Quality of Candidate Flights and Submission Prediction in Collaborative Digital Departure Reroute

Sarah Youlton*
Universities Space Research Association
at NASA Ames Research Center, Moffett Field, CA, USA

Alexandre Amblard†
Universities Space Research Association
at NASA Ames Research Center, Moffett Field, CA, USA

William J. Coupe‡
NASA Ames Research Center
Moffett Field, CA, USA

Collaborative Digital Departure Reroute (CDDR) enables the reroute of flights using a flight operator proposed set of alternative route options, referred to as Trajectory Option Set (TOS), in order to reduce delay on the airport's surface and in the Metroplex environment. The reroute functionality is enabled through NASA's Digital Information Platform (DIP). TOS candidate flights are defined as flights with an alternative route with delay savings greater than the flight operator defined relative trajectory cost. This paper analyzes the TOS candidate flights at Dallas/Fort Worth International Airport (KDFW) in the North Texas Metroplex to gain insight into which candidate flights are higher quality through a scoring method. This insight will inform refinements to help CDDR focus on high quality reroute opportunities. Binary classification models for predicting the flight operator's submission of candidate flights are also explored in this paper.

I. Introduction

In partnership with the Federal Aviation Administration (FAA) and commercial airlines, NASA has been developing various decision support tools and systems to improve operations in both single airport and Metroplex environments. During Phase 3 of the Airspace Technology Demonstration 2 (ATD-2) project [1], the ability to reroute flights was implemented in the North Texas Metroplex which includes two major airports, Dallas/Fort Worth International Airport (KDFW) and Dallas Love Field Airport (KDAL). In alignment with the FAA's vision for an Info-Centric National Airspace System (NAS) [2], this reroute capability along with other ATD-2 systems underwent a digital transformation and were deployed on NASA's Digital Information Platform (DIP). The Collaborative Digital Departure Reroute (CDDR) concept supports the reroute decision process by predicting delay savings on each alternative route in the flight operator's Trajectory Option Set (TOS), where TOS is defined as the collection of alternative routes for a flight that the flight operator will consider [1, 3–5]. Using CDDR, flight operators and Air Traffic Control (ATC) can assess reroute options with the main goals of reducing delay, fuel burn, and CO_2 emissions. For CDDR, focus is placed on candidate flights where the TOS alternative route meets set delay savings requirements at push back from the gate, referred to as the OUT event [1, 4, 5].

The reroute functionality is provided on the DIP system through the CDDR Service. This paper studies CDDR candidate flights in the North Texas Metroplex at KDFW during an operational field evaluation, taking place from April 28, 2022 through March 27, 2023, with the objective of gaining insight into which candidate flights are higher quality through a scoring method. Binary classification models for predicting the flight operator's submission of candidate flights are also explored in this paper. Predicting candidate flight submission can indicate if a candidate flight is high quality and also give insight into operational planning. Section II will give more background information about the CDDR Service. Section III will provide more details on the data and metrics used. Section IV will discuss results of the

*Software Engineer, NASA Ames Research Center.
†Data Scientist, NASA Ames Research Center.
‡Aerospace Engineer, NASA Ames Research Center.

scoring method developed to determine the quality of candidate flights. Section V will explore the binary classification models. Section VI will provide concluding remarks and future applications.

II. Collaborative Digital Departure Reroute

The North Texas Metroplex consists of two major airports, KDFW and KDAL, and is within the D10 Terminal Radar Approach CONtrol (TRACON) shown in Fig. 1. This TRACON is composed of sixteen departure fixes along the terminal boundary, organized in groups of four in each cardinal direction. Each group of four fixes is referred to as a departure gate. The departure fix capacity can be impacted and reduced from severe weather and separation requirements from Traffic Management Initiatives (TMIs) within the TRACON, leading to increased delay on the surface of the airports [1, 3–5]. The CDDR Service provides decision support to flight operators and ATC to reroute flights using a TOS alternative route on a different departure fix to avoid restrictions and reduce delay [1, 3, 5]. Although the alternative route may reduce delay, it often is a longer route [1, 3–5]. The flight operators must consider this trade-off and decide when it is best to choose the alternative route over the original filed route.

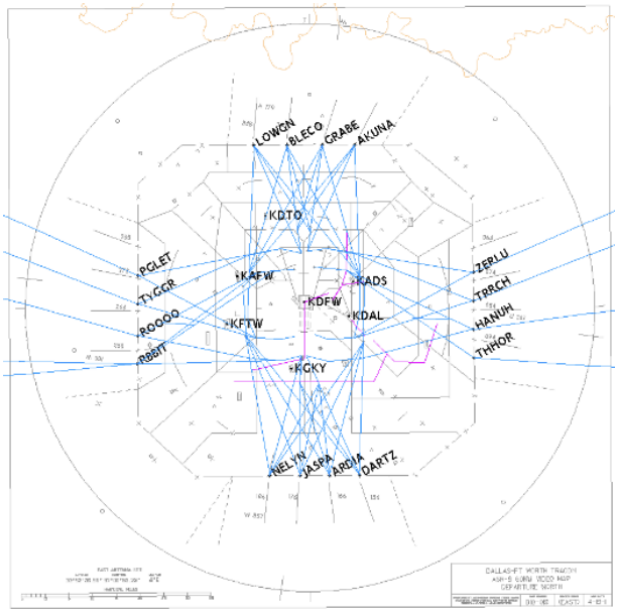

Fig. 1 D10 TRACON.

The predicted delay on the filed route and TOS alternative routes are determined using the estimated take-off time (ETOT) generated by the Terminal Scheduler, a component of the CDDR system [1, 3–5]. To help the flight operators assess the alternative routes, the CDDR Service computes delay savings on each alternative route in the flight operator's TOS relative to the filed route [1, 3–5]. Each alternative route is also given a Relative Trajectory Cost (RTC), a duration cost in minutes, that is assigned by the flight operators. The RTC value often reflects the increased distance of the alternative route and the desire of flight operators to utilize the alternative route [1, 3–5]. When the predicted delay savings for an alternative route is greater than the RTC value, the flight is considered a candidate for reroute [1, 4, 5]. In addition to the delay savings of the rerouted flight, the CDDR Service calculates the aggregate savings from rerouting the flight [1, 3–5]. Upon notification of candidate flights, if the flight operator determines that the reroute should be implemented, the flight operator submits a reroute request to ATC and the flight status is recorded as Submitted. In this paper, the Submitted status is used as an indication that a candidate flight was high quality.

III. Data and Metrics

The data used to develop the quality of candidate flights scoring method and prediction models were collected between April 28, 2022 and March 27, 2023 from the operational CDDR system at KDFW. During this time range, 927 candidate flights were identified and 107 of these candidates were submitted as reroutes by flight operators. The breakdown of data for overall and monthly submission status is shown in Fig. 2.

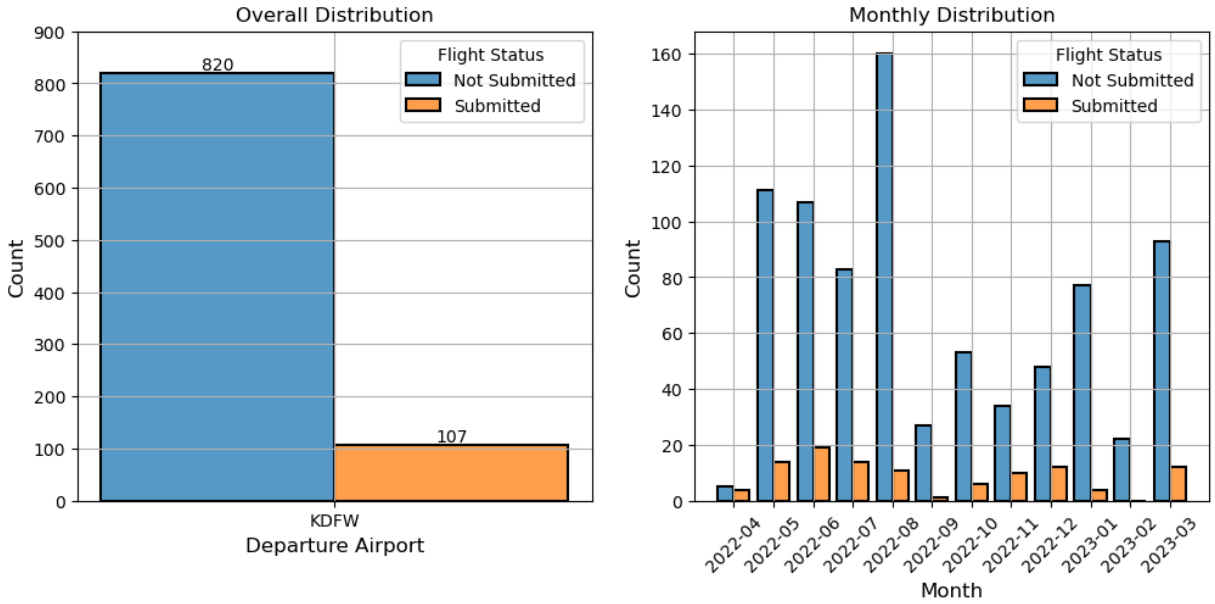

Fig. 2 Overall and monthly submission distributions.

Based on our operational experience and observations, six data elements were identified as important features to consider for flight submission: 1) candidate duration, the amount of time the TOS route was considered a candidate flight (in minutes), 2) OFF delay savings, the estimated delay savings at the runway (in minutes), 3) IN delay savings, the estimated delay savings for the TOS route relative to the Reference Filed Route at the arrival gate (in minutes), 4) aggregate delay savings for the airline, estimated aggregate delay savings for all subsequent flights from the same carrier if the flight used the TOS route, 5) Controller-Pilot Data Link Communication (CPDLC) equipment, whether the aircraft has equipment to communicate reroutes digitally, and 6) delay savings over RTC probability, the probability that the OFF delay savings on the TOS route will be greater than the RTC value. These six features were used to create the scoring method and the binary classification models.

The OFF delay savings, IN delay savings, aggregate delay savings for the airline, and delay savings over RTC probability were sampled both at the flight's push back time from the gate, referred to as the OUT event, and at the last time the flight operator submitted the reroute, referred to as the final submission event. If a reroute was submitted, the data element was taken at the final submission event. Otherwise, the data element was taken at the OUT event. For submitted reroutes, the difference between the OUT event and the final submission event ranged from -62.93 minutes to 246.38 minutes where a positive value indicates the final submission event was before the OUT event and a negative value indicates the final submission event was after the OUT event. The mean, median, and interquartile range of the difference were computed and found to be 24.61 minutes, 20.20 minutes, and 29.13 minutes, respectively.

A handful of data engineering steps were taken to prepare the data for the different models. The Boolean data elements for CPDLC equipment and Submitted were converted to integers so "False" became 0 and "True" became 1. All of the delay savings columns were adjusted so a positive value indicated savings, meaning a positive value signifies a reroute saved time and a negative value signifies a reroute lost time. Candidate flights with null values in any of the six features were removed. During this time period, 11 candidate flights with null values were identified and are not included in the counts in Fig. 2.

Figure 3 shows the distributions of the z-score values of each of the six data elements for submitted and not submitted candidate flights using ten bins. The mean and standard deviation used to compute these z-score values were calculated during an operational field evaluation [6], taking place from April 28, 2022 through September 16, 2022. Data in this date range are used as the training data set for the scoring method and all models. To better view the submitted flights distribution, Fig. 4 shows the distributions of the z-score values of each of the six data elements with percentages where the submitted and not submitted groups are determined independently.

Outliers were identified for each of the six normalized features. For this paper, an outlier was defined as a value outside of the range of the median of the normalized feature in the full data set ±3.5 ×IQR where IQR is the interquartile

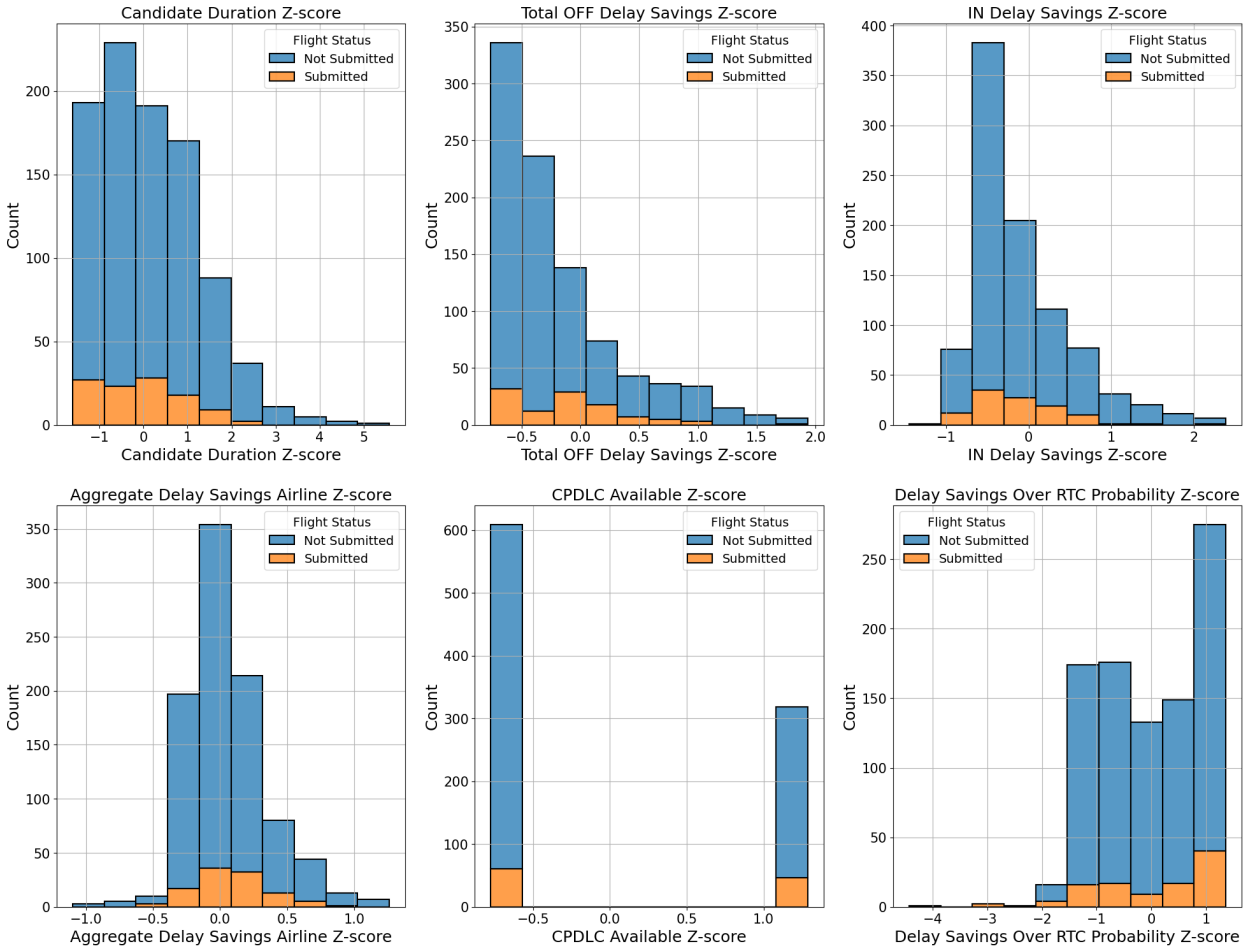

Fig. 3 Z-score distributions for submitted and not submitted candidate flights in 10 bins.

range. The full data set was used to determine outliers due to the exploratory nature of this analysis project and the large amount of noise in this type of data. A total of 68 outlier candidate flights were identified and removed from the full data set where 67 of the outlier candidate flights were not submitted and 1 of the outlier candidate flights was submitted. The outliers are also not included in the counts in Fig. 2. All analysis was performed using the data set filtering out the outliers.

As mentioned above, the operational field evaluation between April 28, 2022 and September 16, 2022 was used as the training data set. After the demonstration period, the CDDR system continued to operate in the field, collecting additional data that were used as a testing set to evaluate models. Table 1 shows a breakdown of the date ranges, percentages of data, and submission counts for the training and testing groups.

Some challenges faced in this paper are both the data collection and interpretation. As an example, candidate flights will reflect high benefits during severe weather events. However, since downstream weather is not taken into account, this type of unknown constraint could be cause for a flight operator to not submit a candidate flight. These unknown constraints make it difficult to analyze data because flights with high predicted delay savings may not actually be good

Table 1 Training and Testing Groups

Group	Date Range	Percentage of Data	Candidate Flights Not Submitted	Candidate Flights Submitted
Training	4/28/22 - 9/16/22	57.8%	473	63
Testing	9/17/22 - 3/27/23	42.2%	347	44

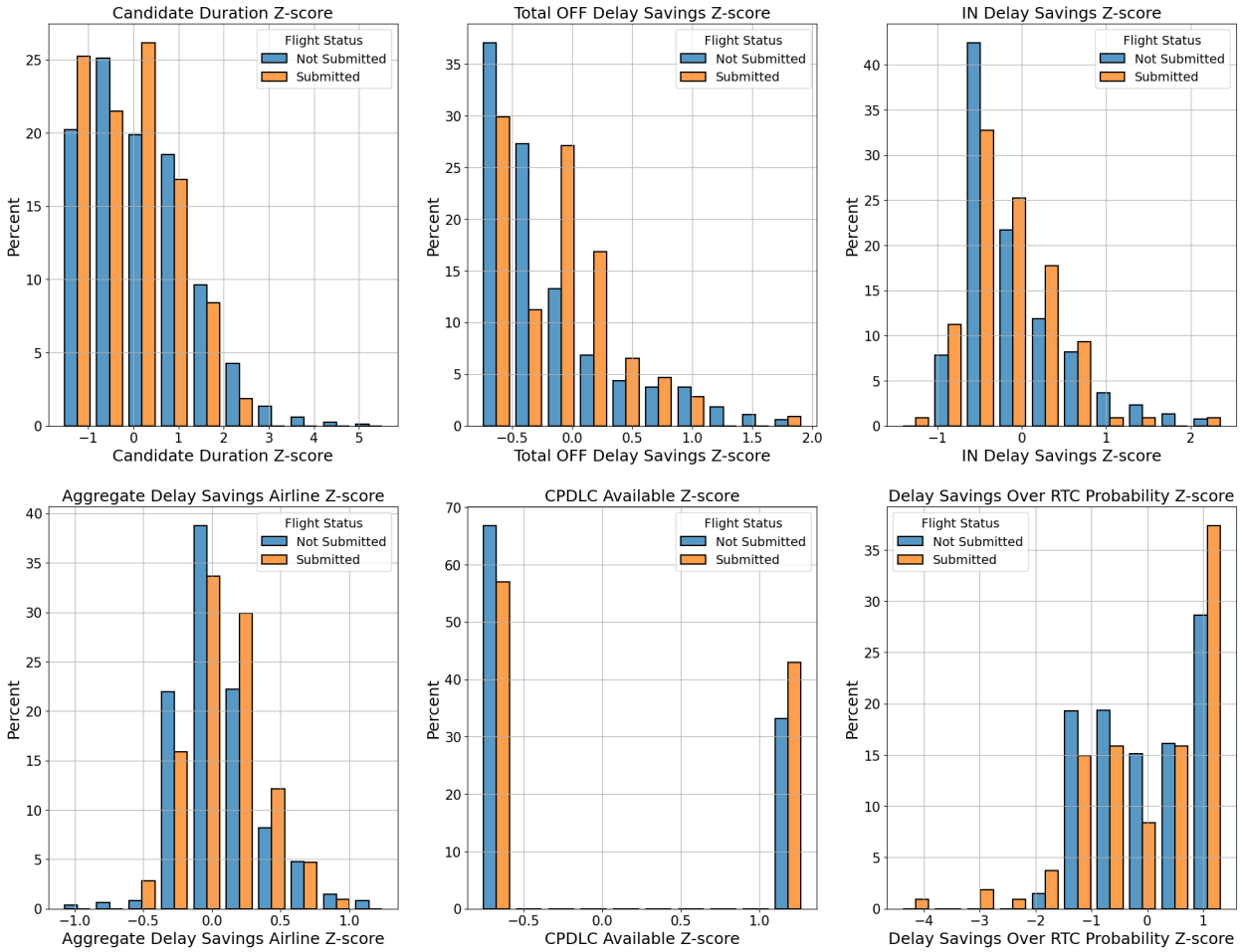

Fig. 4 Z-score distributions for submitted and not submitted candidate flights in 10 bins with percentages.

candidate flights. In addition, the way the data are sampled may introduce some bias. These types of obstacles help explain some of the less intuitive results.

Various metrics were used to assess the performance of the scoring method and classification models. For the quality of candidate score, the residuals and R^2 values were computed for the best fit line along with slope. Accuracy and f1-score were computed for the binary classifiers with emphasis on the macro f1-score due to the imbalanced classes. The following sections outline the method used to score the quality of candidate flights and the models built to predict if a candidate flight will be submitted.

IV. Quality of Candidate Score

To gain more insight into the quality of a candidate flight, a technique to assign a score to each candidate flight was developed. This scoring method was performed post-operation and gives insight into whether the candidate flight was a good option for submission. One approach to score candidate flights is to calculate a summation of weighted normalized data features where each determined weight, w_n, is multiplied by each normalized feature, f_n, using the six operationally important data elements mentioned above. Combining multiple important features into a single score could help users easily identify which candidate flights to focus on. This score will be computed for each candidate flight. The linear combination is seen in Eq. (1).

$$\text{Quality Score} = \sum_{n=1}^{6} w_n f_n \tag{1}$$

The main challenge in this task is that there is no objective truth to compare the score against to determine the best weights to use. Since there is no true score, the submission event is used as an indicator that a candidate flight was high quality. Each candidate flight was assigned a bin based on its computed quality score. For each bin, the fraction of submitted candidate flights over the total number of candidate flights in the bin, or submitted fraction, was determined and used to assess performance of the scores. This submitted fraction corresponds to the empirical probability of submission for each bin.

As a baseline, the quality scores were computed with all weight values set to 1 and then the scores were normalized between 0 and 1. Figure 5 shows box plots of the normalized scores for the submitted and not submitted groups. Based on the normalized score, each candidate flight was assigned a bin, using a total of ten bins. For each bin, the mean normalized score and submitted fraction were evaluated. Using the mean normalized score as the x axis and the submitted fraction as the y axis, best fit lines were drawn for the training and testing data sets. Figure 6 shows the linear combination results using the weights of 1, where each bin is labeled with candidate flight count. The training data set produced a best fit line with residuals = 0.0785, R^2 = 0.1082, and slope = 0.1084. With weights of 1, the best fit line from the training data was applied to the testing data, resulting in residuals = 1.7649 and R^2 = -0.2984. When applying an independent best fit line to the testing data, the weights of 1 produced the best fit line with residuals = 0.5393, R^2 = 0.6033, and slope = -0.9772. It is clear from these metrics that the six data elements should not be weighted equally due to very different trends in the training and testing data sets. These results provided motivation into exploring weight permutations for each of the six features.

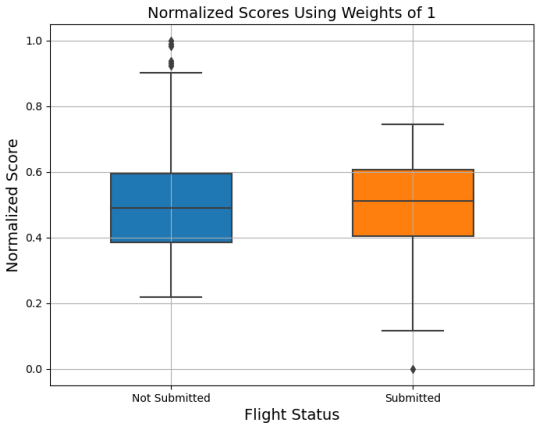

Fig. 5 Box plot of normalized scores for submitted and not submitted groups.

All possible weight permutations for the six features were generated using a nested loop with a list of predefined values ranging from -1 to 1 in 0.2 increments, resulting in 1,771,561 options. A subsequent loop cycled through each of these possible weight permutations and performed various calculations. For every permutation of weight values, the quality scores were computed and normalized between 0 and 1. Each candidate flight was assigned a bin based on its score, using a total of ten bins. The submitted fraction and mean normalized score were computed for each bin. Some weight permutations produced empty bins so performance was only assessed if at least six out of the ten total bins were defined. To assess performance, a best fit line was created using the mean normalized score as the x axis and the submitted fraction as the y axis. Residuals, R^2, and slope were all calculated for the best fit line as evaluation metrics. The polyfit function from NumPy [7] was implemented to draw the best fit line and determine the residuals which uses the least squares method by finding the sum of the squared residuals.

Within the loop, these steps above were performed with the full training data set and with scikit-learn stratified 3-fold cross validation on the training data set. The evaluation metrics were calculated for each training fold, validation fold, and the full training data set. If the slope of the best fit line had equivalent signs for each training fold, each validation fold, and the full training data set, the average residuals and average R^2 were computed. If the sign of the slope did not match, the weight permutation was not considered due to the trend being inconsistent. The weight permutations that produced the minimum average residuals and maximum average R^2 were reported.

Three sets of weights were identified using this method, two results from the minimum average residuals and one result from the maximum average R^2. The two results with the minimum average residuals were the same set of weights

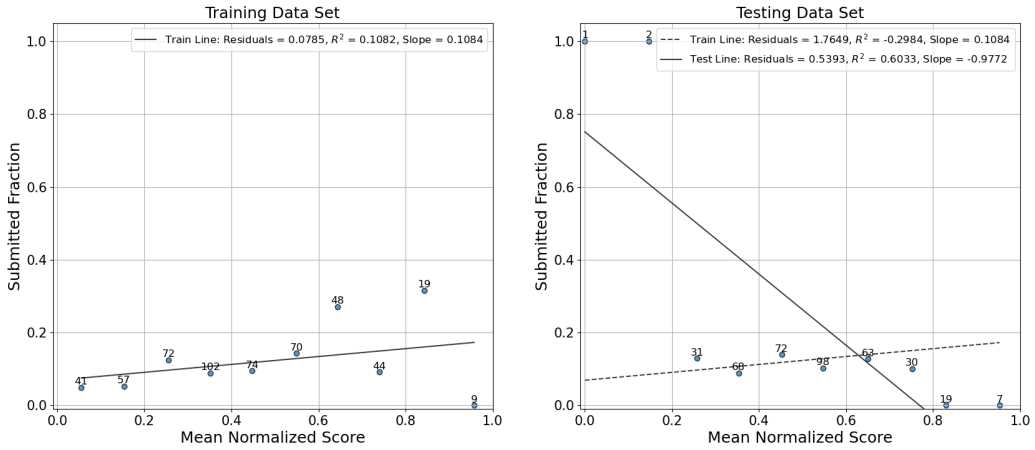

Fig. 6 Baseline linear combination results for training and testing data sets.

but with opposite signs. Since the performance was equivalent, the set of weights that produced a positive slope, where higher scores have a higher submitted fraction, was chosen to make the scoring more intuitive. This set of weights along with the set of weights that produced the maximum average R^2 were applied to the testing data. The same steps were performed with the testing data set. The quality score was calculated for each candidate flight and candidate flights were then grouped into ten bins. For each bin, the mean normalized score and the submitted fraction were computed. Residuals and R^2 values of the best fit line were determined. The weight permutation that produced the maximum average R^2 during training had opposite trends in the training and testing data set and therefore, had unreliable performance. The weight permutation that produced the minimum average residuals during training performed decently on the testing data set and was selected as the best result.

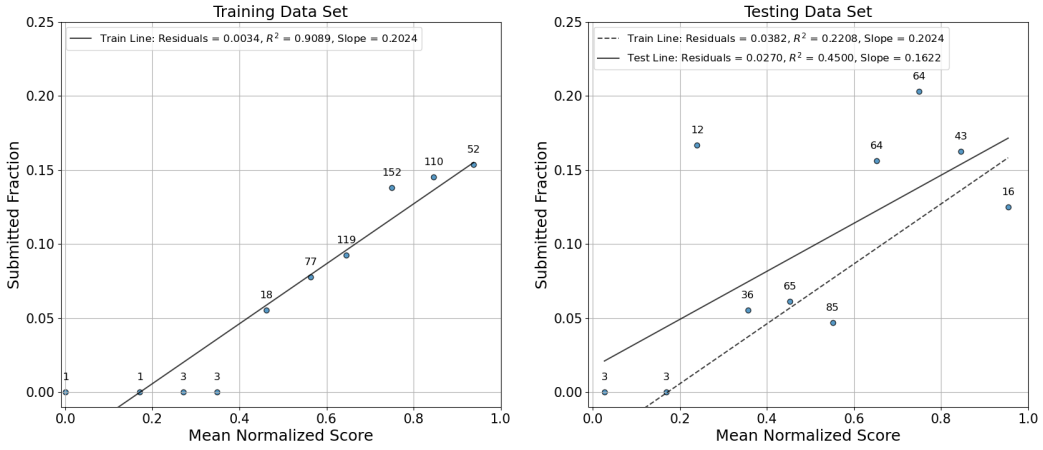

Fig. 7 Best linear combination results for training and testing data sets.

The best weights were $w_1 = 0$, $w_2 = 1$, $w_3 = -1$, $w_4 = 0$, $w_5 = 0.2$, and $w_6 = 0.4$ for features candidate duration, OFF delay savings, IN delay savings, aggregate delay savings for the airline, CPDLC equipment, and delay savings over RTC probability, respectively. These weights suggest the candidate duration and the aggregate delay savings for the airline are not important in the scoring method. The graphs of the mean normalized score vs the submitted fraction for the training and testing data sets using these weights are shown in Fig. 7, where each bin is labeled with candidate flight count. On the training data, the weights produced a best fit line with residuals = 0.0034, R^2 = 0.9089, and slope = 0.2024. These weights and the best fit line from the training data were then applied to the testing data, resulting in residuals = 0.0382

and R^2 = 0.2208. When applying an independent best fit line to the testing data, the weights produced the best fit line with residuals = 0.0270, R^2 = 0.4500, and slope = 0.1622.

Another obstacle with developing the weighted sum was how computationally expensive it was to analyze the performance of each weight permutation. Because of the high number of permutations and the run time needed for each one, it was difficult to add in additional weights to the list of predefined values for further exploration.

V. Binary Classification Model

The second part of this paper was to build a binary classification model using historical data to predict whether a candidate flight will be submitted by flight operators. The same six normalized data elements were used as model features. As shown in Table 1, the operational field evaluation between April 28, 2022 and September 16, 2022 was used to train the model and the remaining data were used as the testing data set.

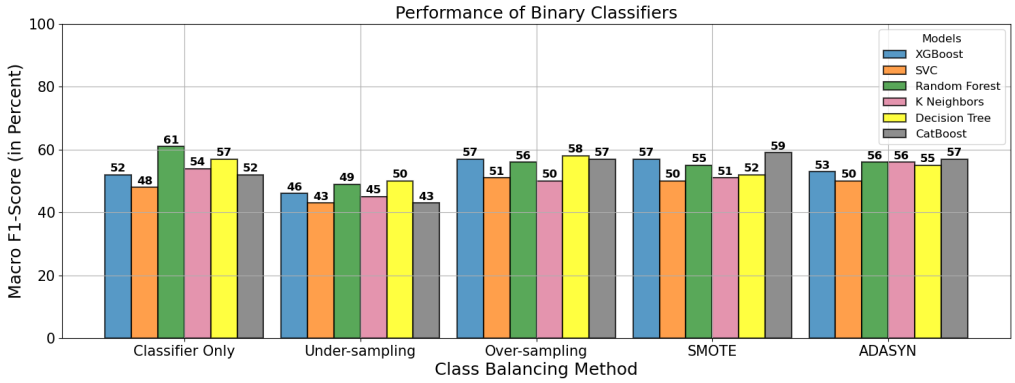

Fig. 8 Performance of binary classifiers and class balancing methods.

A number of classification models were explored including XGBoost [8] Classifier, CatBoost [9] Classifier, and scikit-learn [10] models Support Vector Classification, Random Forest Classifier, K Neighbors Classifier, and Decision Tree. Additional models were explored that did not perform as well. Due to the large difference in submitted vs not submitted class sizes, various class balancing methods from imbalanced-learn [11] were considered such as random under-sampling, random over-sampling, Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic (ADASYN). An imbalanced-learn pipeline [12] was constructed for every combination of class balancing technique and classifier pair in addition to every classifier individually as pipeline steps. Each pipeline was then fed into scikit-learn GridSearchCV to tune the hyperparameters of the classification models using scikit-learn stratified 5-fold cross validation on the training data. Due to the imbalanced classes, the macro f1-score metric was focused on. The hyperparameters that produced the best macro f1-score in GridSearchCV were used to train a model on the training data and then predict the classes for the testing data. The performance of the testing data set is shown in Fig. 8 for each model pipeline.

The highest performing model was the Random Forest classifier without a class balancing method which produced a macro f1-score of 61% and accuracy of 88%. From hyperparameter tuning, this model used entropy to measure the quality of a split, no set maximum tree depth, minimum samples leaf of 1, minimum samples split of 2, and 300 estimators without bootstrapping. Other models that had close performance were CatBoost with SMOTE producing a macro f1-score of 59% and accuracy of 79% and Decision Tree with random over-sampling producing a macro f1-score of 58% and accuracy of 80%. For the class balancing methods, random under-sampling didn't help any of the models overall. To learn which features are more significant to the prediction model, the built-in scikit-learn feature importance function was applied to the Random Forest classifier with the results shown in Fig. 9. In this figure, the delay savings over RTC probability and the OFF delay savings are the top two significant features, which is in agreement with our predicted most important data elements.

When comparing the feature importance from the Random Forest classifier with the best weights from the quality score, some of the results seem contradictory. The quality score forced a linear relationship in the data and focused on the overall submitted fraction for a bin. The Random Forest classifier is a nonlinear model predicting the individual submission for each candidate flight. These fundamental differences between the methods may not allow direct comparison of the results. Furthermore, the uncertainty and potential bias in the data could contribute to the opposed

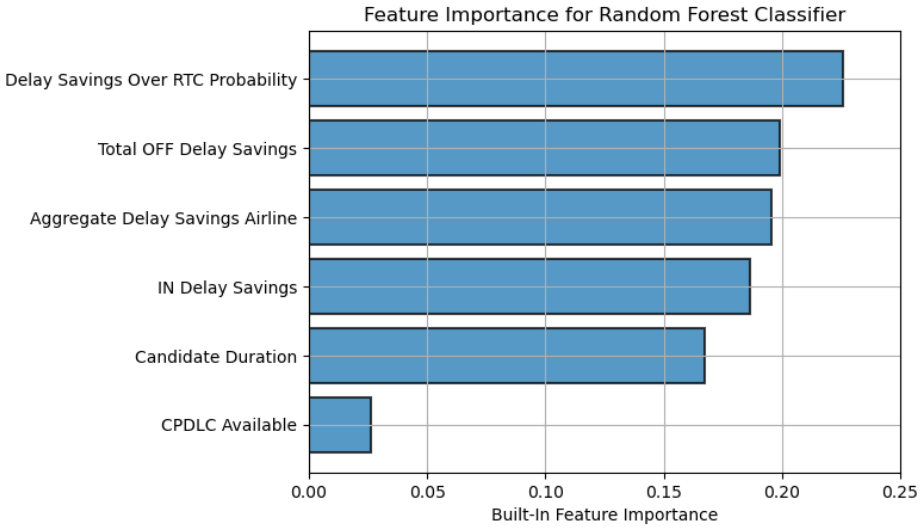

Fig. 9 Feature importance for binary classification.

results. Another possible explanation is that since some of the features are correlated, the scoring method and the classifier simply choose to use certain features over others. These differing results motivate continued future investigation into these methods.

VI. Conclusion

The CDDR system supports the reroute decision process enabled by DIP. In this service, candidate flights are identified and potential delay savings are computed. Through the use of a quality of candidate score, users can gain insight into which candidate flights are higher quality. Six operationally important features were identified: candidate duration, OFF delay savings, IN delay savings, aggregate delay savings for the airline, CPDLC equipment, and delay savings over RTC probability. To score candidate flights, a weighted sum of these six important features was explored by cycling through a predefined set of weights. The quality score was computed using each weight permutation and candidate flights were divided into ten bins based on quality score. The best fit line was drawn for each weight permutation using the mean normalized bin score as the x axis and the submitted fraction as the y axis. Metrics were computed on the best fit lines to determine the best weight permutation. Using this method, the best weights were w_1 = 0, w_2 = 1, w_3 = -1, w_4 = 0, w_5 = 0.2, and w_6 = 0.4 for features candidate duration, OFF delay savings, IN delay savings, aggregate delay savings for the airline, CPDLC equipment, and delay savings over RTC probability, respectively. Using these weights, the training data set resulted in a best fit line with residuals = 0.0034, R^2 = 0.9089, and slope = 0.2024. Applying the weights and best fit line from training to the testing data set resulted in residuals = 0.0382 and R^2 = 0.2208. The independent best fit line on the testing data using these weights produced the best fit line with residuals = 0.0270, R^2 = 0.4500, and slope = 0.1622.

In addition to the quality of candidate flight score, binary classification models were explored to predict whether a candidate flight will be submitted by flight operators. A handful of models along with class balancing techniques were implemented using an imbalanced-learn pipeline in GridSearchCV. The best performing binary classifier was the Random Forest classifier without a class balancing technique. This model produced a macro f1-score of 61% and accuracy of 88%. From this best model, the six features were also analyzed to assess feature importance. As expected, the delay savings over RTC probability was the most important feature followed by OFF delay savings and aggregate delay savings for the airline.

Overall, this paper outlines a method to assign a quality score to each candidate flight and a binary classification model to predict flight operator reroute submission. Due to the computationally expensive nature of developing a linear combination without truth data, only a set list of predefined weights was explored. Future work could assess other weight possibilities or investigate other methods to develop a quality score. Additionally, other data elements could be tested as features for both the scoring method and the classification model. Further extension of this work could be incorporating the scoring method and submission prediction model into real-time operations. The quality of candidate

score could enhance the CDDR service by providing a single value flight operators can focus on in real-time to aid in submission decisions rather than comparing multiple data elements individually. The submission prediction model could be implemented into an additional service to help assess the quality of a candidate flight and assist in real-time operational planning.

Acknowledgments

The material in this paper is based upon work supported by the National Aeronautics and Space Administration under Contract Number NNA16BD14C, managed by the Universities Space Research Association (USRA). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration.

References

[1] Coupe, W. J., Bhadoria, D., Jung, Y., Chevalley, E., and Juro, G., "ATD-2 Field Evaluation of Pre-Departure Trajectory Option Set Reroutes in the North Texas Metroplex," *2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC)*, IEEE, 2022, pp. 1–10.

[2] "Charting Aviation's Future: Operations in an Info-Centric National Airspace System," `https://www.faa.gov/sites/faa.gov/files/Charting-Aviations-Future-Operations-in-ICN_0.pdf`, 2022. Accessed: 2023-05-15.

[3] Coupe, W. J., Jung, Y., Chen, L., and Robeson, I., "ATD-2 Phase 3 Scheduling in a Metroplex Environment Incorporating Trajectory Option Sets," *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)*, IEEE, 2020, pp. 1–10.

[4] Chevalley, E., Juro, G. L., Bakowski, D., Robeson, I., Chen, L. X., Coupe, W. J., Jung, Y. C., and Capps, R. A., "NASA ATD-2 Trajectory Option Set Prototype Capability For Rerouting Departures in Metroplex Airspace," *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)*, IEEE, 2020, pp. 1–10.

[5] Coupe, W. J., Bhadoria, D., Jung, Y., Chevalley, E., and Juro, G., "Shadow Evaluation of the ATD-2 Phase 3 Trajectory Option Set Reroute Capability in the North Texas Metroplex," *Fourteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM 2021)*, 2021, pp. 1–10.

[6] Coupe, W. J., Amblard, A., Youlton, S., and Kistler, M., "Machine Learning Airport Surface Model," *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*, IEEE, 2023, pp. 1–10.

[7] "numpy.polyfit," `https://numpy.org/doc/stable/reference/generated/numpy.polyfit.html`, accessed: 2023-11-01.

[8] "XGBoost Documentation," `https://xgboost.readthedocs.io/en/stable/`, accessed: 2023-11-01.

[9] "CatBoost," `https://catboost.ai/en/docs/`, accessed: 2023-11-01.

[10] "scikit-learn Machine Learning in Python," `https://scikit-learn.org/stable/`, accessed: 2023-11-01.

[11] "imbalanced-learn documentation," `https://imbalanced-learn.org/stable/index.html`, accessed: 2023-11-02.

[12] "Imbalanced Learn Pipeline," `https://imbalanced-learn.org/stable/references/generated/imblearn.pipeline.Pipeline.html`, accessed: 2023-11-02.