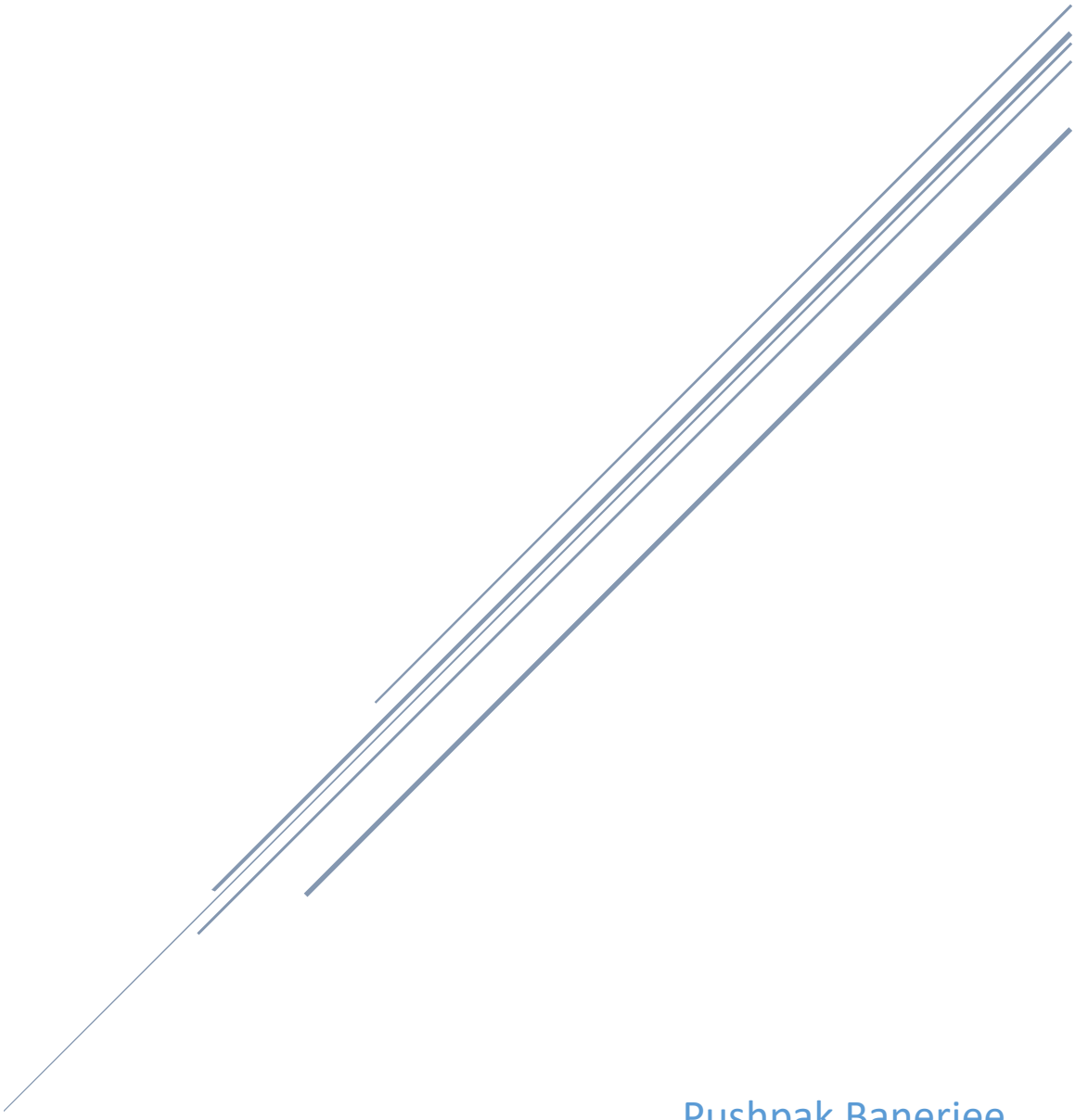
SOCIAL MEDIA TOURISM PROJECT

Project Notes - 2



Pushpak Banerjee
Post Graduate Program – DSBA – September A

Contents

1. Model building and interpretation.

a. Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes)

We initially convert all categorical values into numeric values using the principles of label encoding. Next we identify the dependent variable as the following: **Taken_product.** The remaining variables are independent variables (X). Hence, for predictive model building purposes, we split the dataset into X variables in one dataframe, and y / dependent variable in another series. Our target is to predict the dependent variable.

We then split the total data into train and test set, 70% data is noted as train – to be used for training the dataset. And the remaining 30% will be used as training dataset – to be used to check the performance of the model.

As we discussed in the last week's session, the entire dataset is broken into 2 datasets – one for the laptops, and the other for the mobiles / tabs / iOS etc, which are collectively identified as mobile dataset

The following models are built for the **Laptop dataset**.

- Logistic Regression
- Random Forest (using pruning technique)
- Decision Tree (using grid search)
- Linear Discriminant Analysis

The accuracy score of the models are as follows:

	Test Accuracy
Logistic Regression	0.813
Decision Tree - Grid Search	0.789
Random Forest - Pruning	0.903
Linear Discriminant Analysis	0.807

For the mobile dataset, the following models are built:

- Logistic Regression
- Decision Tree (using grid search)
- Linear Discriminant Analysis

The accuracy score of the models are as follows:

	Test Accuracy
Logistic Regression	0.864
Decision Tree - Pruning	0.922
Linear Discriminant Analysis	0.864

Purely from the accuracy point, we see that the Random Forest is the best performer for the **Laptop** dataset, and the Decision Tree seems to be the best performer for the **Mobile** dataset.

However, we will check in the next section whether that really makes a good model for our purpose or not.

b. Test your predictive model against the test set using various appropriate performance metrics

For each of the models created, we test the following scores:

- Accuracy,
- Recall score (from classification report)
- Confusion matrix
- AUC score.

This is done for testing and training dataset. The output for test data is captured in the tables below:

	Model (Laptop data)		
	Test Accuracy	Recall score for 1	AUC
Logistic Regression	0.813	0.46	0.857
Decision Tree - Grid Search	0.789	0.55	0.819
Random Forest - Pruning	0.903	0.77	0.983
Linear Discriminant Analysis	0.807	0.44	0.851

	Model (Mobile data)		
	Test Accuracy	Recall score for 1	AUC
Logistic Regression	0.864	0.20	0.784
Decision Tree - Pruning	0.942	0.74	0.982
Linear Discriminant Analysis	0.864	0.21	0.784

Based on the above tables, it can be said that the Random Forest is the most suited model for the **laptop** datatset, since it has both high test accuracy as well as recall score for the 1 (Taken_product = True) data.

For the **mobile** dataset, the Decision Tree model is the best, as it has high accuracy and recall score for test data.

c. Interpretation of the model(s)

The accuracy score of the models show how many True Positives and True Negatives the model is able to correctly predict. High accuracy means a good and strong model.

However, if the accuracy levels reach close to 100%, this indicates that the models are becoming overfit. In other words, the model is able to predict the train dataset extremely well, but when it will be put to real life scenarios, the models might fail badly.

Models like Decision Trees and Random Forests are extremely prone to overfitting. Hence, we had to prune these models to make them more appropriate.

Another very important aspect is the Recall score of the model. A recall is the model's capacity to identify positives as positives. For example, when we say in the previous table that Decision Tree has a recall value of 0.74 for 1, this means that 74% of the time, the model is able to predict correctly a person who has taken the product. For the airline industry advertising in the social media, this is a very important criteria, since advertising in social media is expensive, and the company would like to target the advertisements to the correct candidates, who have a high probability of actually buying the ticket.

2. Model Tuning

a. Ensemble modelling, wherever applicable

We use Bagging as an ensemble technique to improve the performance of the models built previously.

In Bagging, smaller models are built on subsets of data, and then they are aggregated together to make the final model. The advantage of this is that it can prevent overfit, and improve accuracy of the data.

The output of each models using bagging is as follows:

	Bagging (Laptop)		
	Test Accuracy	Recall score for 1	AUC
Logistic Regression	0.825	0.51	0.854
Decision Tree - Grid Search	0.789	0.55	0.789
Random Forest - Pruning	0.888	0.61	0.974
Linear Discriminant Analysis	0.819	0.48	0.851

	Bagging (Mobile)		
	Test Accuracy	Recall score for 1	AUC
Logistic Regression	0.863	0.2	0.784
Decision Tree - Pruning	0.947	0.66	0.973
Linear Discriminant Analysis	0.865	0.23	0.784

b. Any other model tuning measures (if applicable)

Boosting is another method used to improve model performance. The output of each model using boosting is as follows:

	Boosting (Laptop)		
	Test Accuracy	Recall score for 1	AUC
Logistic Regression	0.795	0.33	0.853
Decision Tree - Grid Search	0.852	0.74	0.92
Random Forest - Pruning	0.945	0.81	0.998
Linear Discriminant Analysis	x	x	x

	Boosting (Mobile)		
	Test Accuracy	Recall score for 1	AUC
Logistic Regression	0.863	0.16	0.785

Decision Tree - Pruning	0.976	0.88	0.977
Linear Discriminant Analysis	x	x	x

c. Interpretation of the most optimum model and its implication on the business

From the above tables, it becomes clear that 2 models are the most optimal, one each for the Laptop dataset and Mobile dataset:

- **Laptop**: ***Random Forest with boosting*** - which has an overall accuracy of 94%, and a Recall value of 81%. This means that the model is able to predict around 94% of the overall values correctly (this includes those who bought the tickets and those who did not). Additionally, this model gives a recall value of 81% for positively identifying people who bought tickets.

- **Mobile**: ***Decision Tree with boosting –*** this model has an overall accuracy of 97% and a Recall value of 88% for test data. However, at such high accuracy values, the model can have a high probability of overfitting. Hence, the next best option would be a ***Normal Decision Tree model***, which has accuracy of 94% and Recall score of 74%

Overall, we can say that if the company follows the models as prescribed above, there is a very high probability that they will be able to identify correctly all persons who will buy the tickets, and all those who will not buy. Among those who are really buying tickets, the models will be able to identify them also very accurately. Social media advertising being very expensive, the company would do good in trying to follow the models created for them.