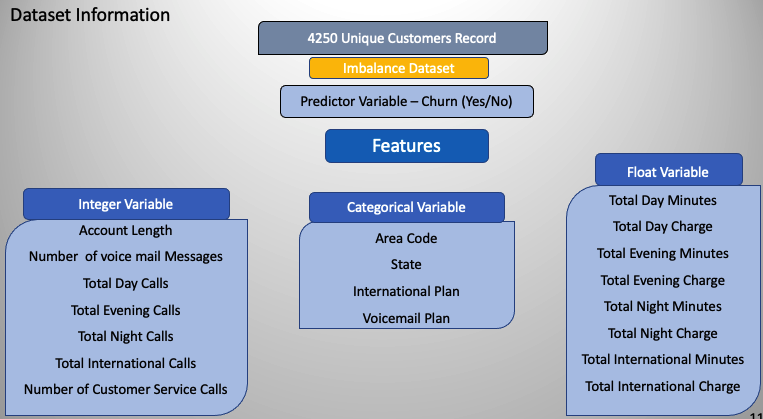
By knowing these factors, the old customers are provided some sort of loyalty benefits contributing customer satisfaction. And also, positive word of mouth attracts new customers.

**Data question**

1.Based on the historical customer data trying various machine learning algorithms which model could give us the highest accuracy on predicting the customers who are more likely to quit the subscription?

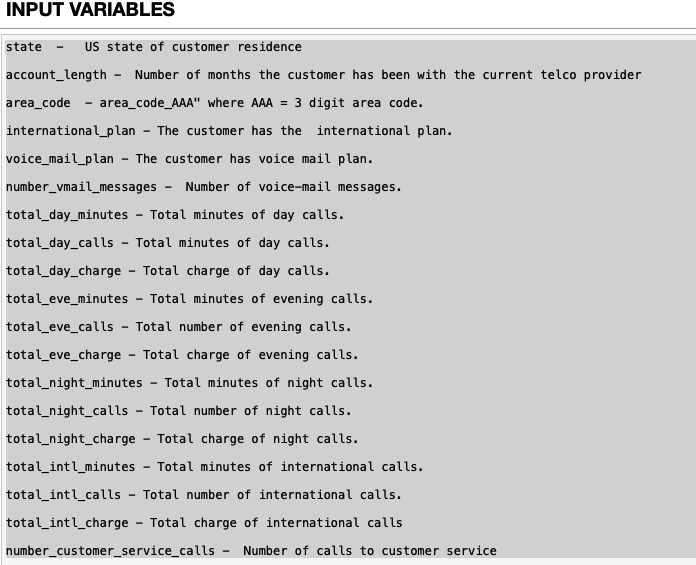
2.What are the underlying factors which influences the customers to churn.

**About the data**

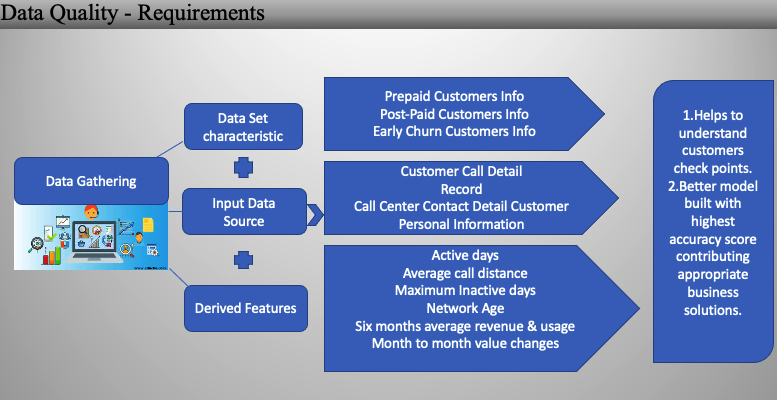


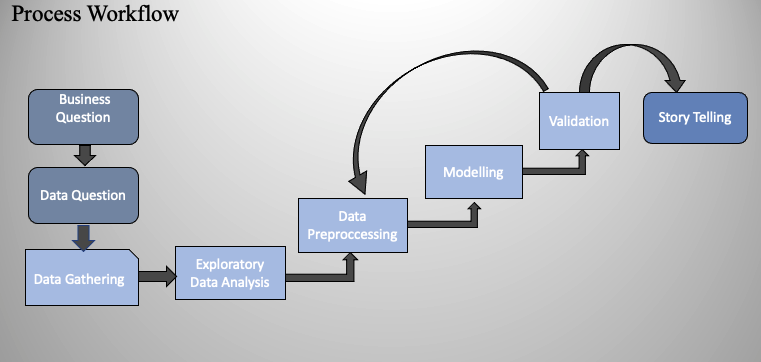
We are able to obtain the data from the Kaggle competition which is more sufficient for this project. Furthermore, the   data gathering will be a challenging part in any kind of data science project for making accurate predictions which in turn contributes to a better solution for the business problem. However, the better solution always depends upon the quality of the data. Clear line of communication and discussion with the domain expertise which plays a pivotal role in the process of quality data gathering.

For the further future enhancement for this project, we require may require more features to obtain the right prediction.

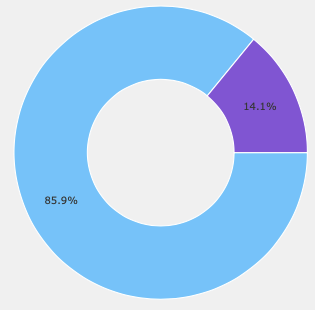


**Data requirements in future**

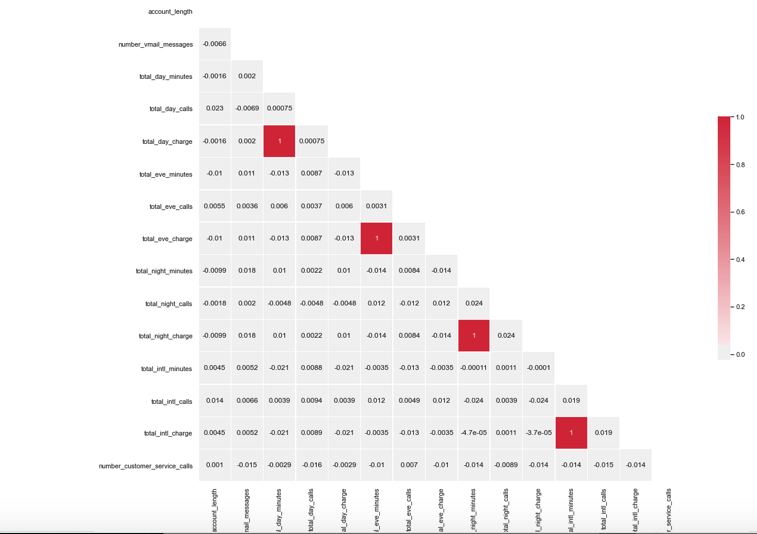
****



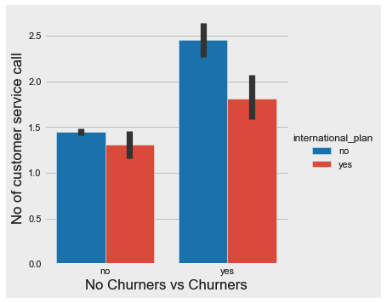
**Highlights of Exploratory Data Analysis**



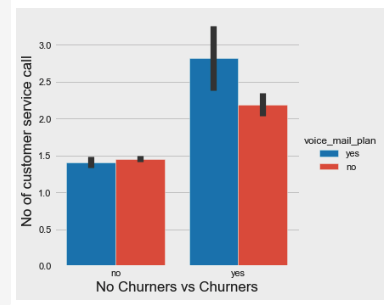
85% of non-churners and 14.1 % of Churners



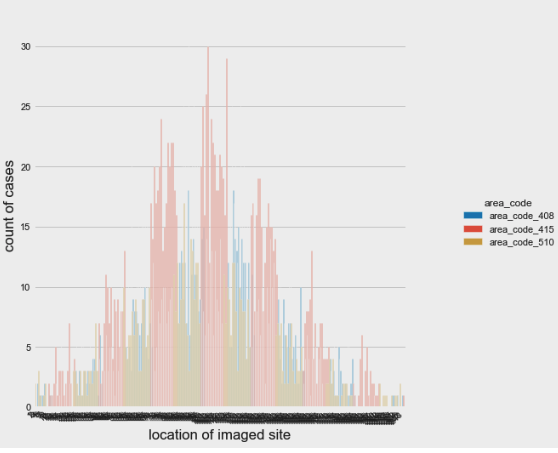
Variables correlated (total\_international minutes is highly correlated with total\_int\_charges, total\_night\_minutes with tatal\_night\_charge, total\_eve\_minutes withtotal evening charge, total\_day\_minutes with total\_day\_charge,total charge with total\_dat\_minutes,total\_day\_charge,total\_eve\_minutes,total\_night\_minutes,total night charge).The light box are positively correlated variable with corresponding variables.



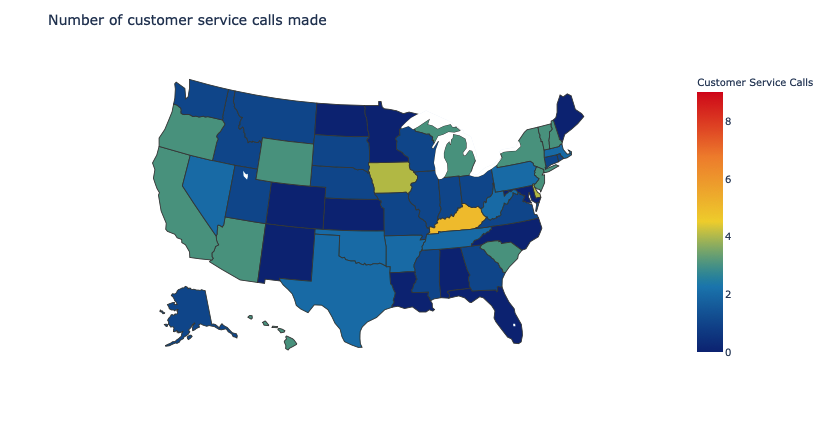
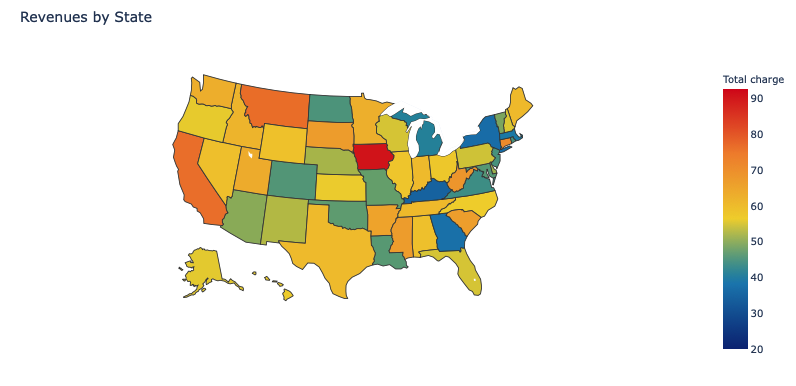
Non-International plan customers are the churners & high customer service call

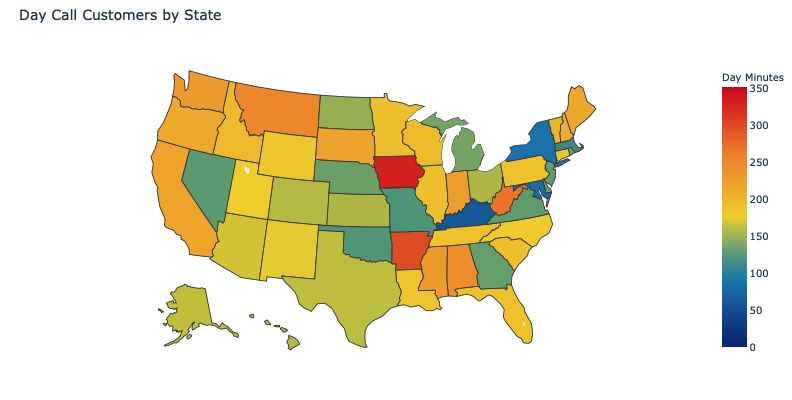


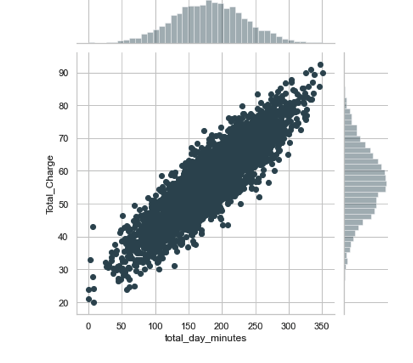
Customers with voice mail plan are churners & with high customer service calls



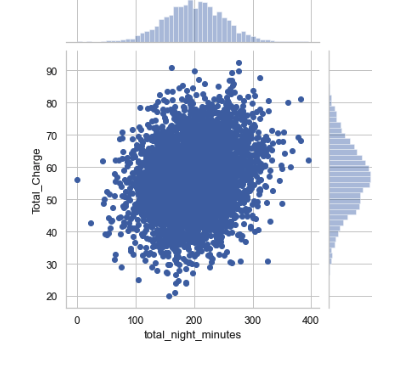
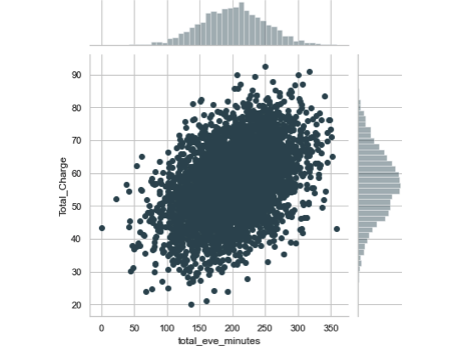
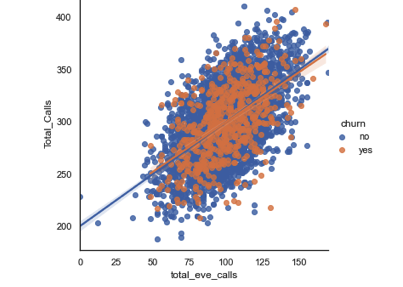
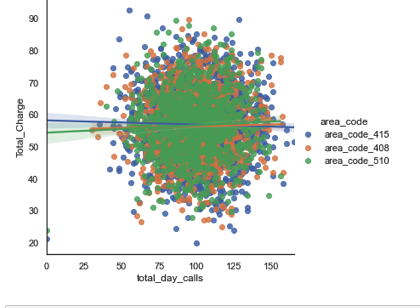
More churners are from the area code 415







There is more correlation between the total day minutes and the total charge. The business has to create a plan to attract the customers who spend more minutes during day time on calls to attract them and turn them to be a loyal customer.



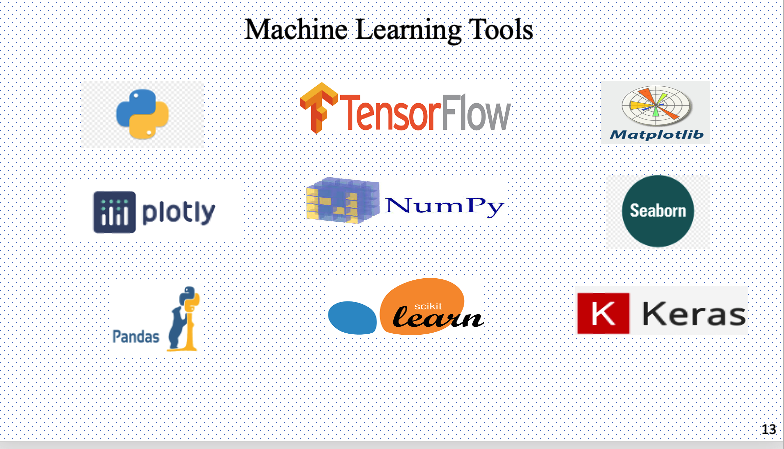
**Data Transformation**

### Standardizing Categorical Columns using label encoder ()

### Normalize Numeric Columns using MinMaxScaler ()

* Create Dummy Variables for state and area code

**Machine Learning Tools**



## **Modelling**

Models experimented including all the features with SMOTE & without SMOTE with best optimal parameters. In total 30 models are performed with all different configuration.

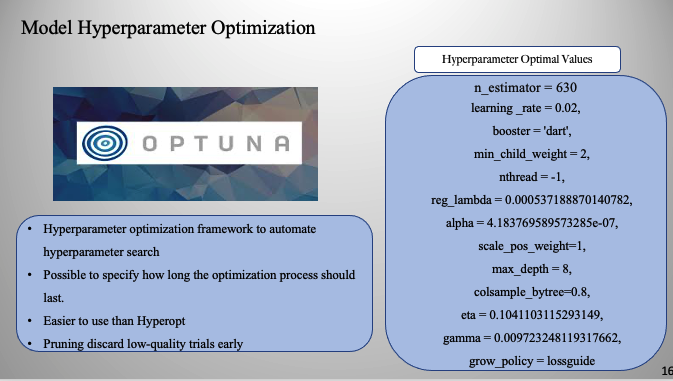
The algorithms namely

1. Support Vector Machines
2. Logistic Regression
3. GaussianNB
4. Decision tree Classifier
5. Random Forest Classifier
6. ADABoost
7. Keras Classifier
8. XGBoost Classifier
9. CatBoost Classifier

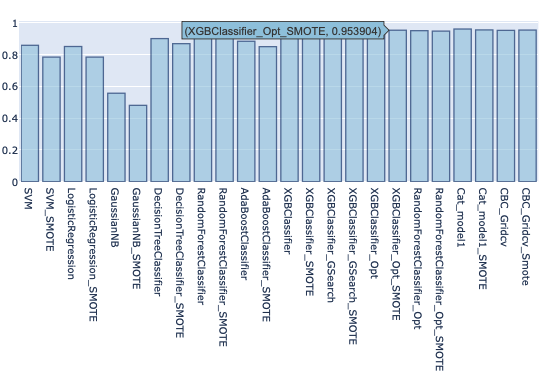
**SMOTE– Synthetic Minority Over-sampling Technique**

* Synthetic samples are generated for the minority class which balances the class distribution and improves the model performance by not being biased to majority class.
* Over sampling the minority class of training data prior to fitting a model.
* Overcomes the overfitting problem posed by random oversampling.

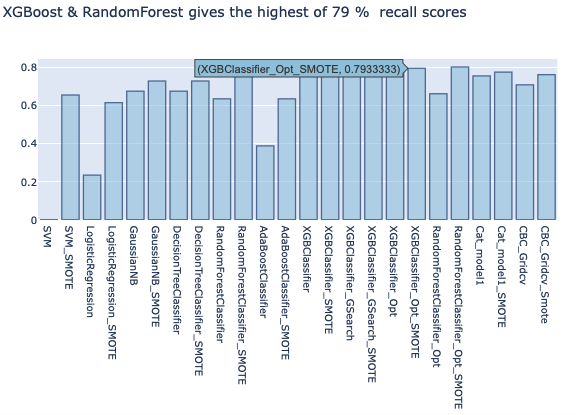
**Model Hyperparameter Optimization using the framework called OPTUNA**

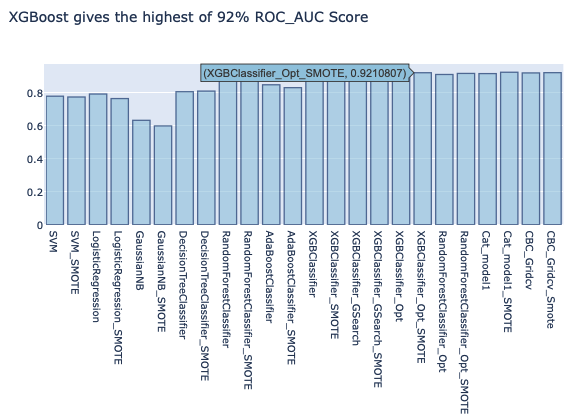
****

**Model Selected**

* XGBoost Model with Smote oversampling technique using optuna hypermeter tuning gave the highest accuracy in classifying churners with the least number of False Negatives. 

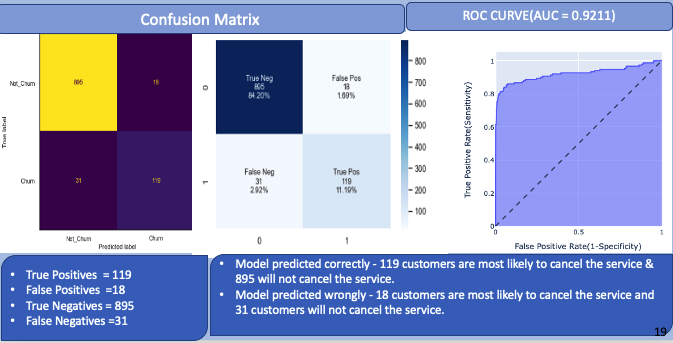
XGBoost, Random Forrest & Cat boost gives the highest of 79% recall scores.



XGBoost, Random Forrest, CatBoost & Keras gives the highest of 91% ROC-AUC Scores

**MODEL EVALUATION AND SELECTION**

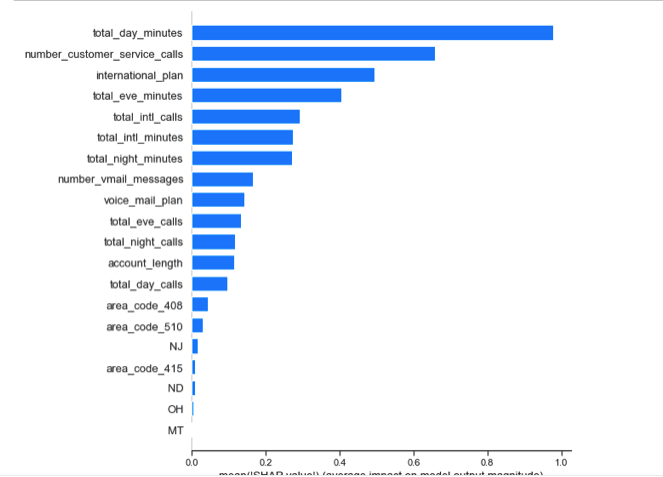
Two metrics are been used to evaluate the model performance. They are confusion Matrix and ROC-AUC curve.

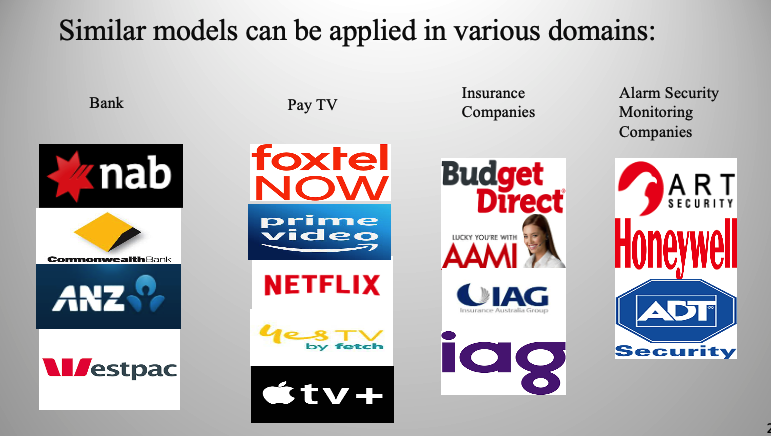


**Shap Values**

Top 13 factors influencing customers to churn using shap values

* Shapley Additive explanation
* Created by computer scientist to enable the accuracy and the interpretability from machine learning models.
* SHAP assigns each feature variable a value
* Higher values indicate higher importance in predicting the target variable.



****

**Summary**

* XGBoost model gave the best accuracy rate, predicting the customers who are most likely to churn. Business can now focus on customer retention campaigns toward targeted high-risk customers.
* Through XGBoost SHAP values I determine the model performance and the feature importance which was influencing the customers to quit the subscription.
* Future Work: Adding more features like customer complain type, resolution time, demographic information of the subscriber & sentiment analysis on customer feedback for better insights.

**References**

* Dataset is from Kaggle Competition 2020: https://www.kaggle.com/c/customer-churn-prediction-2020
* https://www.strategyand.pwc.com/m1/en/reports/hitting-the-target.pdf
* https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/
* https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d
* https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/