Project 1: Classification

Bhumit Shah(1001765834)
Kaustubh Rajpathak(1001770219)

Overall Status

As a start, we first looked at the data and tried to understand the data i.e. attributes. Then we dropped the attributes with the unknown values and required by project description. After analyzing the data i.e. histogram analysis we decided on 4 attributes which can be dropped before further processing. After removing attributes and sampling, we calculated information gain and gini index for remaining attributes. Based on that we developed decision trees. For naive bayes, conditional probabilities were calculated and developed model for naive bayes. After training the models we tested it using test data and generated a confusion matrix to calculate F1 Score, Precision and Recall. We trained and tested on 2 splits which are 80/20 and 50/50 respectively. We were able to generate decision trees and naive bayes. All models use 10-fold cross validation for accurate results.

File Descriptions

There are 2 files in this project. Project_It_1.R file has implementations of Decision Tree using GINI Index and Information Gain and also Naive Bayes.

Division of Labour

Name	Work Done	No. Of Hrs. Spent
Bhumit Shah	Implementation of Naive Bayes, Visualization of Results	10
Kaustubh Rajpathak	Implementation of Decision Tree(Information Gain and GINI Index)	10

Problems Encountered

1) Identifying attributes for removal
   During the identification for attributes which may not be relevant, we were not sure which methods to use for the same. We were confused on what we should use as there were many options for that such as information gain, frequency distribution etc.

2) Understanding of attributes related to domain
   There were some attributes which we did not know as we were not aware about the domain and how they are related to the domain.

3) What do you mean by withholding the column. What should be the criteria for withholding the column?Is it from the attributes identified in the milestone or something else. Is column identification different for different splits? Project description does not seem to be clear about these questions.

Analysis

Comparison of Different Metrics F1 Score, Confusion Matrix , Precision, Recall for Decision Tree and Naive Bayes

Attribute Drop Analysis

1. Information Gain

The information Gain metric for the decision tree uses entropy of each node to decide which split to take. It chooses the split with the highest decrease in entropy which corresponds to the most information gain. Using this metric we identified Duration having the highest variable importance, campaign having the lowest variable importance and cons.price.idx in the middle. We tested the model eliminating each attribute one at a time and observing the change in metrics.

This is what we got by eliminating duration :

Information	Accuracy	Precision	Recall	F1 Score

Gain Tree				
Metrics before Withholding Duration(80/20 Split)	0.8935	0.592	0.476	0.5277
Metrics After Withholding Duration(80/20 Split)	0.8835	0.6197	0.176	0.2741
Metrics before Withholding Duration(50/50 Split)	0.8982	0.6956	0.2894	0.4088
Metrics After Withholding Duration(50/50 Split)	0.8866	0.7063	0.1785	0.2849

As we can see the model performs poorly without duration for yes values as the training dataset is highly imbalanced with a large number if no values and without the dominating factor the F1 score drastically reduces.

This is what we got by eliminating cons.price.idx :

Information Gain Tree	Accuracy	Precision	Recall	F1 Score
Metrics before Withholding cons.price.idx(80/20 Split)	0.8935	0.592	0.476	0.5277
Metrics After Withholding cons.price.idx(80/20 Split)	0.892	0.6102	0.4322	0.5060
Metrics before Withholding cons.price.idx(50/50 Split)	0.8982	0.6956	0.2894	0.4088

Metrics After Withholding cons.price.idx(50/50 Split)	0.8952	0.6300	0.4170	0.5019

With the dominating attribute present eliminating a fairly important attribute doesn't make a lot of a difference and the F1 scores are fairly similar

This is what we got by eliminating campaign :

Information Gain Tree	Accuracy	Precision	Recall	F1 Score
Metrics before Withholding Campaign(80/20 Split)	0.8935	0.592	0.476	0.5277
Metrics After Withholding Campaign(80/20 Split	0.8935	0.592	0.476	0.5277
Metrics before Withholding Campaign(50/50 Split)	0.8982	0.6956	0.2894	0.4088
Metrics After Withholding Campaign(50/50 Split	0.8982	0.6956	0.2894	0.4088

As we can see the campaign makes absolutely no difference in the model hence we might as well drop it completely.

2. **Gini Index**

Gini Index is a similar concept to information gain but instead of using the difference in entropy it uses the gini value of a node which in principle uses the same method of associating a number between 0-1 for each attribute wrt the amount of variation in the class. A value of 0.5 means that the attribute has equal yes and no instances which correspond to their respective values. We use the same method of testing as IG.

The effects are the same for all 3 attributes as IG attribute ranking principle is similar to GINI.

3. **Naive Bayes**

The naive bayes classifier calculates conditional probability of each attribute with the target attribute y and then calculates the joint probability of each entry in test data with the target as yes or no. The classifier compares the two and whichever is the highest it assigns it to that class.

This is what we get by dropping duration :

Naive Bayes	Accuracy	Precision	Recall	F1 Score
Metrics before Withholding Duration(80/20 Split)	0.8435	0.4107	0.58	0.4809
Metrics After Withholding Duration(80/20 Split)	0.834	0.3808	0.524	0.441
Metrics before Withholding Duration(50/50 Split)	0.8444	0.3997	0.5575	0.4656
Metrics After Withholding Duration(50/50 Split)	0.83	0.3599	0.5115	0.4725

duration

Y	[,1]	[,2]
no	221.8571	220.5994
yes	504.1259	366.1912

As we can see, dropping the duration attribute doesn't affect the model much as much even though the model has 2 very different mean values of duration for yes and no. These values are obtained by printing the model in R Studio. It shows that the NB model isn't dependent on the duration attribute to make an accurate classification.

This is what we get by dropping campaign :

Naive Bayes	Accuracy	Precision	Recall	F1 Score
Metrics before Withholding Campaign(80/20 split)	0.8435	0.4107	0.58	0.4809
Metrics After Withholding Campaign(80/20 split)	0.8545	0.4365	0.564	0.4921
Metrics before Withholding Campaign(50/50 split)	0.8444	0.3997	0.5575	0.4656
Metrics After Withholding Campaign(50/50 split)	0.8542	0.4204	0.5263	0.4674

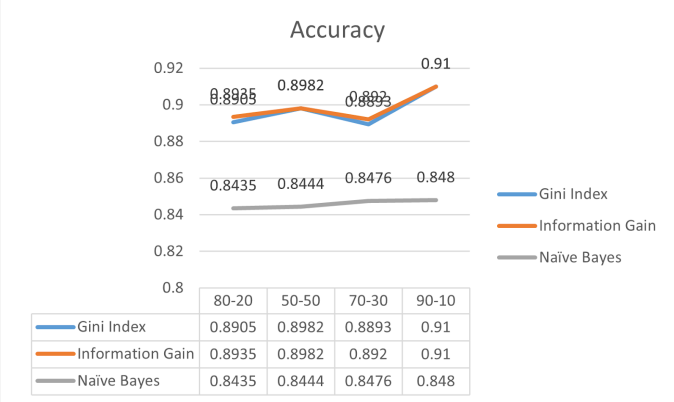
We see quite a surprise here as the model performs better without the campaign attribute suggesting its values do not necessarily correlate with the output. We see this in the decision tree model as well where it has very low variable importance suggesting it has an equal distribution of values corresponding to yes / no. But we know that the data only has 13% yes values meaning campaign introduces false data in the model decreasing its performance.

This is what we get by dropping cons.price.idx:

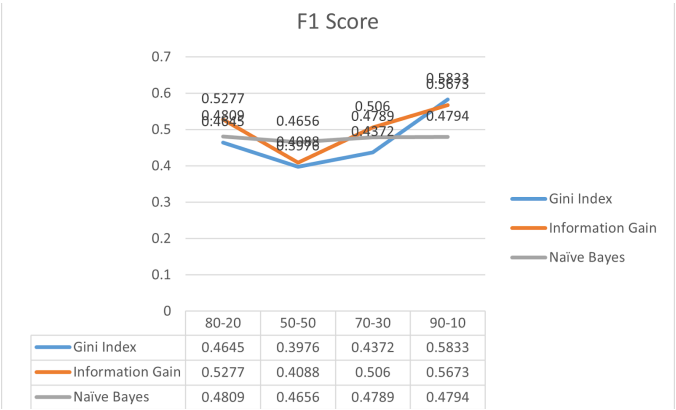
Naive Bayes	Accuracy	Precision	Recall	F1 Score
Metrics before Withholding cons.price.idx (80/20 split)	0.8435	0.4107	0.58	0.4809
Metrics After Withholding cons.price.idx(80/20 split)	0.8495	0.4205	0.54	0.4728
Metrics before Withholding cons.price.idx(50/50 split)	0.8444	0.3997	0.5575	0.4656
Metrics After Withholding cons.price.idx(50/50 Split)	0.852	0.4142	0.5246	0.4629

We see similar performance with and without the attribute which goes by our analysis that the model is not dependent on individual attributes but the collection of them. Attributes mentioned in milestone, which were selected as candidates for removal based on initial analysis are kept as importance for those attributes namely day_of_week,month,conf.price.udx and age turns out be higher than expected for both decision trees and naive bayes respectively.. Removing those attributes results in reduction in F1 Score, Accuracy, Recall and Precision.

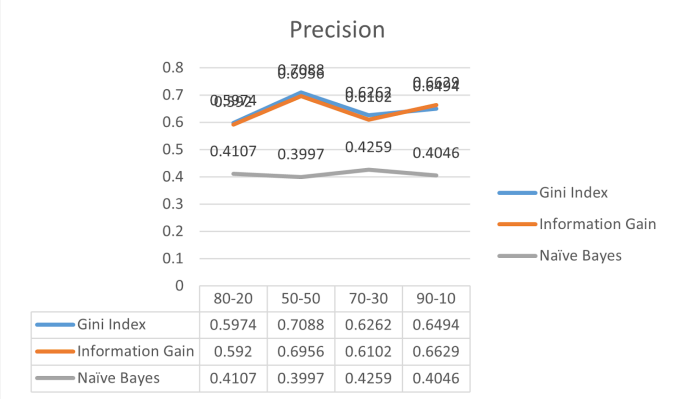
Split Analysis on Model



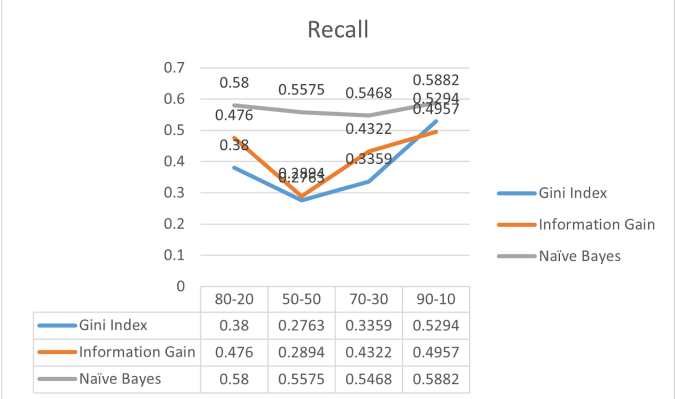
	80-20	50-50	70-30	90-10
Gini Index	0.8905	0.8982	0.8893	0.91
Information Gain	0.8935	0.8982	0.892	0.91
Naïve Bayes	0.8435	0.8444	0.8476	0.848



	80-20	50-50	70-30	90-10
Gini Index	0.4645	0.3976	0.4372	0.5833
Information Gain	0.5277	0.4088	0.506	0.5673
Naïve Bayes	0.4809	0.4656	0.4789	0.4794



	80-20	50-50	70-30	90-10
Gini Index	0.5974	0.7088	0.6262	0.6494
Information Gain	0.592	0.6956	0.6102	0.6629
Naïve Bayes	0.4107	0.3997	0.4259	0.4046



	80-20	50-50	70-30	90-10
Gini Index	0.38	0.2763	0.3359	0.5294
Information Gain	0.476	0.2894	0.4322	0.4957
Naïve Bayes	0.58	0.5575	0.5468	0.5882

1. Accuracy

We can see that the accuracy stays the same for all the splits except 90/10 where it increases in decision trees whereas it stays relatively the same in naive bayes. From this we can conclude that naive bayes require lesser training data than decision trees to reach optimal accuracy. Our limiting factors in accuracy are classification of yes values due to the imbalance in the dataset. Overfitting for decision trees was checked by replacing test data with training data. The accuracy for this was 92 %, therefore we can be sure that the statistics are independent of the dataset and only depend on the quality of the model. With larger training data in the 90/10 split the accuracy is closer to the max possible accuracy of 92 % as expected.

2. Precision

Again the naive bayes model performs fairly consistently and we can see an interesting story with the decision tree models. The precision is the highest at a 50/50 split and drops significantly to the lowest at the 80/20 split after which it improves near to the max found at 50/50. We attribute this behaviour to the quality of the dataset as the other metrics recall, F1 and accuracy do not go by this trend.

3. **Recall**

It is again a similar story with naive bayes being fairly consistent across splits but we see an increasing trend in decision trees with increase in training data size. This can be approved to the model being better trained. It is proved by seeing recall values of the model when run on training data which are significantly higher at the 50/50 split at 0.564 for gini and 0.682 for IG. The model is therefore not sufficiently trained to handle new instances of test data. Recall values therefore improve as the model gets better trained.

4. **F1 Score**

The decision trees again show a trend of improving with higher training data as with recall and naives bayes continues to remain consistent.

Confusion Matrices

1) For Information Gain Decision Tree and GINI Index Decision Tree

First Conf. Matrix in each image below is for Information Gain Decision Tree and Second is for GINI Index Decision Tree

80/20 Split

```
          infomodelpredict
           Predicted:no Predicted:yes
Actual:no          1668             82
Actual:yes          131            119
print(conf_matrix_gini)
          ginimodelpredict
           Predicted:no Predicted:yes
Actual:no          1686             64
Actual:yes          155             95
```

50/50 Split

```
          infomodelpredict1
           Predicted:no Predicted:yes
Actual:no          4315             77
Actual:yes          432            176
print(conf_matrix_gini1)
          ginimodelpredict1
           Predicted:no Predicted:yes
Actual:no          4323             69
Actual:yes          440            168
```

2) For Naive Bayes

First Conf. Matrix is for 80/20 Split and Second is for 50/50 Split

```
          banknbpredict
           Predicted:no Predicted:yes
Actual:no          1551            199
Actual:yes          109            141
print(conf_matrix_nb1)
          banknbpredict1
           Predicted:no Predicted:yes
Actual:no          3905            487
Actual:yes          275            333
```