Data Analytics

On Customer Churn Data

Link to Notebook: https://colab.research.google.com/drive/1fLA8P0twYx2-q0K7OZHL6e7P5gF3zgyz

Agenda Style

01 Project Objective and Business Problem

02 Descriptive Analytics

03 Predictive Analytics

04 Prescriptive Analytics

05 Cognitive Analytics

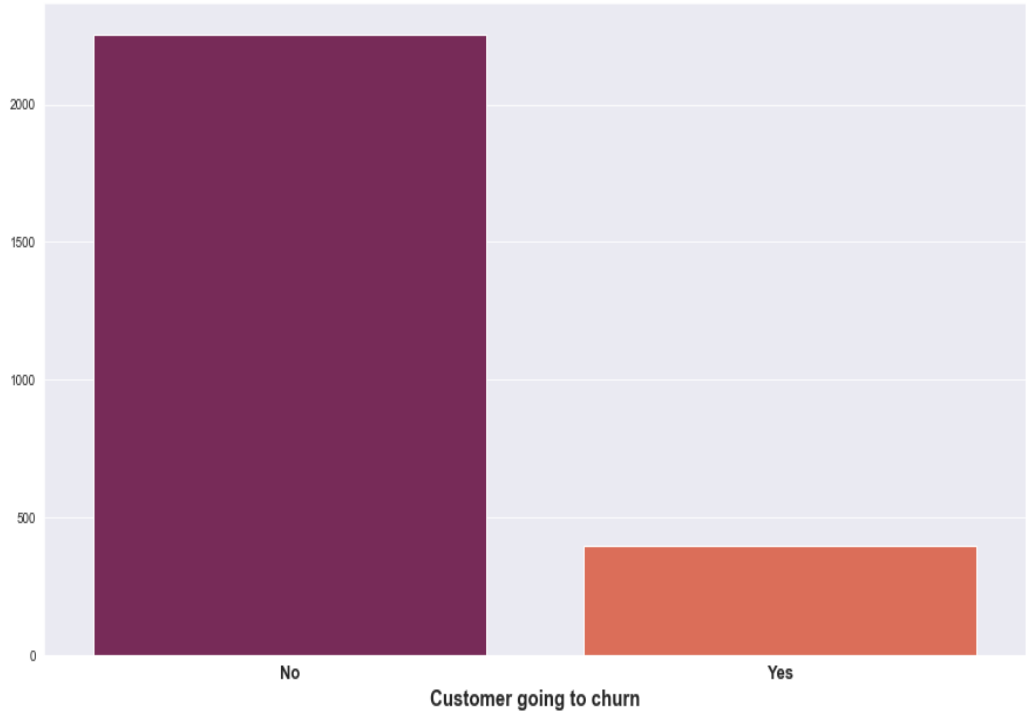
Business Problem

And Project Objectives

The Database for this Project comprises of two datasets. One used for training while the other used for testing purposes.

The Data has ‘Churn’ information about customers, which is dependent on 20 other independent attributes.

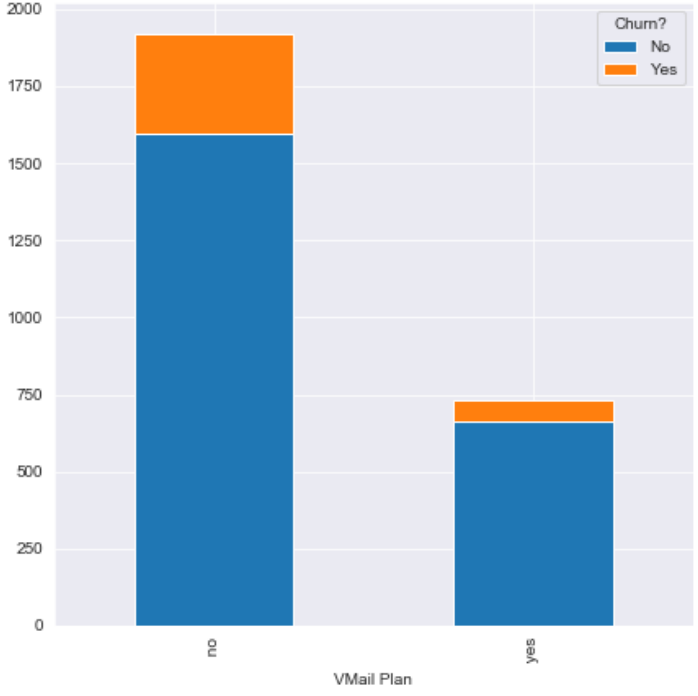
The main business problem is to predict this Churn behavior and deduce plausible explanations for customer churn.



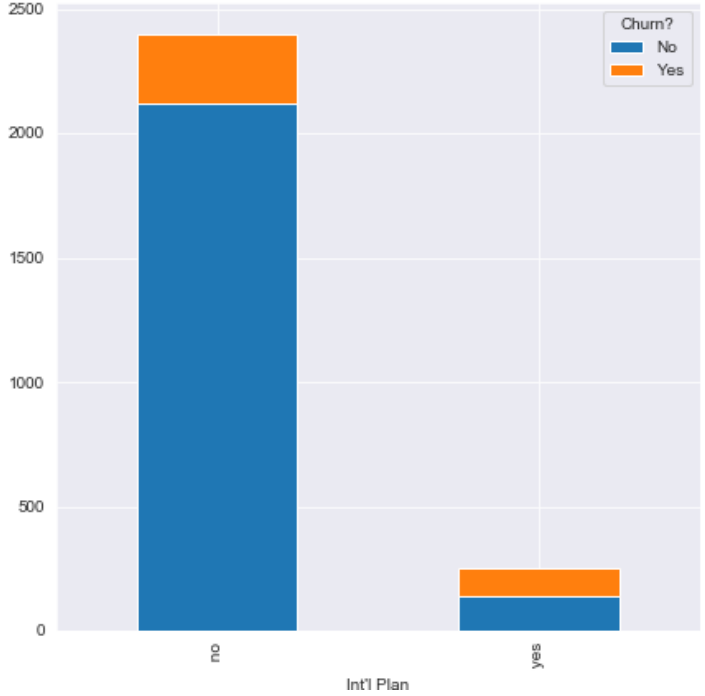
Descriptive Analytics

The story so far…

17% of customers having no Voice Mail Plan churned as opposed to 9% of those who had.

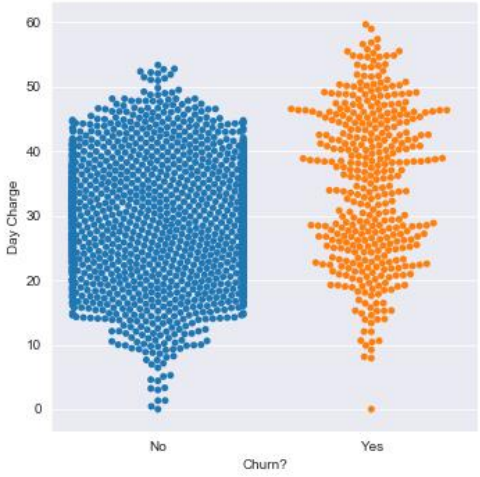


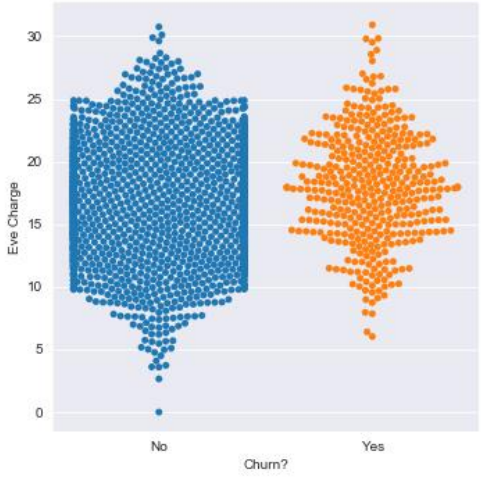
A staggering 45.2% of customers having International Plan churned as opposed to 11.75% of customers who did not have International Plan.

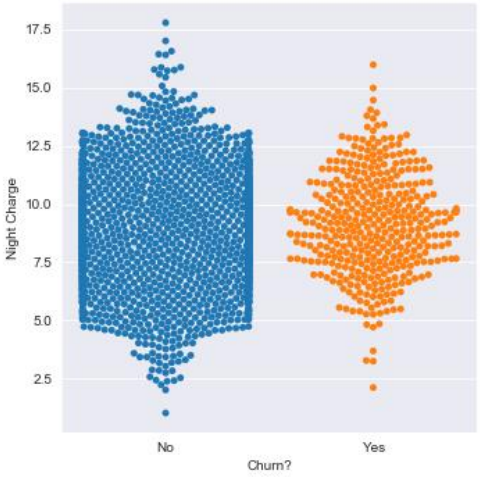


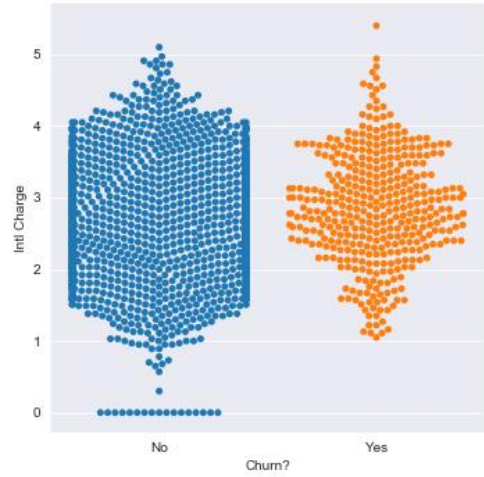
The data is a highly imbalanced dataset with only 14.9% of the classes pertaining to customers who have churned

The distribution of data shown in the plots here, indicate that the rate of churn does not have a high correlation with the call charges.



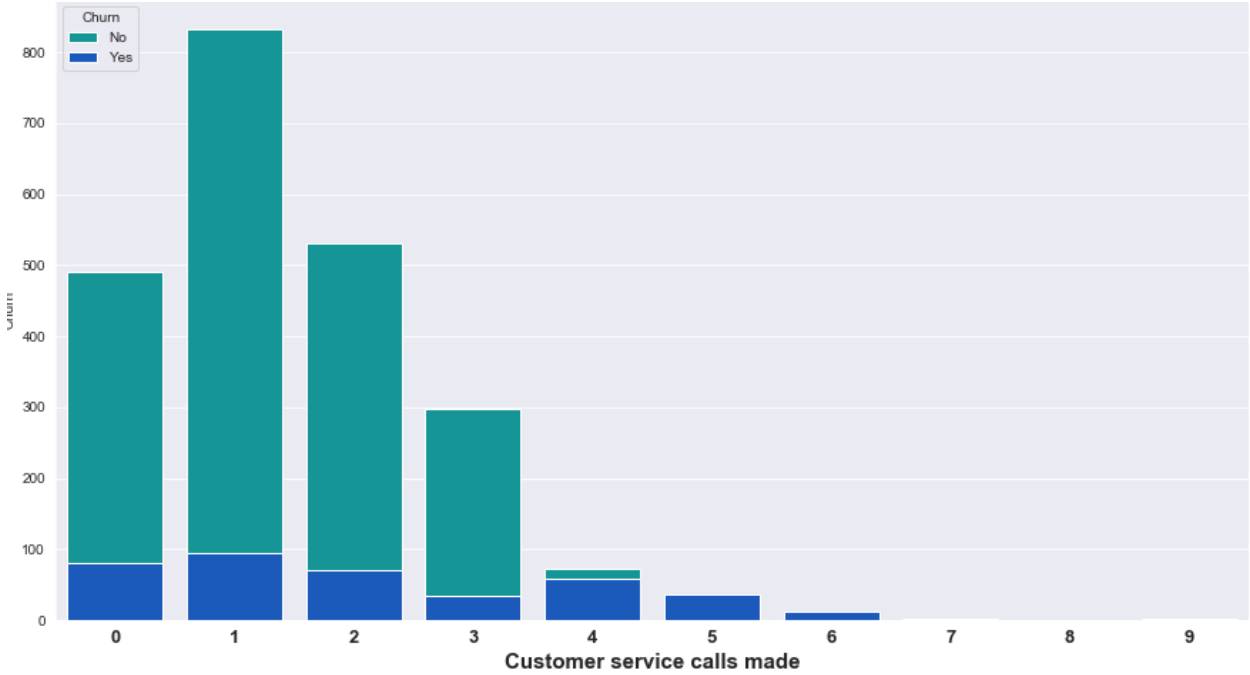






Customers, who have had to call the Customer Service 4 times or more have churned in large proportions (at least 44.7%).

CustServ Calls	No	Yes	Ratio
0	491	80	0.14
1	833	94	0.10
2	531	71	0.11
3	297	35	0.10
4	73	59	0.44
5	19	37	0.66
6	7	13	0.65
7	3	3	0.50
8	1	1	0.50
9	0	2	1.00



Predictive Analytics

Using machine learning

I

Data-Set

The data set comprises of Training and testing Data. ‘Churn’ is the dependent variable while there are 20 independent variables.

II

Data Preprocessing

Label-encoding for ordinal data, Scaling of variables and making dummy variables.

The imbalanced representation of classes in the data set was removed by using SMOTE oversampling.

III

Model Selection

To select which model performs best for the given dataset, we train a host of different models, including logistic regression, support vector machines, random forest etc.

IV

Model Performance

Model performance is calculated on the Testing data with various metrics like accuracy, precision, recall etc.

Models Used

01

Gradient boosting is a machine learning technique for regression and classification problems, which produc es a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

02

Support-Vector Machines are supervised learn ing models that use a set of mathematical func tions that are defined as the kernel. The functi on of kernel is to take data as input and transf orm it into the required form.

03

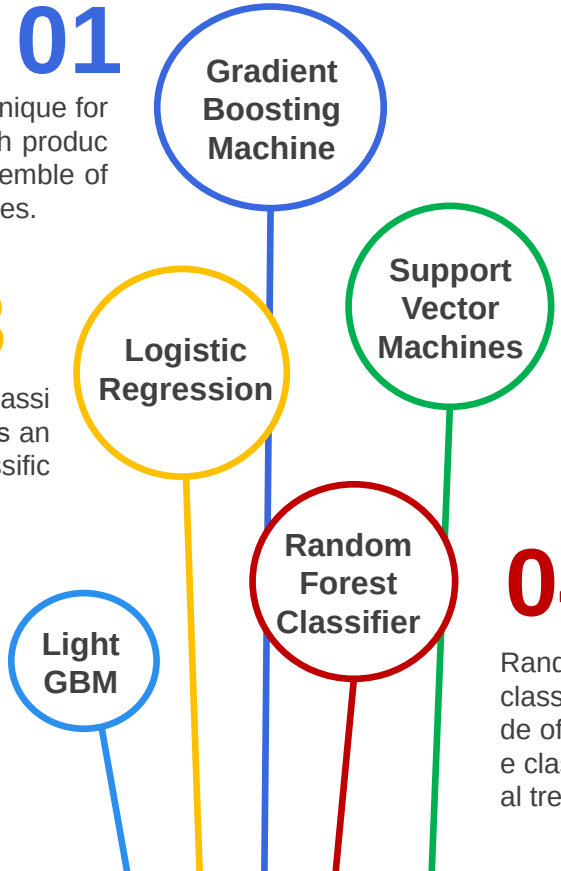
Logistic regression models the probabilities for classi fication problems with two possible outcomes. It's an extension of the linear regression model for classific ation problems.

04

Random forests is an ensemble learning method for classification, that operates by constructing a multitu de of decision trees at training time and outputting th e class that is the mode of the classes of the individu al trees.

05

Light GBM is a gradient boosting framework that uses tree based learning algorithms. It is design ed to be distributed and efficient.



Model Performance

This data is obtained by performing testing on the test data set

Models	Accuracy	Recall Majority	Recall Minority	AUC Score
Gradient Boosted Machine	43%	95%	35%	89%
Logistic regression	76%	74%	76%	82%
Support Vector Machine	86%	80%	87%	89%
Random Forest Classifier	82%	85%	82%	90%
Light GBM	41%	98%	33%	85%

Highest Accuracy

The highest Accuracy score is achieved on **SVM** Model

Highest Majority Recall

The highest Minority recall score is achieved on **Light GBM**

Highest Minority Recall

The highest Minority recall score is achieved on **SVM**

AUC Score

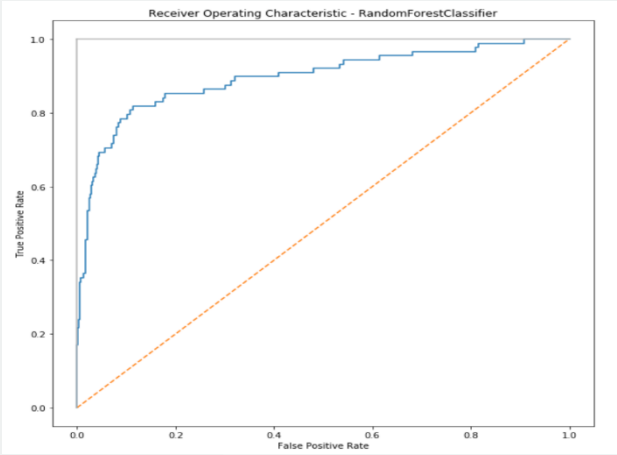
The best AUC score is achieved from **Random Forest Classifier**

Overall Performance

When it comes to overall performance, **Random Forest Classifier** is the best

Model Selection

Random Forest Classifier is the best model out of all the models tested as it has the best auc-roc score as well as the best overall Recall for both the minority as well as majority class. So, it can overall correctly predict people who will churn & those who will not churn.



Best Auc Roc Score

ROC is a probability curve and AUC re presents degree or measure of separa bility. It tells how much model is capabl e of distinguishing between classes.

Best overall Recall Score

Recall score tells us out of all the people who were going to churn (or not), how many were we able to accurately predict.

Classification Report:

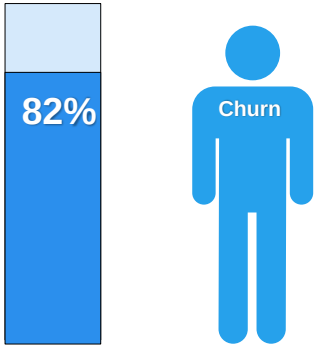
	precision	recall	f1-score	support
0	0.97	0.82	0.89	595
1	0.41	0.85	0.56	88
accuracy			0.82	683
macro avg	0.69	0.84	0.72	683
weighted avg	0.90	0.82	0.85	683

Confusion Matrix:

```
[[488 107]
 [ 13  75]]
```

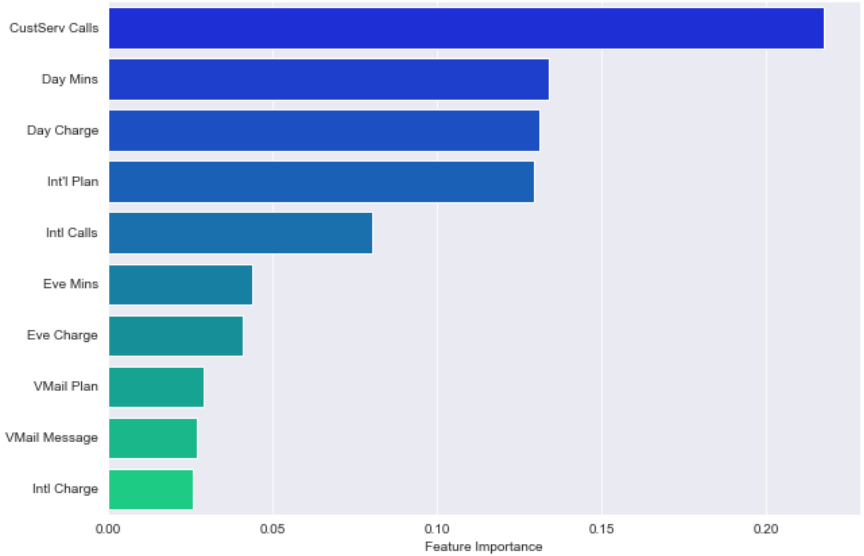
Prescriptive Analytics

Conclusion of Project and Outcomes



- We are able to predict whether a Customer is going to churn or not, with an accuracy of 82%.
- We were able to develop Machine Learning models and critically compare them on the basis of various performance metrics.

- We are also able to identify 83-85% of all the people who are going to churn from our customer base.
- We were also able to identify the most important features that have high correlation with our dependent variable.
- From the chart shown here, we see that CustServ Calls, Daily Mins and Day Charge are the most important features governing the customer behavior.



Cognitive Analytics

The next best course of action

1

We have seen that Customers with an international call plan are much more likely to churn We must plan to offer a better **pricing policy** to customers with international plans.

2

Customers who have not used our voice mail services show a higher propensity to churn out. Thus more customers needs to be introduced to voice mail.

3

Customers who had the need to call the customer service more than 4 times, are extremely likely to churn. We must work on our troubleshooting and **grievance redressal mechanism** and come up with a better approach.

4

Finally we have seen that day charges and day minutes are highly correlated with churning, thus we need to **monitor** these variable more carefully for every customer

Thank You