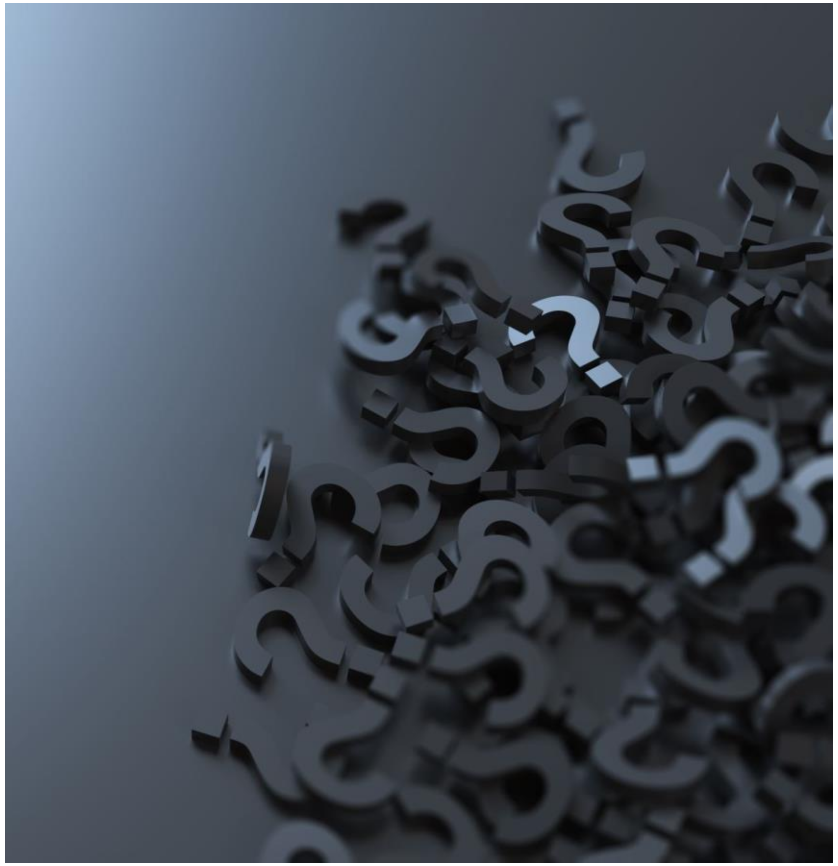
ONLINE SHOPPERS PURCHASING INTENTION CLASSIFICATION

Problem Description

- This is a classification problem with the target variable "Revenue" = True or False.
- The goal is to predict whether a session will produce Revenue with the help of features containing several types of session information.



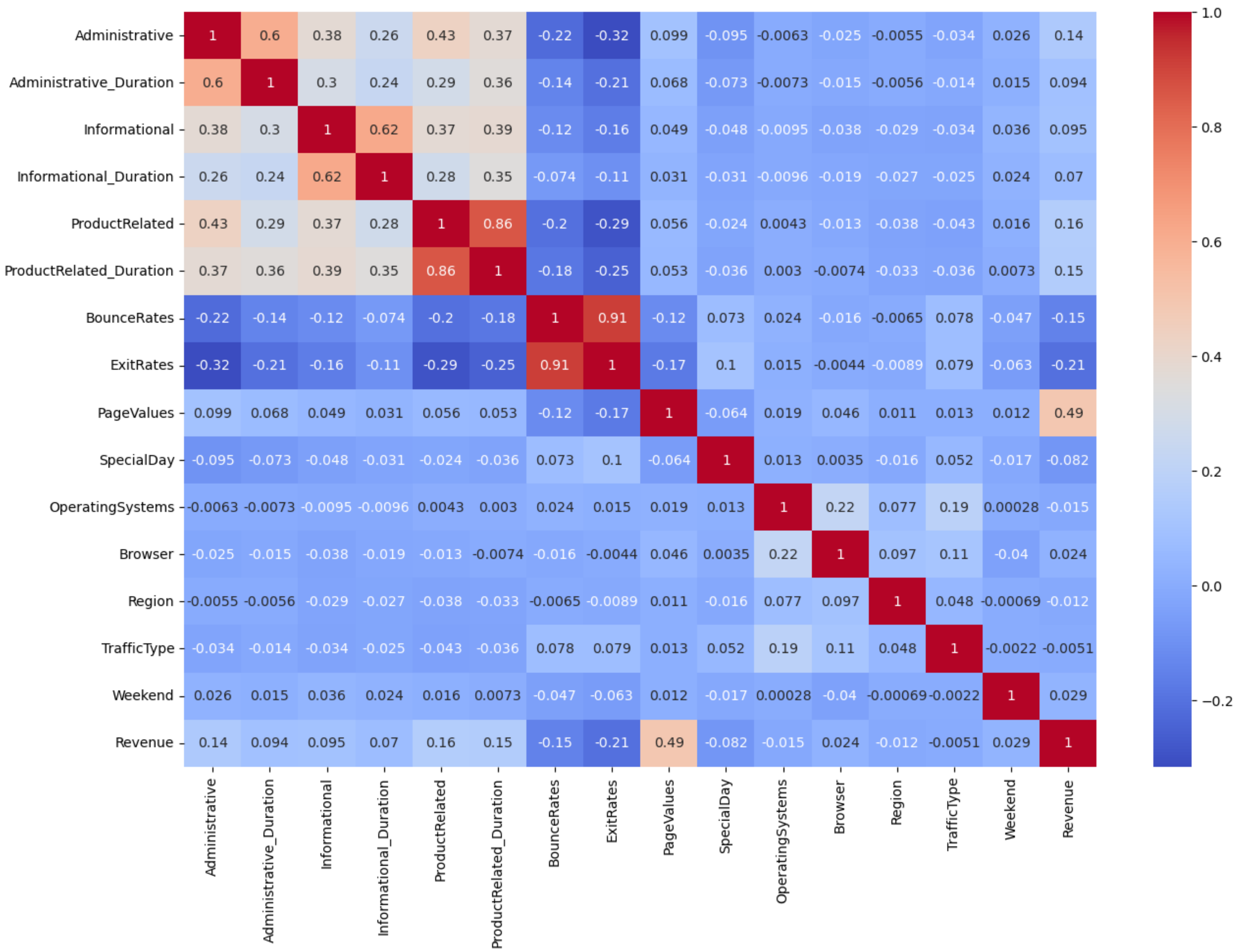
Data Description

- The dataset contains 18 attributes, including 10 numerical and 8 categorical.
- The "Revenue" attribute is the class label.
- The dataset includes information on page visits and time spent in different categories.
- The "Bounce Rate" represents the percentage of visitors who leave without further interaction.
- The "Exit Rate" is the percentage of pageviews that were the last in the session.
- The "Page Value" is the average value of a visited page before an e-commerce transaction.
- The "Special Day" feature indicates the likelihood of a session being finalized with a transaction.
- The dataset provides insights into e-commerce user behavior for marketing strategies and improving user experience.

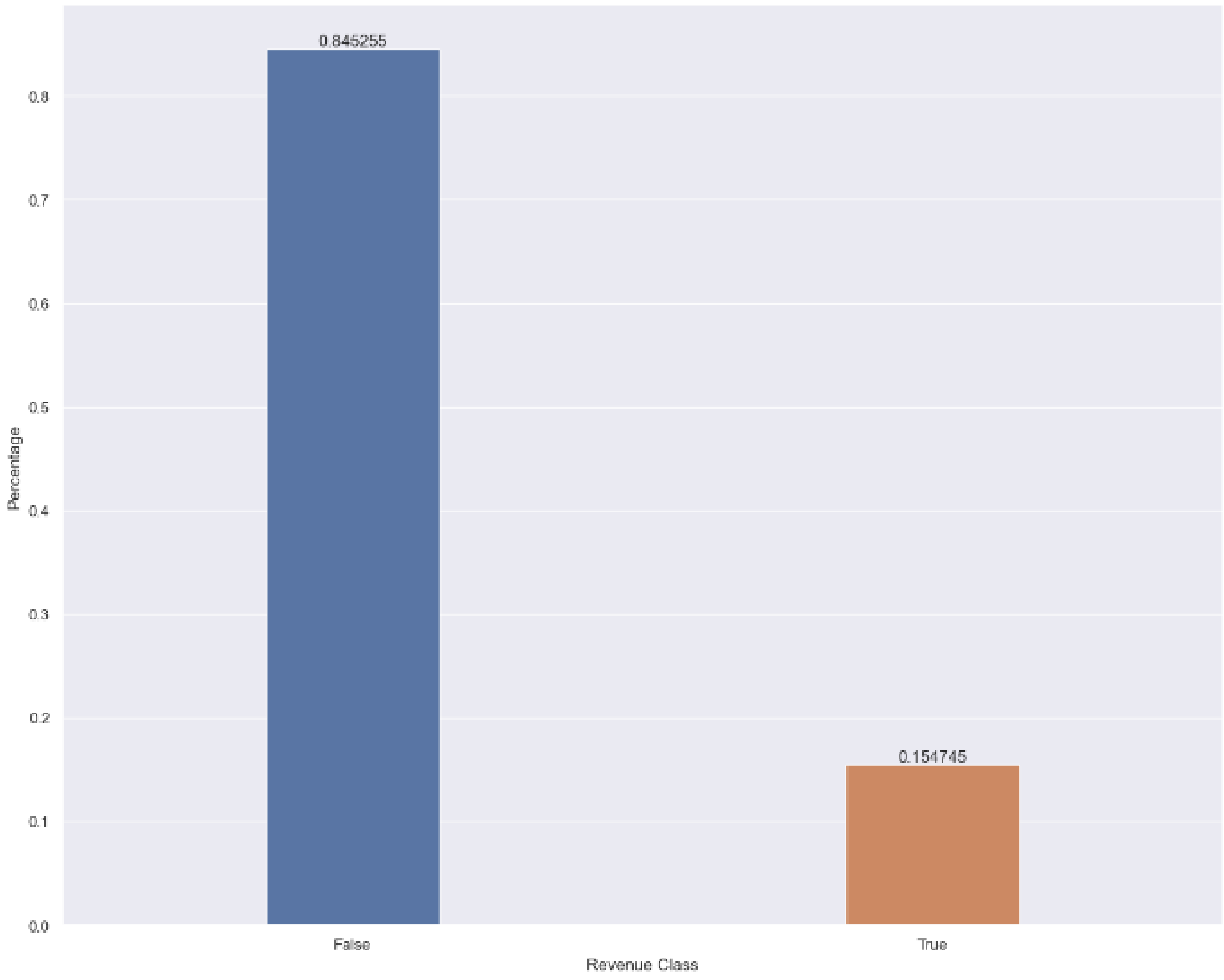
Feature Name	Feature Description
Administrative	This is the number of administrative-type pages visited by the user on the website.
Administrative Duration	Duration of time (in seconds) that the user spent on the administrative pages of the website.
Informational	This is the number of informational pages visited by the user on the website.
Informational Duration	Duration of time (in seconds) that the user spent on the website's informational pages.
Product-Related	This is the number of product-related pages visited by the user on the website.
Product-Related Duration	Duration of time (in seconds) that the user spent on the website's product-related pages.
Bounce Rate	Bounce Rate for a webpage is the percentage of users who entered the website from that page and bounced/left the website without triggering any other requests during that session.
Exit Rate	The value of "Exit Rate" feature for a web page is calculated as for all pageviews to the page, the percentage that were the last in the session.
Page Value	Page Value of a web page represents the average value for the web page that a user visited before completing an e-commerce transaction.
Special Day	The value of Special Day feature indicates the closeness of the site visiting time to a specific special day (e.g., Father’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction.
Operating System	Categorical Feature showing the User’s Operating System.
Browser	Browser information for the user.
Region	Categorical feature containing user’s region.
Traffic Type	This is the type of traffic source for the user (e.g., search engine, social media, etc.).
Visitor Type	Shows if the user is returning or a new visitor
Weekend	Boolean value indicating whether it is the weekend.
Month	Month of the year.

EDA and Data Preprocessing

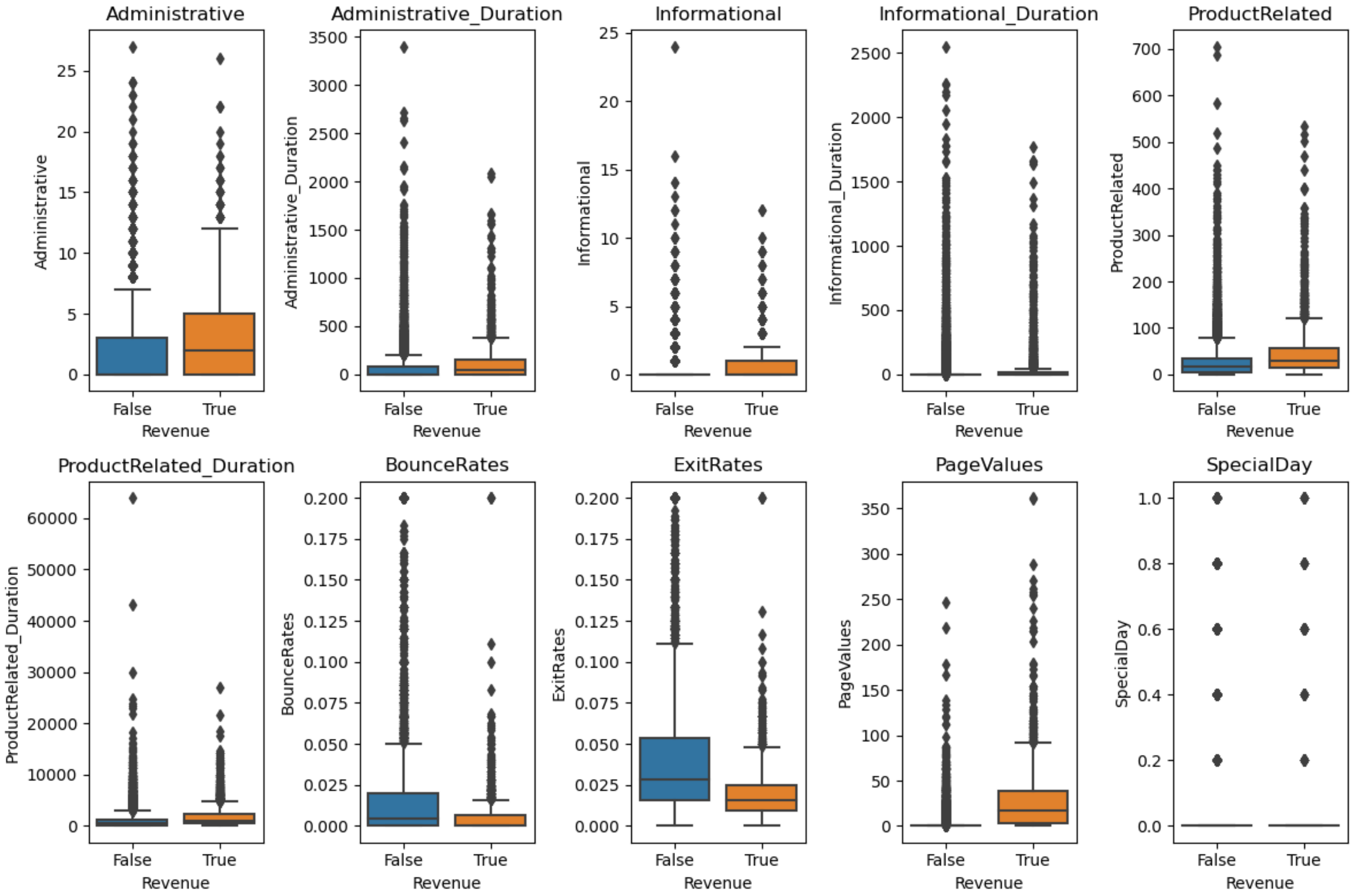
Correlation Analysis:



Class Imbalance:



Boxplots:



Administrative
Administrative_Duration
Informational
Informational_Duration
ProductRelated
ProductRelated_Duration
BounceRates
ExitRates
PageValues
SpecialDay
False
True
Revenue

Exploratory Data Analysis:

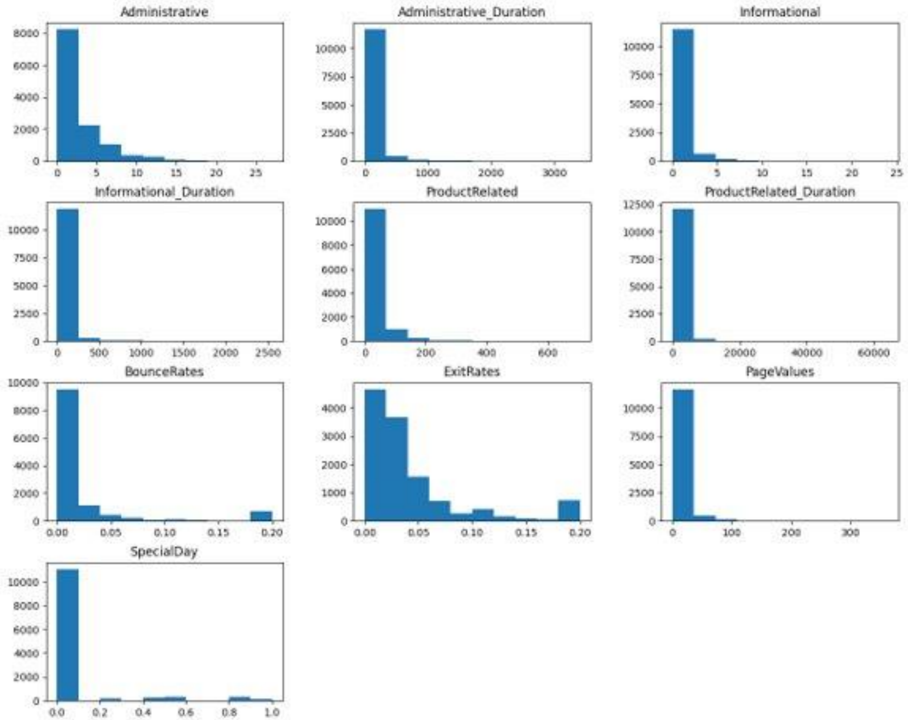
- For numerical variables, we look at the distribution of each column and note the mean, standard deviation and min-max values.

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	ExitRates	PageValues	SpecialDay
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
mean	2.315166	80.818611	0.503569	34.472398	31.731468	1194.746220	0.043073	5.889258	0.061427
std	3.321784	176.779107	1.270156	140.749294	44.475503	1913.669288	0.048597	18.568437	0.198917
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	7.000000	184.137500	0.014286	0.000000	0.000000
50%	1.000000	7.500000	0.000000	0.000000	18.000000	598.936905	0.025156	0.000000	0.000000
75%	4.000000	93.256250	0.000000	0.000000	38.000000	1464.157214	0.050000	0.000000	0.000000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	0.200000	361.763742	1.000000

For categorical variables, we look at the number of categories within each feature.

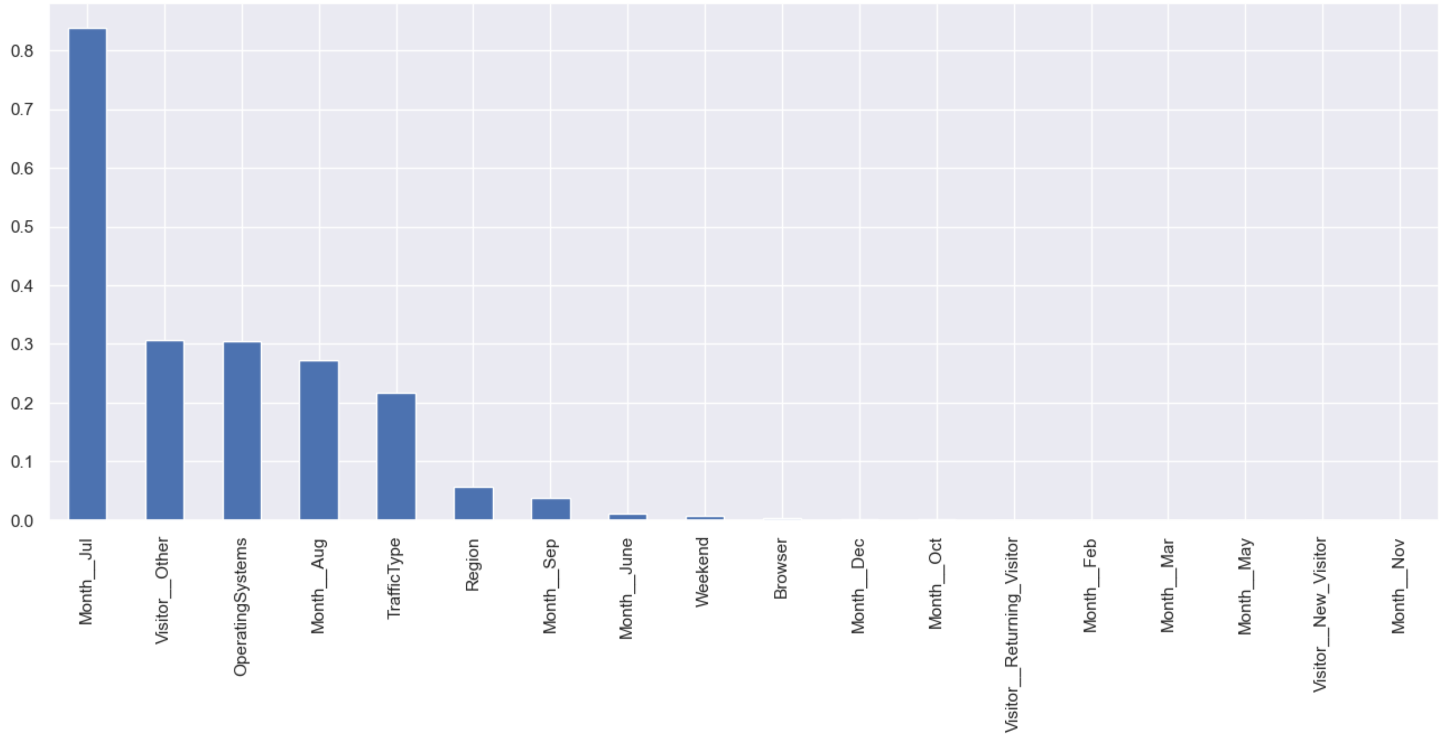
```
Month              10
OperatingSystems    8
Browser            13
Region              9
TrafficType        20
VisitorType         3
Weekend             2
Revenue             2
```

To understand the distribution of the numerical variables, a histogram of all the numerical variables was plotted. It was seen that there is skewness in every variable.



Feature Selection:

- We use the Pearson's R for assessing the relation between numerical features and use the chi squared test to assess the relation between categorical variables.



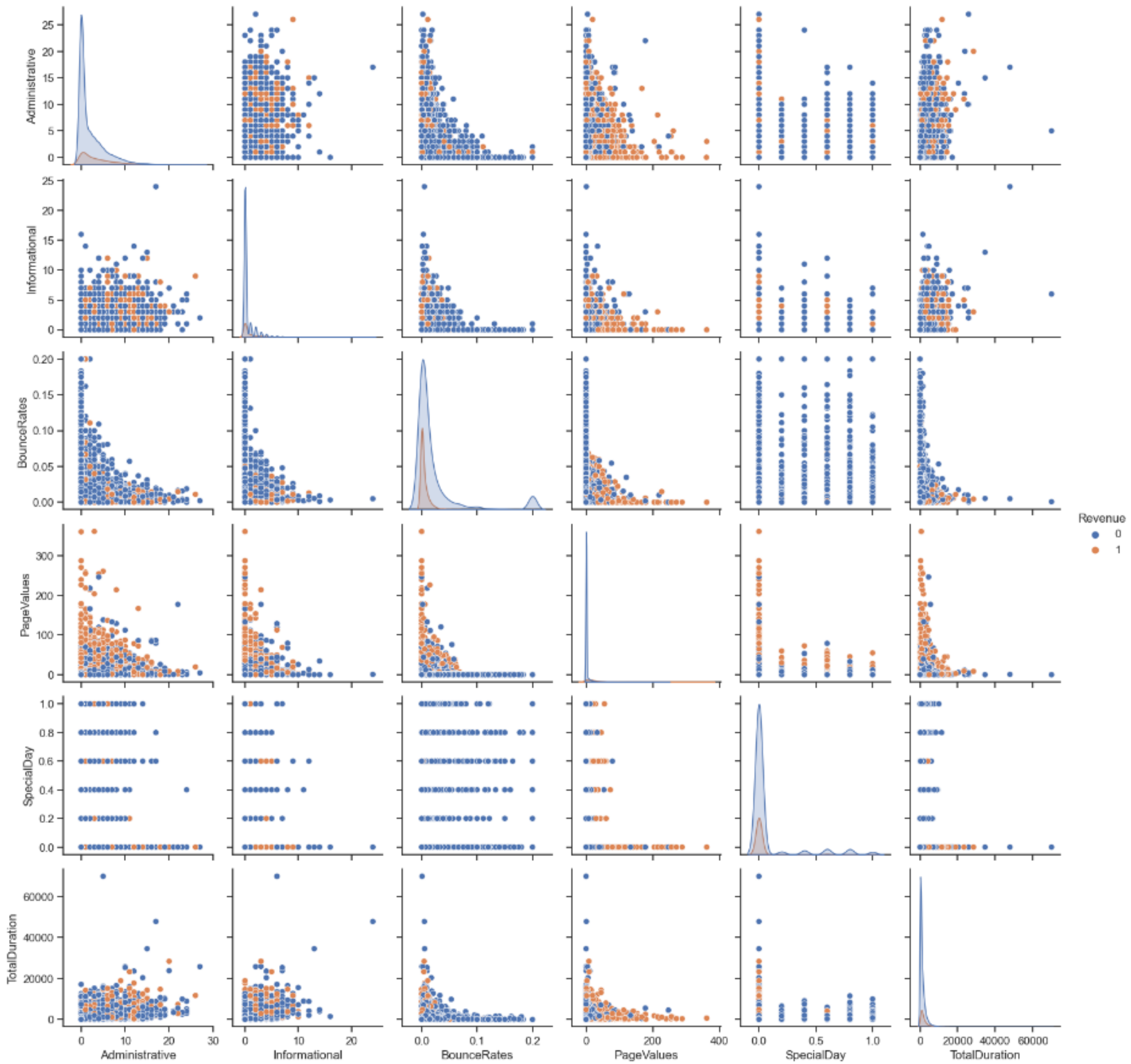
Feature Selection:

- The p-values obtained from the chi-squared test are:
- Looking at the domain knowledge, we can see that the features Administrative Duration, Informational Duration, and Product related Duration combine to give the total time spent by the customer on the website. So, we combine the three features into one, namely TotalDuration.
- With these feature selection techniques, we reduce the number of features for models from 28 to 19. This would enable the model to learn faster, and its complexity also decreases in the process.

Month__Jul	8.389294e-01
Visitor__Other	3.072289e-01
OperatingSystems	3.048965e-01
Month__Aug	2.715249e-01
TrafficType	2.168395e-01
Region	5.720218e-02
Month__Sep	3.781416e-02
Month__June	1.118278e-02
Weekend	7.347547e-03
Browser	3.083995e-03
Month__Dec	7.253287e-04
Month__Oct	6.066376e-04
Visitor__Returning_Visitor	1.478870e-05
Month__Feb	2.251791e-07
Month__Mar	2.858176e-10
Month__May	1.195512e-13
Visitor__New_Visitor	6.348514e-26
Month__Nov	1.155523e-49

Linear Separability

- hvadjcvs



MODEL IMPLEMENTATION

Gaussian Naïve Bayes

- Easiest to implement, uses Bayes theorem to calculate probability of each instance for each class.
- Assumes independence between dependent variables
- Assumes that the continuous variables come from a normal distribution.

Training Data Metrics

	Predicted 0	Predicted 1
0	7294	1
1	1331	5

```
Precision of the model is : 0.8333333333333334
Recall of the model is : 0.003742514970059802
F1_score of the model is : 0.007451564828614009
```

Test Data Metrics

	Predicted 0	Predicted 1
0	3111	16
1	549	23

```
Precision of the model is : 0.5897435897435898
Recall of the model is : 0.04020979020979021
F1_score of the model is : 0.07528641571194762
```

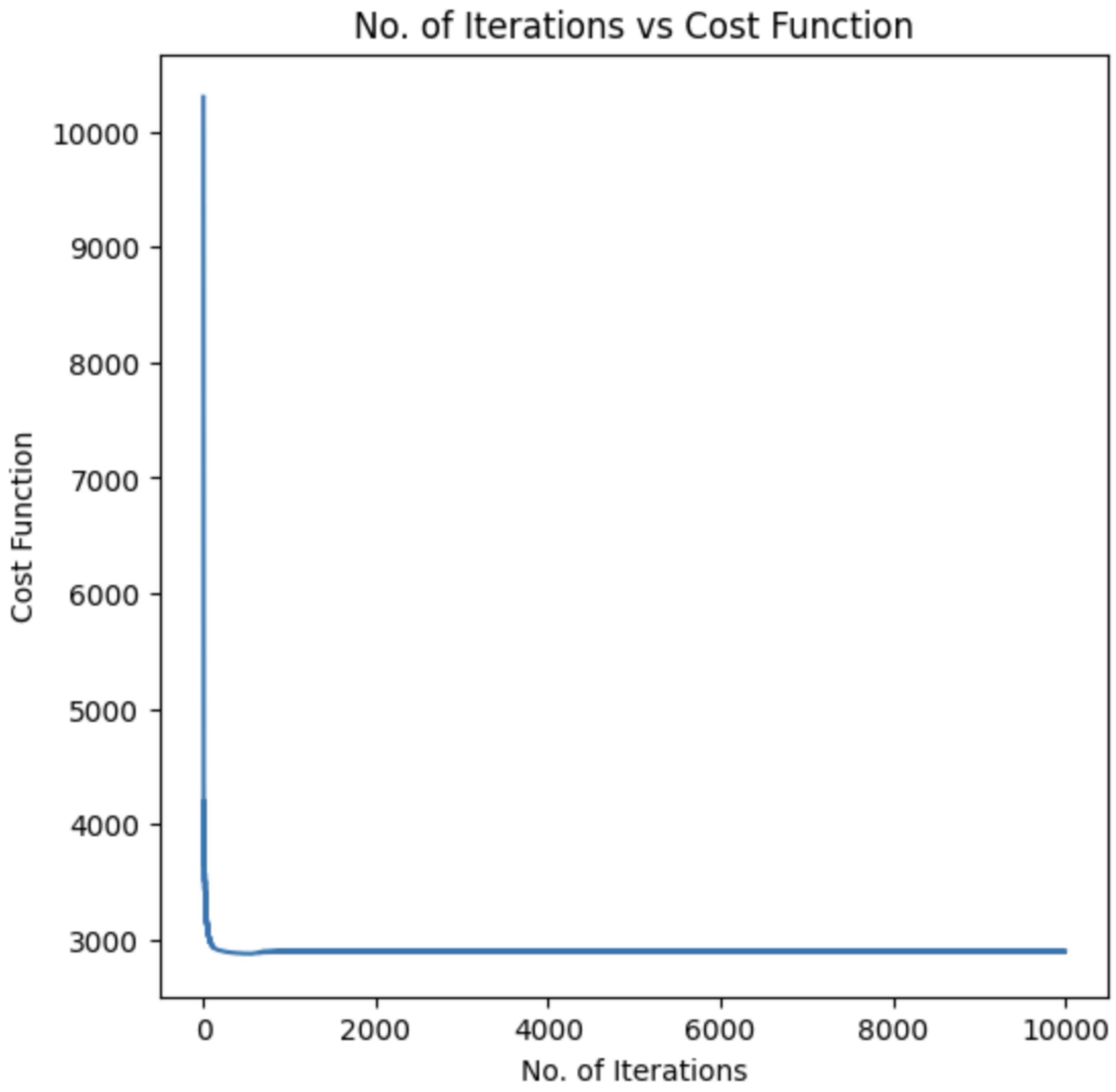
Results:

- Since Gaussian Naïve Bayes calculates probabilities using Bayes Theorem, the likelihoods are multiplied by the priors of each class. Due to class imbalance, the prior of the dominant class (0) drives the probabilities for that class higher, giving us more false negatives and driving the precision higher.
- This model is, however, not accurate as the features are not independent and show dependence.

LOGISTIC REGRESSION

- Most common Classification model
- Used since our data outcome is categorical in nature and the predictors are not linearly coupled

Cost Function Graph



TRAINING DATA METRICS

	Predicted 0	Predicted 1
0	8165	173
1	1005	521

	precision	recall	f1-score	support
0	0.89	0.98	0.93	8338
1	0.75	0.34	0.47	1526
accuracy			0.88	9864
macro avg	0.82	0.66	0.70	9864
weighted avg	0.87	0.88	0.86	9864

TEST DATA METRICS

	Predicted 0	Predicted 1
0	2043	41
1	240	142

	precision	recall	f1-score	support
0	0.89	0.98	0.94	2084
1	0.78	0.37	0.50	382
accuracy			0.89	2466
macro avg	0.84	0.68	0.72	2466
weighted avg	0.88	0.89	0.87	2466

SOFT MARGIN SVM

- Based on maximum margin classification
- Uses the most computational power out of the classical ML algorithms
- Uses the Sequential Least Squares programming to obtain the decision boundary

Model Training

- Trained on 500 samples.
- Run on Intel (R) Core (TM) i7-5500U CPU @ 2.40GHz (4 CPUs)

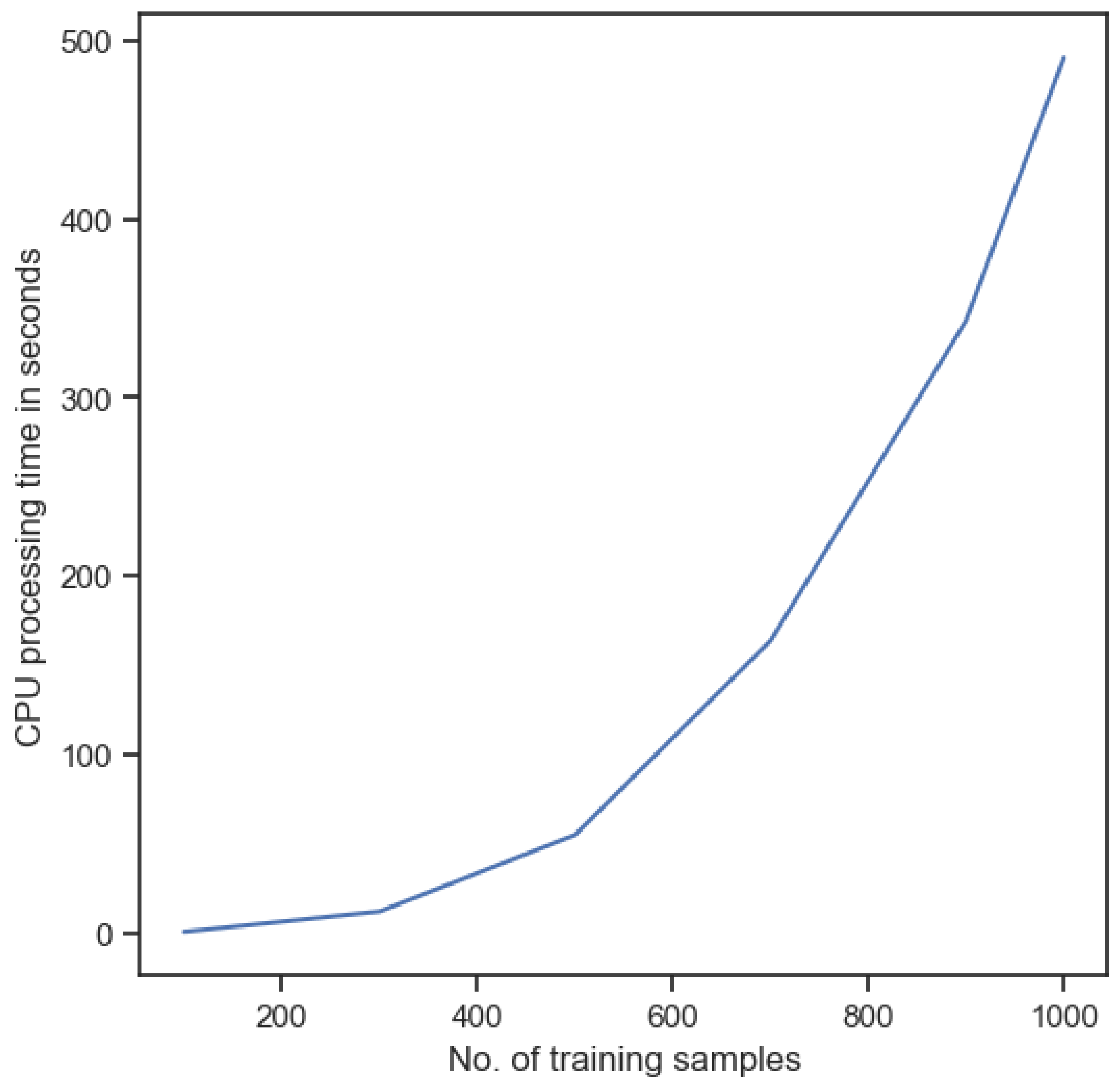
- The f1 score of the model was obtained as 64.18 %
- The precision of the model was obtained as 72.88 %
- The recall of the model was obtained as 57.3%

Time for training a sample of 500:

CPU processing Time:	54.140625 seconds
Total training time:	59.5065 seconds

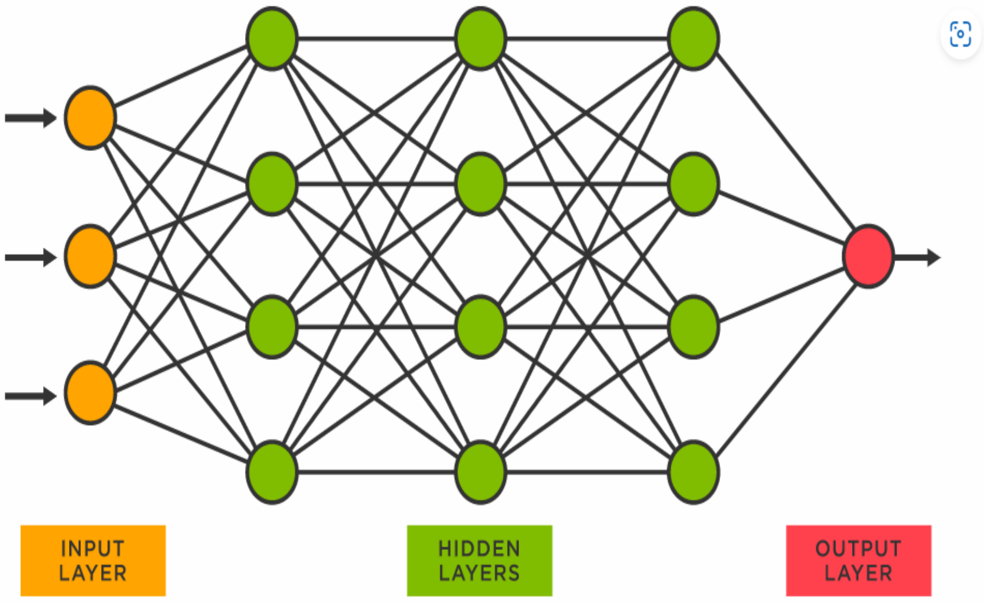
Time Complexity of SVM:

$O(n^2)$



NEURAL NETWORKS

- Achieve state-of-art performance compared to other models
- Is complex and has low interpretability
- Is a non-linear model
- Uses backpropagation algorithm to improve model performance



A fully connected Neural network Network

A Simple Neural network model

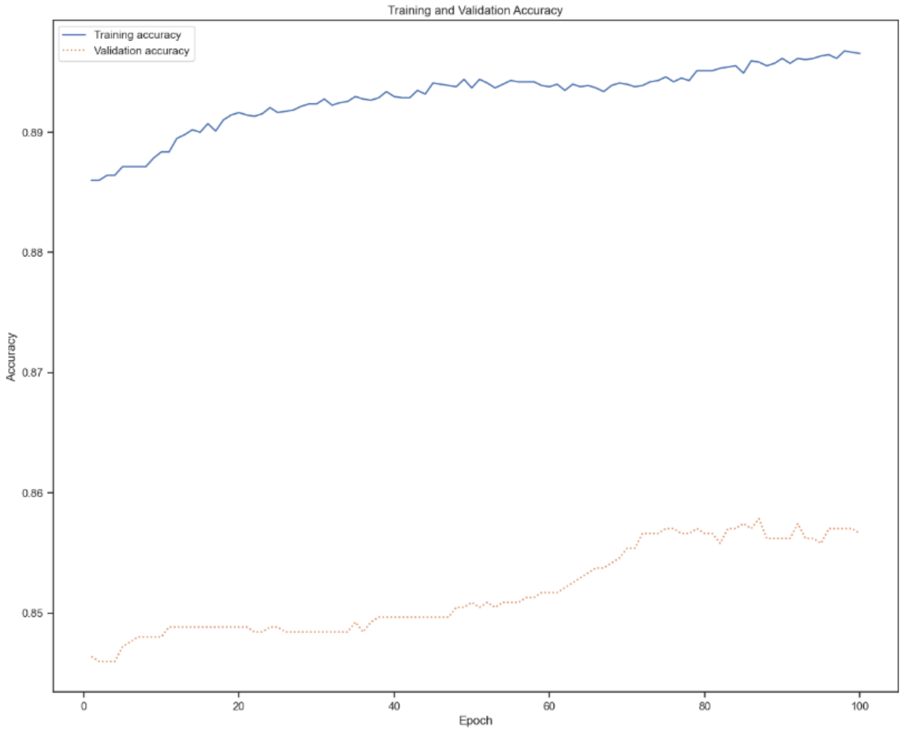
```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 Hidden_layer_1 (Dense)      (None, 8)                 152

 Output_layer (Dense)        (None, 1)                 9

=================================================================
Total params: 161
Trainable params: 161
Non-trainable params: 0
_________________________________________________________________
```

Evaluation on the test set

Cut-off Analysis

	cutoff	precision	accuracy
0	0.10	0.903014	0.725625
1	0.20	0.711009	0.862249
2	0.50	0.329620	0.884678
3	0.80	0.024902	0.868292
4	0.95	0.002621	0.851188

Threshold selected: 0.2



Training Set: 89.8%, Precison: 71.83%, Recall: 57.14%

Test Set: 85.5 %, precision: 71.88%, Recall:

FINAL RESULTS

	Precision	Recall	F1 Score
Naive Bayes	0.59	0.04	0.075
Logistic	0.78	0.37	0.5
Neural Networks	0.71	0.57	0.63

Conclusion

Neural Networks outperform classical machine learning algorithms but they risk overfitting while training.

Out of the Naïve Bayes, Logistic regression, and SVM, SVM gives the best performance but is computational inefficient.

Questions?