**5. Prediction Model**

In this part, the prediction models are trained with a set of flight data and then another set of data is used to validate the model. Since there are two kinds of delay, departure delay and arrival delay, we applied machine learning techniques for both kinds of delay. Moreover, delay could be a binary class (delay or not) or a multi-level class (delay class). Thus, prediction models are applied to predict both binary class and multi-level class for delay.

**5.1 Predict delay or not with binary class (True or False)**

The ultimate goal of this study is to accurately predict if a flight will be delayed or not, and in some situations, we cannot obtain complete and accurate weather attributes such as wind gust speed, dew point temperature, and atmosphere pressure. While missing weather data, it may be hard to predict delay. However, we use a portion of our data with missing weather data for machine learning to predict delay and research the results. This section contains four sub-sections: departure delay prediction with weather data, arrival delay with weather data, departure delay prediction without weather data, arrival delay prediction without weather data.

**-Departure delay prediction using binary class (including weather data)**

In Fig. 5.1 below, the 10-fold cross-validation results of prediction models during training is shown. Decision Tree and Random Forest preform the best, while Gradient Boosting method becomes the third best. KNN, Gaussian Naïve Bayes, and SVM classifiers do not perform well. This results is anticipated because many attributes of our data are categorical like airplane manufacturer, airline, origin, and destination. Thus, SVM, Gaussian Naïve Bayes, and KNN may not perform well.

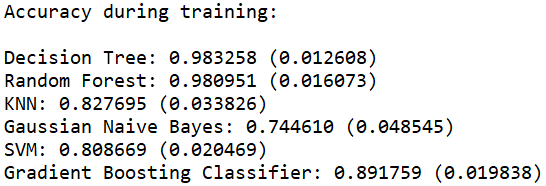


Figure 5.1 Cross-validation results of prediction models (departure delay prediction with weather data).

In Table 5.1 below, the testing results of six classifiers are shown. From the confusion, matrix, Random Forest classifier accurately predict all on time and delayed flights, and Decision Tree classifier also performs well with high accuracy. Gradient Boosting Classifier is the third best, which is the same as the results during training. The other three classifiers are unreliable though they still have accuracy score up to about 0.8.

Table 5.1 Testing results of prediction models (departure delay prediction with weather data).

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |
|  |  |
| KNN | Gaussian Naïve Bayes |
|  |  |
| SVM | Gradient Boosting Classifier |

The ROC curves of all six models are shown in Fig. 5.2 below. By using the area under the curve and combining the results above, it is not hard to conclude that when we have all the attributes including weather data, **Random Forest is the best classifier to predict departure delay.**

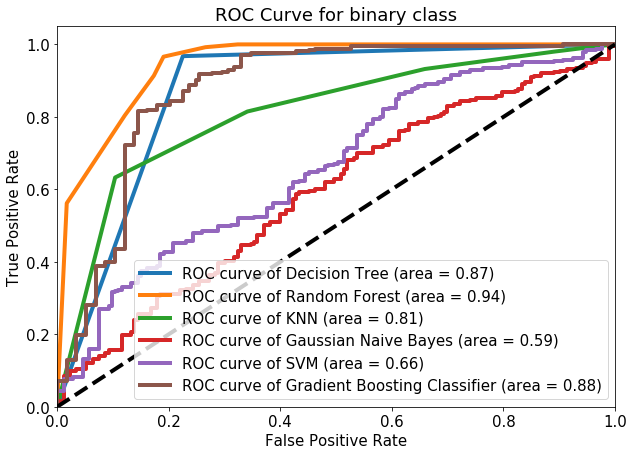


Figure 5.2 ROC curves of prediction models (departure delay prediction with weather data).

**-Arrival delay prediction using binary class (including weather data)**

In Fig. 5.3 below, the 10-fold cross-validation results of prediction models during training is shown. Decision Tree, Random Forest, and Gradient Boosting classifiers preform the best. KNN and SVM yield acceptable results during cross-validation. Gaussian Naïve Bayes classifier performs the worst because none of our attributes follow Gaussian distribution. (Other Naïve Bayes classifiers like complement Naiive Bayes are tested but results are also unacceptable).

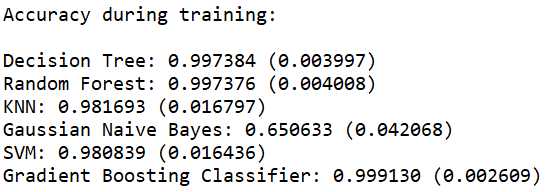


Figure 5.3 Cross-validation results of prediction models (arrival delay prediction with weather data).

From testing results shown in Table 5.2 below, the confusion matrices and accuracy reports are useful to rank the classifiers. It is worth to notice that the accuracy score is not useful here because we need to keep cost in mind because most of the flights in test dataset are not delayed here. Thus, though some models like SVM and KNN predict most on time flights correctly, they do not predict 8 delayed flights correctly. Random Forest and Gradient Boosting classifiers predict all flights accurately, while Decision Tree classifier perform relatively worse by mis-classifying 3 delayed flights as on time.

Table 5.2 Testing results of prediction models (arrival delay prediction with weather data).

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |
|  |  |
| KNN | Gaussian Naïve Bayes |
|  |  |
| SVM | Gradient Boosting Classifier |

The ROC curves of prediction models is shown in Fig. 5.4 below. If the accuracy of these models are ranked only by area, Gradient Boosting is the best while SVM becomes the second place. However, combining the confusion matrices above, it is not hard to see that SVM is not suitable for flight delay prediction. Though Gradient Boosting classifier performs well on predicting arrival delay here, it is not the best while predicting departure delay (discussed above). Thus, for both departure and arrival delay, Random Forest classifier is the first choice while all desired attributes including weather data are available.

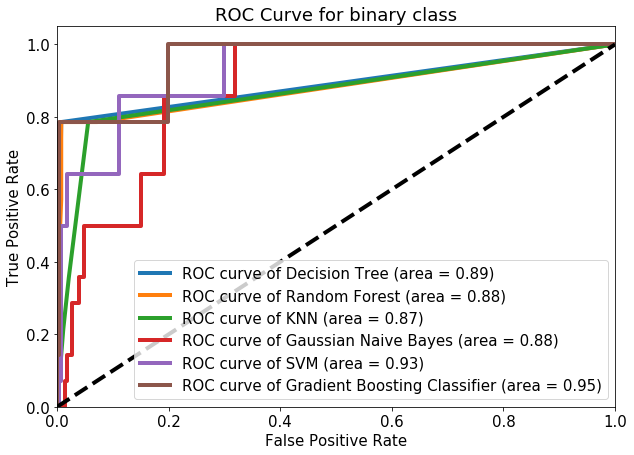


Figure 5.4 ROC curves of prediction models (arrival delay prediction with weather data).

**-Departure delay prediction using binary class (without weather data)**

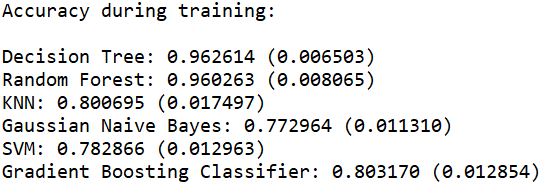


Figure 5.5 Cross-validation results of prediction models (departure delay prediction without weather data).

While weather data are missing, the classifier models are perform relatively worse compared to their performance while weather data are provided. However, Decision Tree and Random Forest are still reliable while 10-fold cross-validation during training as well as testing shown in Fig. 5.5 and Table 5.3 below. The other four classifiers including Gradient Boosting, are not reliable though they may still have high accuracy score around 0.8. The confusion matrices show their disability to predict departure delay while weather data are not provided.

Table 5.3 Testing results of prediction models (departure delay prediction without weather data).

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |
|  |  |
| KNN | Gaussian Naïve Bayes |
|  |  |
| SVM | Gradient Boosting Classifier |

The ROC curves for all six models are plotted in Fig. 5.6 below. According to the shape of the curves and the area calculated under the curve, Random Forest and Decision Tree both performs well in predicting departure delay while weather attributes are missing.

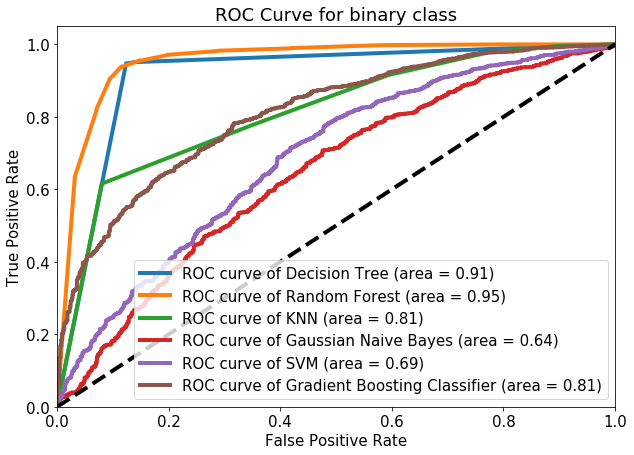


Figure 5.6 ROC curves of prediction models (departure delay prediction without weather data).

**-Arrival Delay Prediction using Binary Class (without weather data)**

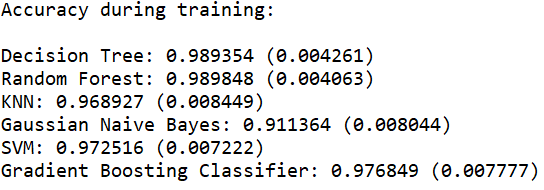


Figure 5.7 Cross-validation results of prediction models (arrival delay prediction without weather data).

In Fig. 5.7, all classifiers except Gaussian Naïve Bayes perform well during 10-fold cross-validation. However, this results is unreliable because of over-fit and other limitations. Combining the confusion matrices of testing in Table 5.4 below, only Decision Tree and Random Forest yield acceptable results while predicting arrival delay without weather data.

Table 5.4 Testing results of prediction models (arrival delay prediction without weather data).

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |
|  |  |
| KNN | Gaussian Naïve Bayes |
|  |  |
| SVM | Gradient Boosting Classifier |

In Fig. 5.8 below, the shape of curve and area under curves below show that Decision Tree, Random Forest, KNN, and Gradient Boosting all yield acceptable result. However, combining the confusion matrices above, we conclude that only Decision Tree and Random Forest are useful to predict arrival delay while weather data are missing.

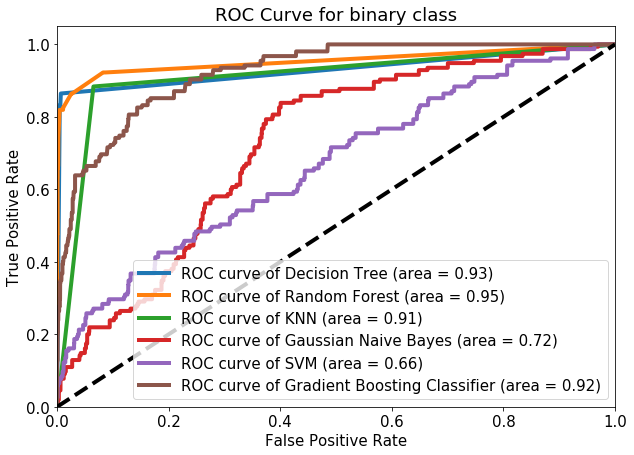


Figure 5.8 ROC curves of prediction models (arrival delay prediction without weather data).

**5.2 Predict delay with multi-type class (Delay Level)**

In Section 5.2 above, all results show that Random Forest and Decision Tree are two useful classifiers for predicting departure and arrival delay no matter if the weather data are available or not. Thus, in this section, we divide delay to four levels: 0. Early flights, 1. Slightly delay (within 10 min), 2. Medium delay (Under 30 min), 3. Long delay (Longer than 30min). Random Forest and Decision Tree classifiers are used for performing this multi-type classification.

**-Departure delay prediction with multi-type class (including weather data)**

The testing results including accuracy scores, confusion matrices, and reports are shown in Table 5.5 below, and multi-type class ROC curves are shown in Fig. 5.9 below. According to these results and curves, it is clear to conclude that both Decision Tree and Random Forest classifiers perform well in predicting departure delay while weather data are completely available. Decision Tree performs even better than Random Forest under this situation.

Table 5.5 Testing results of Decision Tree and Random Forest classifier (departure delay prediction with weather data).

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |

Figure 5.9 ROC curves of multi-type class (departure delay prediction with weather data).

**-Arrival delay prediction with multi-type class (including weather data)**

The testing results including accuracy scores, confusion matrices, and reports are shown in Table 5.6 below, and multi-type class ROC curves are shown in Fig. 5.10 below. According to these results and curves, it is clear to conclude that both Decision Tree and Random Forest classifiers perform well in predicting departure delay while weather data are completely available. Though Random Forest classifier’s ROC curves shown some advantage compared to the curves of Decision Tree, their confusion matrices show that they perform the same in predicting arrival delay level.

Table 5.6 Testing results of Decision Tree and Random Forest classifier (arrival delay prediction with weather data).

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |

Figure 5.10 ROC curves of multi-type class (arrival delay prediction with weather data).

**-Departure delay prediction with multi-type class (without weather data)**

The testing results including accuracy scores, confusion matrices, and reports are shown in Table 5.7 below, and multi-type class ROC curves are shown in Fig. 5.11 below. According to these results and curves, while weather data are missing, both classifiers perform slightly worse compared their performance while full weather data attributes are available. However, there is no doubt that both Decision Tree and Random Forest classifiers still perform well, and the prediction results are acceptable.

Table 5.7 Testing results of Decision Tree and Random Forest classifier (departure delay prediction without weather data).

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |

Figure 5.11 ROC curves of multi-type class (departure delay prediction without weather data).

**-Arrival delay prediction with multi-type class (without weather data)**

The testing results including accuracy scores, confusion matrices, and reports are shown in Table 5.8 below, and multi-type class ROC curves are shown in Fig. 5.12 below. According to these results and curves, it is clear to conclude that both Decision Tree and Random Forest classifiers yield acceptable results while predicting arrival delay level while weather data are missing. However, different than three situations above, Decision Tree shows obvious robust accuracy while predicting class 1 short delay than Random Forest.

Table 5.8 Testing results of Decision Tree and Random Forest classifier (arrival delay prediction without weather data).

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |

|  |  |
| --- | --- |
|  |  |
| Decision Tree | Random Forest |

Figure 5.12 ROC curves of multi-type class (arrival delay prediction without weather data).

**5.3 Logistic Regression**

In order to rank and select from all the features, recursive feature elimination (RFE), which works by recursively removing attributes and building a model on those attributes that remain, is implemented. In another word, RFE uses the model accuracy to identify which feature contributes the most to predicting the target class. Finally, based on RFE test results, the ranking of attributes for departure and arrival delay are shown below in Table 5.9 and 5.10. From Table 5.9, for departure delay, the weather attributes are highly ranked, and airline is the second important attribute that delay. For arrival delay shown in Table 5.10, it is expected that origin and destination are important, but air manufacturer becomes the first place that affecting arrival delay. Though this observation may has some limitations because our data like aircraft manufacturer are categorical and has been normalized for logistic regression. We still draw some useful conclusion from the RFE logistic regression. The logistic regression results shown in Fig. 5.13 confirms the conclusions from ranking attributes by RFE. The attributes x1 to x10 in the logistic regression report in Fig. 5.13 are the ten corresponding flight attributes in Table 5.9 and 5.10.

Table 5.9 Attributes ranking for departure delay.

|  |  |  |  |
| --- | --- | --- | --- |
| **Flight Attributes** | **Rank** | **Weather Attributes** | **Rank** |
| Aircraft Type | 7 | Cloud Altitude | 4 |
| Planned Cruise Speed | 12 | Temperature | 9 |
| Planned Cruise Altitude | 14 | Dew Point | 3 |
| Origin | 5 | Visibility | 10 |
| Destination | 15 | Wind Speed | 1 |
| Departure Week Day | 16 | Wind Gust Speed | 8 |
| Departure Time Hour | 6 |  |  |
| Arrival Time Hour | 13 |  |  |
| Airline | 2 |  |  |
| Aircraft Manufacturer | 11 |  |  |

Table 5.10 Attributes ranking for arrival delay.

|  |  |  |  |
| --- | --- | --- | --- |
| **Flight Attributes** | **Rank** | **Weather Attributes** | **Rank** |
| Aircraft Type | 9 | Cloud Altitude | 6 |
| Planned Cruise Speed | 5 | Temperature | 13 |
| Planned Cruise Altitude | 7 | Dew Point | 11 |
| Origin | 3 | Visibility | 2 |
| Destination | 4 | Wind Speed | 12 |
| Departure Week Day | 14 | Wind Gust Speed | 10 |
| Departure Time Hour | 8 |  |  |
| Arrival Time Hour | 16 |  |  |
| Airline | 15 |  |  |
| Aircraft Manufacturer | 1 |  |  |

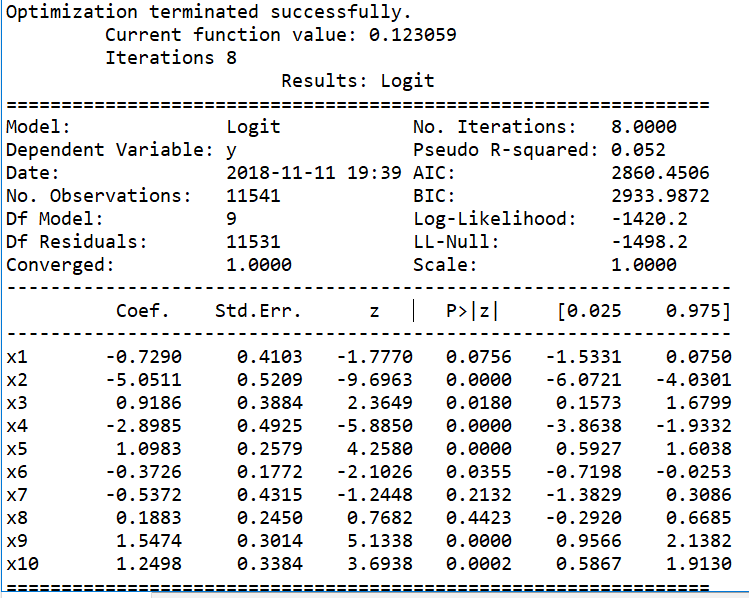


Figure 5.13 Logistic Regression Result.

Another point to research for logistic regression is that after selecting attributes based on logistic regression results, how will classifiers perform. Thus, we show ROC curves (predicting arrival delay) of all classifiers in Fig. 5.14 and 5.15 for comparison before and after feature elimination. Eight attributes are kept for data with weather and six attributes are kept for data without weather attributes. Both sets of figures show that the feature elimination does not affect the classifiers performance.

|  |  |
| --- | --- |
|  |  |
| Before Feature Elimination | Selecting 8 attributes |

Figure 5.14 ROC curves of classifiers with weather data for arrival delay.

|  |  |
| --- | --- |
|  |  |
| Before Feature Elimination | Selecting 6 attributes |

Figure 5.15 ROC curves of classifiers without weather data for arrival delay.

**5.4 Conclusion on prediction models.**

In this section, prediction models are implemented on our data set. Both binary class and multi-type class to be predicted have been applied. Situations that weather data are missing is also researched. According to all the results, we can draw the conclusion that for flight delay prediction, if all attributes listed above including weather attributes are provided, Random Forest is the first choice. If weather attributes are missing, both Random Forest and Decision Tree perform well. However, Decision Tree is preferred based on its advantage on accurately predict delay level.