**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning |
| **Assessment Title:** | Capstone Project Proposal - Machine Learning for Airlines Delay results prediction |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Jose Matias, Sertac Battal |
| **Student Number:** | 2023391, 2023472 |
| **Assessment Due Date:** | Sunday, 26 November 2023 |
| **Date of Submission:** | Sunday, 26 November 2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**1. Introduction**

Summer is the period of the year where most people enjoy their holidays and try to get some rest from their jobs, and wasting time with flights and time travelling from one destination to another can be really annoying. In order to help both the people travelling and the companies that offer the service, this project is going to look at the data of flights in the United States during the summer months, from June to August of 2023 for the companies that represent at least 1% of the market share.

The goal of this project is to analyse the data provided by the Bureau of Transportation Statistics and try to predict which flights are going to have a delayed departure of more than 15 minutes in relation to the original departure time. In order to do so, the data is going to be preprocessed and used as inputs for machine learning models and try to predict if the flight is going to be delayed or not, based on the airport of origin, the destination, the time when the flight is expected to take off and the carrier.

If the models provide a good result, the companies can utilise this information and try to implement solutions to reduce the delays, increasing their customer satisfaction and thus, their profit. On the other hand, the customers can utilise this data to avoid companies and airports that have a higher probability of having delayed flights and avoid wasting time during their holidays.

**2. Machine Learning Approaches**

**k parameters and cross validation**

**metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

In this graph, the cross-validation accuracy scores of the k-Nearest Neighbors (k-Nearest Neighbors - kNN) algorithm are shown for different k (number of neighbors) values.

The line on the graph shows the effect of k values on accuracy scores.

as the k value increases, the accuracy score begins to increase again and reaches a peak point. At this point, the optimum k(20) value of the model has been found.

As a result, by examining the maximum point on the graph, you can determine the optimal k value for your model, and using this value, it can be used for a more generalizable k-NN model.

**a. Choice of Machine Learning Approaches**

**k-The Nearest Neighbors (kNN)**

Unlike some other algorithms, which assume a complex relationship between properties, kNN simplicity is in Decadence. It works according to the principle that similar examples tend to exhibit similar behaviors. In the area of flight delays, this means that flights with similar characteristics, such as origin, destination and carrier, may experience comparable delays.

An accuracy score of 0.73 is the target column, which implies that the model correctly predicts flight delays about 73% of the time. This success rate shows that KNN is strong in distinguishing patterns in the Airline Delay dataset.

Comparison of model performance metrics.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

In the complex field of flight delays, where various variables are intertwined, k-Nearest Neighbors turn out to be a reliable ally. Its simplicity, adaptability and impressive accuracy make it a valuable asset in predicting and understanding flight delays.

**Selected Hyperparameters and Their Causes:**

The Number of K Neighbors (n\_neighbors): 5 was selected. A smaller K value allows the model to be more flexible, but it can lead to excessive compliance. 5 is a general value choice.

Accuracy: Measures the proportion of samples classified correctly.

**Random Forest**

The Random Forest model is a machine learning approach chosen to understand and predict flight delays. This preference is based on important reasons:

Random Forest is effectively designed to deal with various features in the Airlines Delay dataset. The large data set, which includes different airline carriers, departure points, destinations and timings, allows the model to handle its complexity. And Random Forest contains a mechanism that reduces the tendency of a tree to over-adapt. Each tree uses different subsets of the dataset, which helps the overall model make more balanced and accurate predictions.

**Random Forest Hyperparameters and Performance:**

Certain hyperparameters that affect the performance of the Random Forest model and the reasons for choosing these parameters are as follows:

**Selected Hyperparameters and Their Causes:**

n\_estimators: Randomly determines the number of trees in the forest. it was chosen as 100 because the combination of many trees usually increases the strength of the model, but reduces the risk of over-learning.

max\_depth: Determine the maximum depth of each tree. it was chosen as a 10 because deeper trees can usually identify more complex patterns, but this can increase the risk of over-adaptation.

**The Resulting Score:**

The Random Forest model attracts attention with the score obtained in this analysis. The highest score of 0.75 obtained reflects the success level of the model in predicting flight delays. This shows that the overall model is reliable and effective.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

The Random Forest model is a powerful tool for understanding the complexity in the Airlines Delay dataset and predicting flight delays. The combination of various hyperparameter selections and trees increases the generalization ability of the model, while reducing the risk of overconformity.

**3. Data Characterization and Cross-Validation**

Different training splits (%10, %20, %30) according to the cross-validation results performed using it, the Accuracy Scores we obtained are as follows.

**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu**

When the accuracy values obtained for each of the Random Forest and k-NN algorithms are examined, interesting findings are as follows;

it is observed that Random Forest has the highest accuracy rate in the 10% training division, while in the 20% and 30% training divisions, the accuracy difference between Random Forest and k-NN becomes more pronounced. Dec. This shows changes in the performance of various algorithms depending on the size and complexity of your data set.

The cross-validation results show that model performance may vary depending on how much of the training data is used. it is seen that Random Forest is more successful in the 30% educational division compared to k-NN. However, a more detailed study of the factors affecting the performance of each algorithm can provide a more in-depth understanding of these results.

**4. Results Interpretation and Discussion**

k-The Nearest Neighbors (kNN):

The kNN model has performed Decently by successfully determining the relationships between flights with similar characteristics. However, depending on the choice of a certain number of neighbors and the metric preference, the model has a risk of over-fitting or under-fitting. Optimal parameter selection helps the model to achieve accurate and overall results.

Random Forest (Random Forest):

The Random Forest model offered a stronger prediction ability with the combination of many decision trees. The risk of over-adaptation is lower because each tree is trained at a certain depth. This allows the generalization to be better. However, it is important to maintain an accurate balance with the combination of a certain number of trees.

Discussion on Overconformity, Underconformity and Generalization:

Overfitting: Can be evident in the decision tree model, especially if the depth level is high. This is an indicator of too much focus on educational data and difficulty adapting to real-world data.

Underfitting: It can be seen in the kNN model or in cases where certain parameters are selected incorrectly. This means that the model cannot fully adapt to the educational data and cannot capture the complexity.

Generalization: Random Forest is more successful in generalization. Because the combination of many decision trees reduces the risk of over-compliance and under-compliance, which allows the model to better adapt to more diverse data sets.

The Rational of the Preferred Model:

Since the Random Forest model shows a balanced result in terms of performance and generalization ability, it stands out as the preferred model in this analysis. The combination of various trees is notable for its ability to exceed KNN's predictions based on similar characteristics, while reducing the risk of over-matching the decision tree.

**5. Code Comments and Conclusions**

Comments in Python code.

Summary of project results and conclusions.

**6. Individual Contributions and Reflective Journal (Sertac)**

As a member of the team, we made decisions and made applications by talking to my other team friend at every stage from the beginning that we had done in our work. First of all, we wanted to do a more advanced study by holding meetings together in determining the data set (physical meetings, online meetings), looking at the problem of our day and recent data. As a result, we discovered the data set that was created according to the delay time of the aircraft(340 carriers). June July, August our aim here is to examine the summer months (June, July, August) which are the most extreme and we have recently examined the delay situation. As the reasons for the delay, we have removed all the columns that we considered unnecessary in the data set with a lot of data and applied ML methods with the minimum values that we will need.

During this process, my other tasks on the team were to determine the appropriate methods, data scaling, PCA and obtaining accurate scores based on the values of train split 10 20 30. In addition, I also explained some of the titles in the report we wrote together.

We spent most of our Surah with data selection and data cleaning. In addition, sura received ML methods due to the large size of our data set in the field.

Below you can find the pie chart of the time I spent

metin, ekran görüntüsü, diyagram, daire içeren bir resim

Açıklama otomatik olarak oluşturuldu

Team member Sertac has made 6 commits since the beginning of the project.

Team member Jose has made 12 commits to add features.

(With screen sharing has also been done in online meetings and improvements have been made together)

8. Conclusion

Summary of key findings.

Closing remarks.

9. References

Harvard Style citations and references.