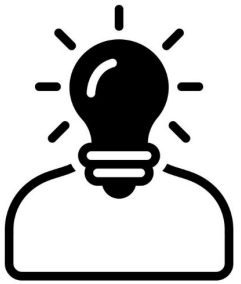
Predictive Modeling (Classification)

Ethan Duong

DECISION TREE ANALYSIS AND NAÏVE BAYES

Data Understanding



Data origin: The dataset belongs to a phone service provider company. It contains information about customers' phong usage and the cost for using their service.

Business problem: Customer churn or customer attrition means the loss of customers for a company. The problem for customer churn is that the company would like to know in advance which customers would churn in near future.

Business goal: help this company in characterizing customer churn through data analytics methods.

Dataset summary: 21 attributes including a binary class attribute about churn.

Test and Validation Sets

- Test and training data sets are two subsets of a larger data set that is used when building and evaluating a machine learning model.
- The training data set is used to train the model, while the test data set is used to evaluate the performance of the trained model.
- The goal of dividing a data set into training and test sets is to ensure that the model is able to generalize well to unseen data.
- This means that the model should not only be able to make accurate predictions on the data it was trained on, but also on new data that it has not seen before.
- Using a test set allows us to evaluate the performance of the trained model on data that it has not seen before, and thus get a better sense of its true predictive power.
- This is important because a model that performs well on the training data may not necessarily perform well on new data, a phenomenon known as overfitting.

General Principles

- In general, the larger the training set, the more information the model has to learn from, and the better it should be able to perform on the test set.
- However, it is also important to have a sufficiently large test set in order to properly evaluate the model's performance.
- In summary, the training and test data sets are two important tools in the machine learning workflow, as they allow us to train and evaluate our models in a way that ensures they can generalize well to unseen data."

Best Practices and Procedure used for creating test and training sets

1. Ensured that the data was representative of the problem being solved. The data accurately reflected the types of inputs and outputs that the model would encounter in real-world use.

2. Used a sufficiently large data set. In general, the larger the data set, the more information the model had to learn from, and the better it was able to perform. However, it was also important to have a sufficiently large test set in order to properly evaluate the model's performance.

3. Avoided sampling bias. Made sure that the training and test sets were drawn from the same population, and that they were representative of the full range of possible inputs and outputs.

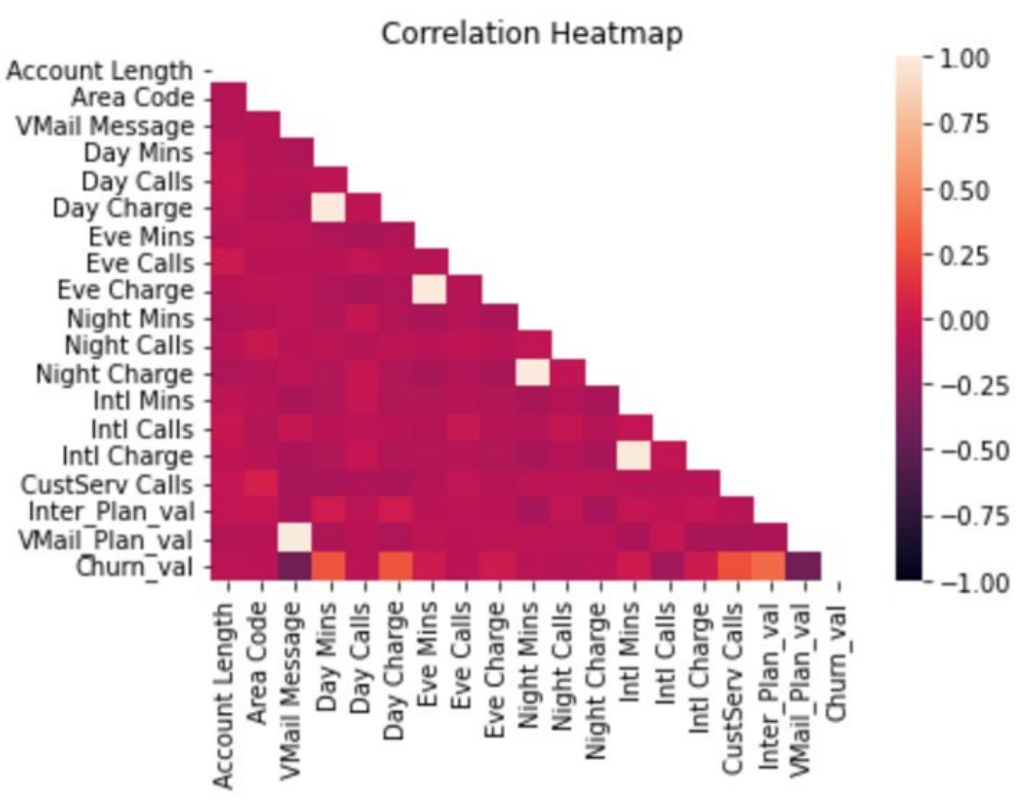
4. Randomly split the data into training and test sets. This ensured that the model was not overfitted to the training data, and that it could generalize well to new, unseen data.

5. Used stratified sampling when splitting the data. This technique ensured that the training and test sets had the same proportions of different classes or categories as the full data set, which was important for certain types of machine learning models.

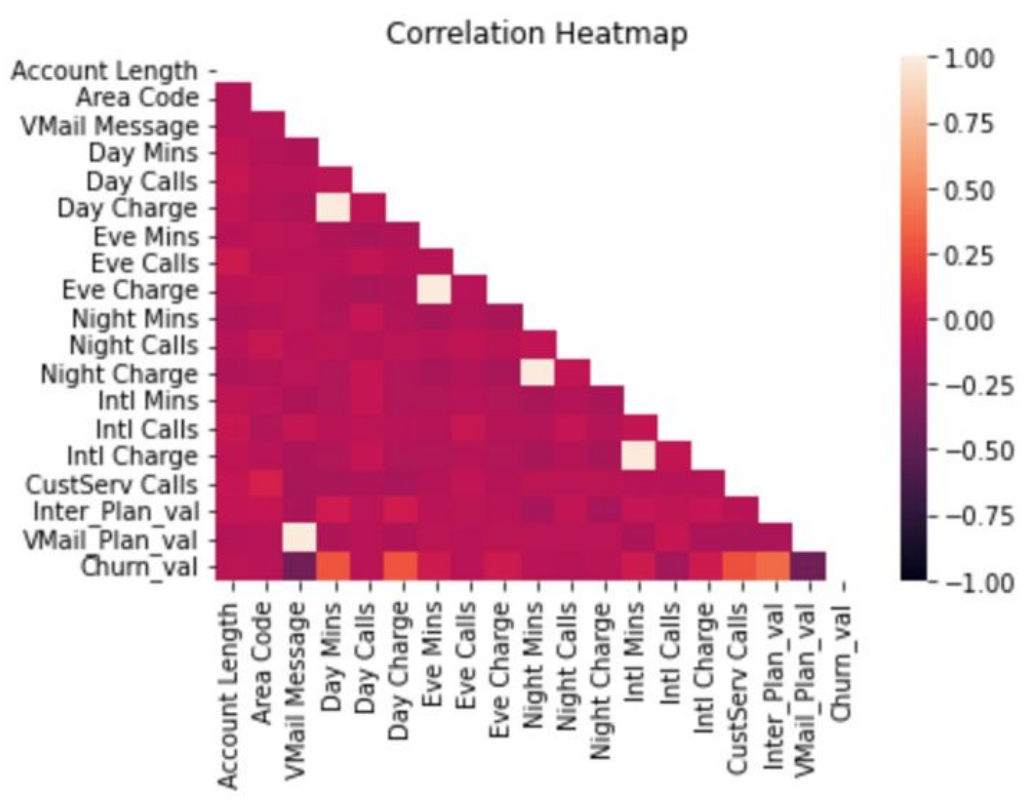
6. Overall, the goal when creating test and training data sets was to ensure that the model was able to generalize well to unseen data, and that its performance could be accurately evaluated. By following these best practices, you could help ensure that your machine learning models were robust and reliable

Correlation Heat map (Test vs Training)

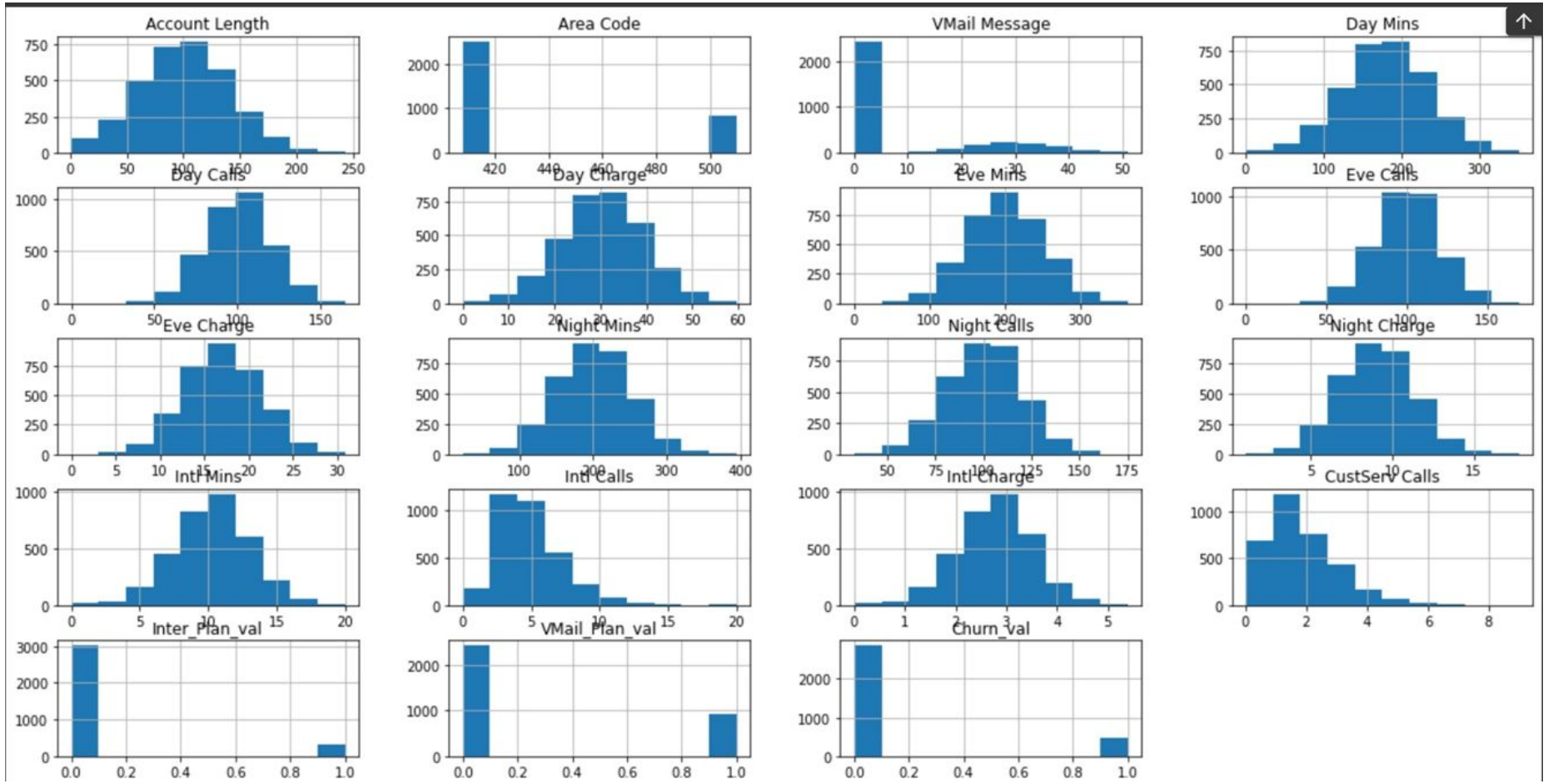
TEST



TRAIN



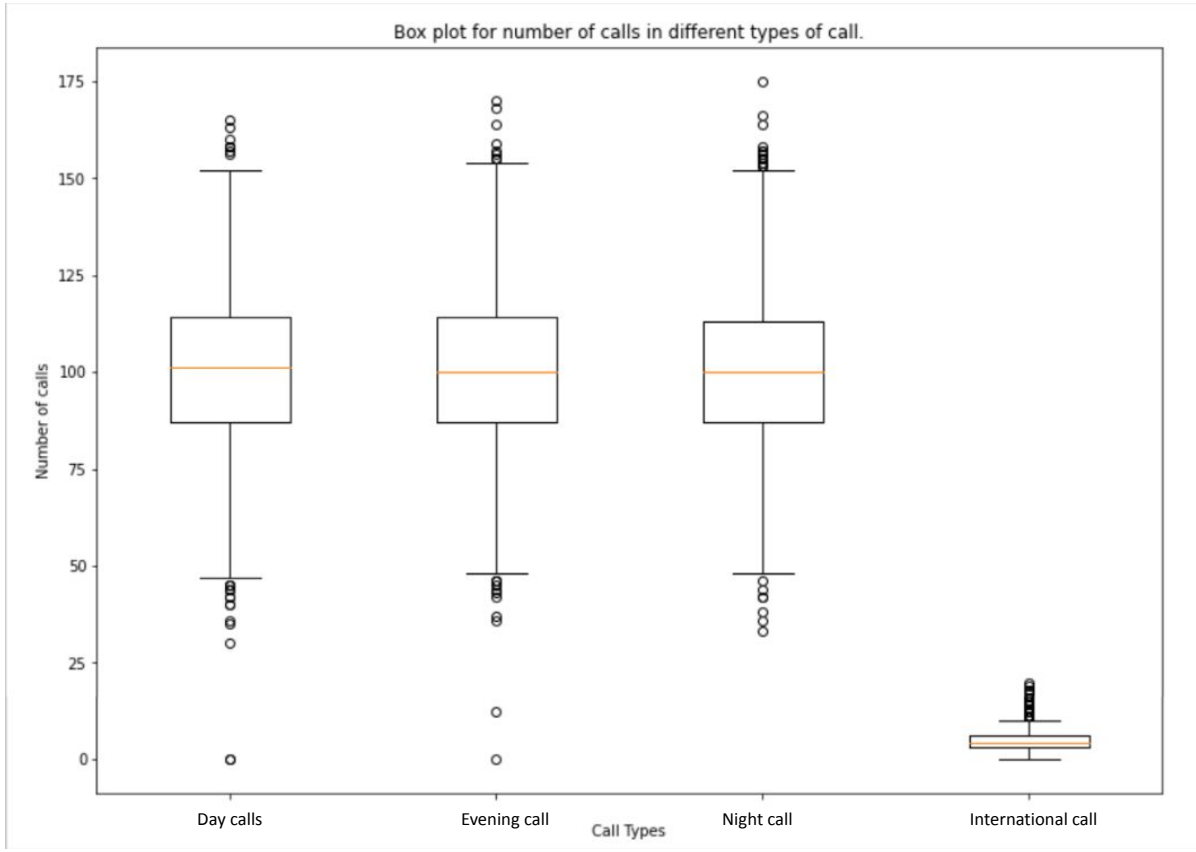
Data Histograms



Outliers

- A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data.
- Outliers for number of calls

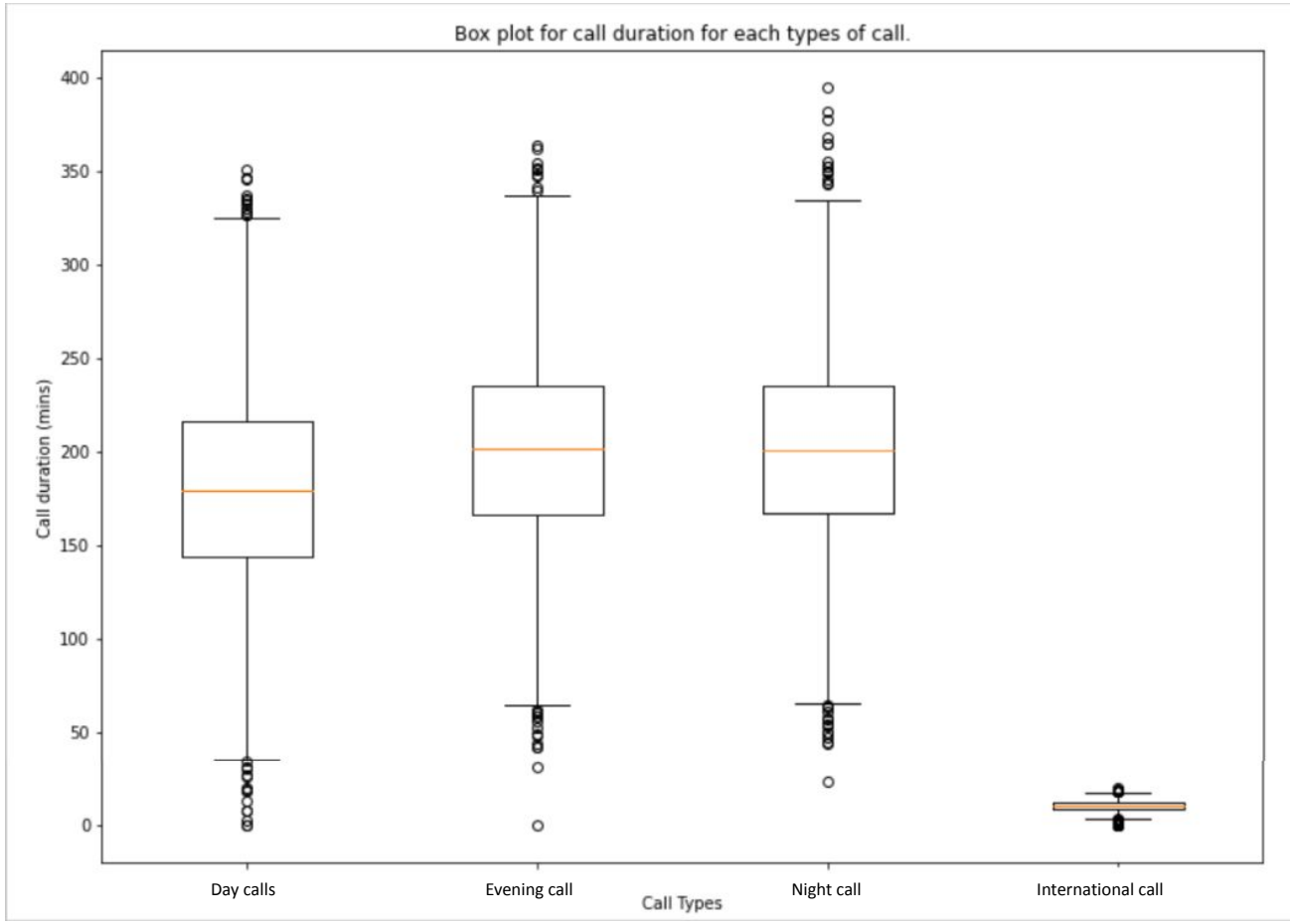
Types of call	Number of lower outliers	Number of Upper outliers
Day	5	4
Evening	4	3
Night	3	3
International	0	50



Outliers

- Outliers for call duration (mins)

Types of call	Number of lower outliers	Number of upper outliers
Day	5	4
Evening	6	3
Night	4	7
International	20	2



What is Classification?

- Machine learning is at the core of Data Science; it helps stakeholders make predictions about the future
- "Classification refers to a predictive modeling problem where a class label is predictive modeling problem where a class label is predicted for a given example of input data" (Machine Learning Mastery, 2022).[1]
- One of the most popular types of classification is "Binary classification." Where the classification tasks have two class labels.
- Customer churn is an example of binary classification, with two states of the predicted results: True and Flase
- Class for normal state = "TRUE"
- Class for abnormal state = "FALSE"

Data splitting strategy

- Splitting existing data is an incredibly important decision data scientists have to make prior to data modeling
- Data will be split into two groups of the training and testing (validation) sets to align with the classification modeling
- The data split strategy consists of 70% training and 30% testing split method

Classification using Decision Tree (Supervised Learning)

- The models used to perform the Classification algorithm are - Decision Tree and Naïve Bayes.
- In the Decision Tree model we initially used the original dataset with all attributes and ran the algorithm, where we got an accuracy of 91%. The original dataset is unbalanced, as churn is considered a rare event compared to not churn and this baseline model is overfitting.
- We then rerun the algorithm with selected attributes (balanced data). The attributes like State, Account Length, Area Code, Phone, were removed as they were deemed not relevant for the churn decision. The accuracy decreased to 86% and the tree looks more fitting with the balanced attributes dataset.

Classification using Decision Tree (Baseline)

Confusion Matrix

		Predicted	
		NO	YES
Actual	NO	TN = 831	FP = 24
	YES	FN = 68	TP = 77

Classification – Performance Result

	Precision	Recall	F1-Score	Support
FALSE	0.92	0.97	0.95	855
TRUE	**0.76**	**0.53**	0.63	145
ACCURACY			**0.91**	1000
Macro avg.	0.84	0.75	0.79	1000
Weighted avg.	0.90	0.91	0.90	1000

Classification using Decision Tree (Balanced)

Confusion Matrix

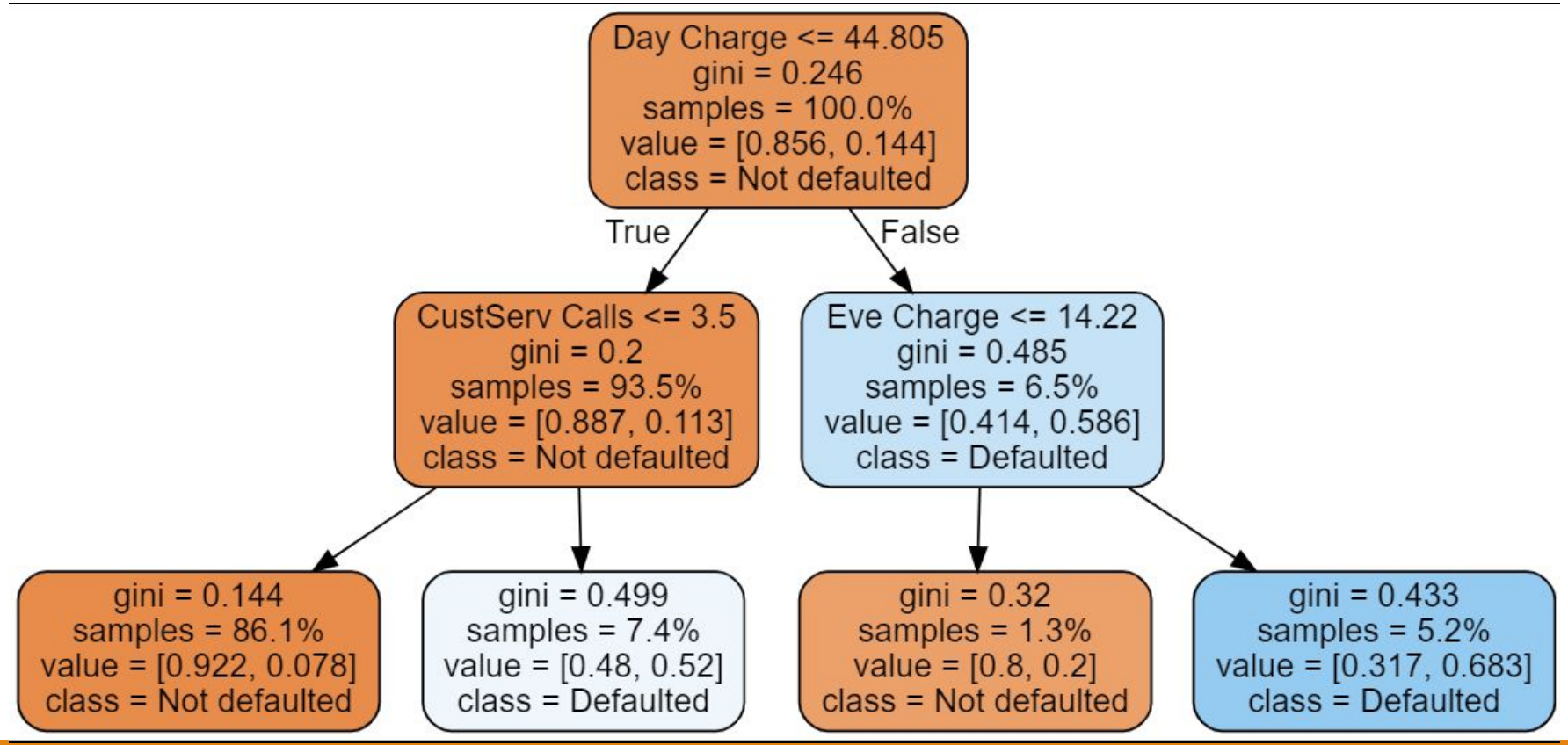
		Predicted	
		NO	YES
Actual	NO	TN = 832	FP = 23
	YES	FN = 75	TP = 70

Classification – Performance Result

	Precision	Recall	F1-Score	Support
FALSE	0.86	0.99	0.92	856
TRUE	**0.75**	**0.48**	0.56	144
ACCURACY			**0.86**	1000
Macro avg.	0.69	0.54	0.53	1000
Weighted avg.	0.82	0.86	0.81	1000

Classification using Decision Tree (Balanced)

Decision Tree



Classification using Naïve Bayes(Unsupervised Learning)

- The Naïve Bayes classification algorithm is based upon Bayes' Theorem
- Essentially Naïve Bayes classifiers measure the conditional probabilities of each class by using their counts/frequencies in each record, and predicts the class with the highest probability
- Here we are using Multinomial Naïve Bayes-" "where the features are assumed to be generated from a simple multinomial distribution. The multinomial distribution describes the probability of observing counts among a number of categories" (Python Data Science Handbook, 2022)[2].

Naïve Bayes Baseline model (All features)

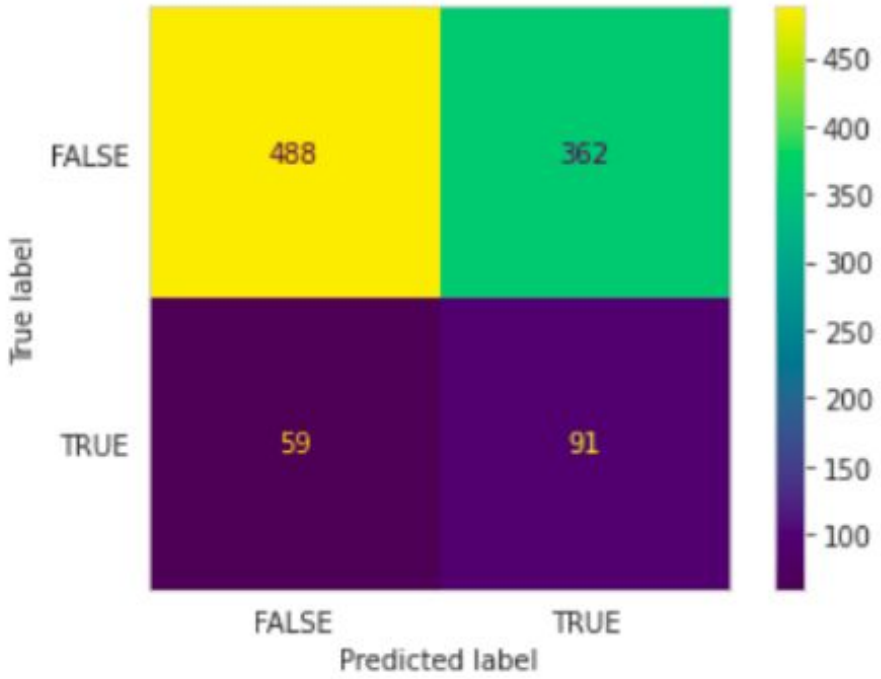
Performance Report

	Precision	Recall	F1-Score	Support
FALSE	0.89	0.57	0.70	850
TRUE	0.20	0.61	0.30	150
ACCURACY			0.58	1000
Macro avg	0.55	0.59	0.50	1000
Weighted avg	0.79	0.58	0.64	1000

Accuracy- 0.58
Precision (True Positive rate)- 0.20
Recall (False positive rate)- 0.61

Confusion Matrix plot

The confusion matrix plot:



Confusion matrix (Highly correlated features)

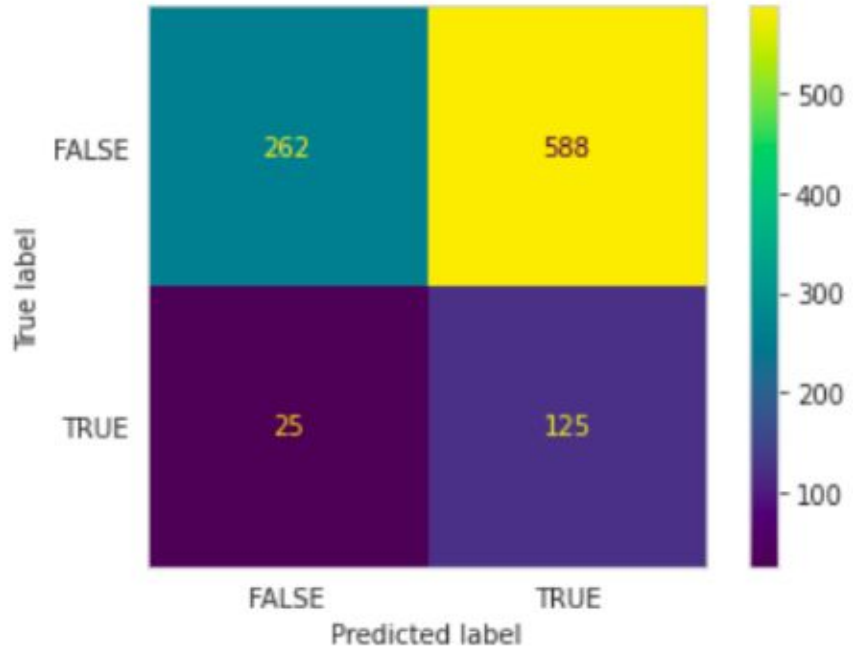
Performance Report

	Precision	Recall	F1-Score	Support
FALSE	0.91	0.31	0.46	850
TRUE	0.18	0.83	0.29	150
ACCURACY			0.39	1000
Macro avg	0.54	0.57	0.38	1000
Weighted avg	0.80	0.39	0.44	1000

Accuracy- 0.39
Precision (True Positive rate)- 0.18
Recall (False positive rate)- 0.83

Confusion Matrix Plot

The confusion matrix plot :



Conclusion

- By far the best predictive result for customer churn characterization and prediction comes from utilizing the balanced dataset to build a **Decision Tree classification model**. The accuracy is 86% which seems fitting as the balanced dataset removed irrelevant attributes like area code, state, phone, account length, etc.
- From the analysis, we were able to conclude that the most important attributes to consider to make a decision were: Inter Plan, Daytime Charges, Voicemail Plan and No of Customer Calls.
- Carrying out iterative cycles of data-preprocessing and feature selection may improve our model's performance.
- Other techniques include k-means clustering may yield better accuracy for more promising results.