IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

FLIGHT DELAY PREDICTION WITH ERROR CLASSIFICATION

[1]Guide – Ms. B. Jyothi

Assistant Professor

[2] R. Pravalika[3]G. Srinidhi [4]P. Viswas

Department of Computer Science and Engineering

ANURAG GROUP OF INSTITUTIONS

ABSTRACT

Accurately predicting flight delays is fundamental to building a more efficient airline business. An airline's core business is customer satisfaction. Our forecasts are critically important in the decision-making process of all civil aviation stakeholders. Bad weather, mechanical reasons and delayed arrival of the aircraft at the point of origin cause flight delays and customer dissatisfaction. Using flight data and weather data, we propose a prediction model for on-time arrival flights**.** This project uses machine learning models such as decision tree regression, Bayesian Ridge, random forest regression, and gradient boosting regression to predict whether a particular flight will arrive late.

INTRODUCTION

A flight delay occurs when an airline lands or takes off later than the scheduled arrival or departure time. Air pollution has been increasing rapidly in recent years. The global airline industry suffers huge losses due to many factors including flight delays. Flight delays are usually covered by all companies involved. H. Airports, airlines, passengers, troubles. Predicting flight delays accurately and thoroughly using factors that play a surprising role is key to minimizing losses and increasing customer satisfaction.

In the United States, the FAA considers a flight delayed if there is a difference of 15 minutes or more between the scheduled time and the actual arrival time. This is a serious problem in the United States, so research is currently underway to analyze and predict flight delays in order to significantly reduce costs. The main reasons for delays on scheduled commercial flights are bad weather, air congestion, delays of aircraft used for flights from previous flights, maintenance and safety issues.

In this, we have used several machine learning algorithms to comparatively study the accuracy of each algorithm. Predict whether a particular flight will be delayed using machine learning models such as decision tree regression, Bayesian ridge, random forest regression, and gradient boosting regression.

LITERATURE REVIEW

[1] Chakrabarty, Navoneel. (2019). A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines.

The current US domestic scenario has resulted in many flight delays and cancellations. American Airlines is one of the most trusted airlines in the United States and the largest airline in the world by number of destinations. However, on domestic flights, AA does not live up to expectations in terms of punctuality or punctuality. Flight delays also cost commercial airlines a lot. They will therefore do their best to prevent flight delays or cancellations or prevent them by taking certain measures. Please allow for possible delays. A gradient boosting classifier model is used to achieve a maximum accuracy of 85.73% through hyperparameter training and tuning. Such intelligent systems are very important for predicting flight accuracy.

[2] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning," IEEE Transactions on Vehicular Technology, vol. 69, No. 1, pp. 140-150, January 2020.

Accurately predicting flight delays is critical to building a more efficient airline business. Recent research has focused on the application of machine learning techniques to predict flight delays. Most of the previous forecasting methods are performed on a single route or airport. This white paper examines the various factors that can affect flight delays and compares several machine learning-based models for common designed flight delay prediction tasks. Automatic Dependent Surveillance Broadcast (ADS-B) messages are received, pre-processed and integrated with other information such as weather conditions, flight schedules and airport information to create the dataset for the proposed scheme. The proposed prediction tasks include various classification and regression tasks. Experimental results show that long-short-term memory (LSTM) can process captured aerial sequence data, but suffers from overfitting problems for limited datasets. Compared with previous schemes, the proposed random forest-based model can achieve higher prediction accuracy (90.2% for binary classification) and overcome the overfitting problem.

[3] Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5.

With the development of computer technology and computer network technology, the amount of data in the information industry is increasing. We need to analyze this huge amount of data and gain useful knowledge. The process of extracting useful knowledge from large amounts of incomplete, noisy, ambiguous and random data is called data mining. The decision tree classification technique is one of the most popular data mining techniques. Decision trees use divide and rule as their basic learning strategy. A decision tree is a structure that contains a root node, branches, and leaf nodes. Each internal node provides a test for an attribute, each branch provides a test result, and each leaf node contains a class specification. The top node of the tree is the root node. This white

paper focuses on various decision tree algorithms (ID3, C4.5, CART), their characteristics, challenges, strengths and weaknesses.

[4] Friedman, Jerome. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis. 38. 367-378. 10.1016/S0167-9473(01)00065-2.

Gradient boosting creates an additive regression model by fitting a simple parameterized function (the base learner) to the current "pseudo" residuals sequentially with least-squares at each iteration. The pseudo residual is the slope of the loss function minimized with respect to the model values at each training data point evaluated at the current step. Integrating randomization into the method has been shown to significantly improve both the approximation accuracy and execution speed of the gradient increase. Specifically, at each iteration, a subsample of the training data is randomly (without permutation) drawn from the full training data set. This random subsample is used instead of the full sample to fit the base learner and compute the model update for the current iteration. This randomized approach will also be more robust against elementary school overcrowding.

EXPERMENTAL

PyCharm:

PyCharm is a purpose-built Python integrated development environment (IDE) that provides a wide range of essential tools for Python developers, tightly coupled to create a working environment for productive Python, web, and data science development.

Scikit-Learn:

Scikit Learn comes with a clean and concise API. It also provides very useful documentation for beginners. There are various algorithms for classification, clustering, regression, etc. It also supports random forests, k-Means, gradient boosting, DBSCAN and more.

Matplotlib:

Matplotlib can produce high quality illustrations suitable for publication. Figures created with Matplotlib are available in paper form on various interactive platforms. MatPlotlib can be used in a variety of toolkits, including Python scripting, IPython shell, Jupyter Notebook, and many other four graphical user interfaces.

Flask:

Flask is used to develop web applications using Python implemented in Werkzeug and Jinja 2**.**

PROPOSED SYSTEM

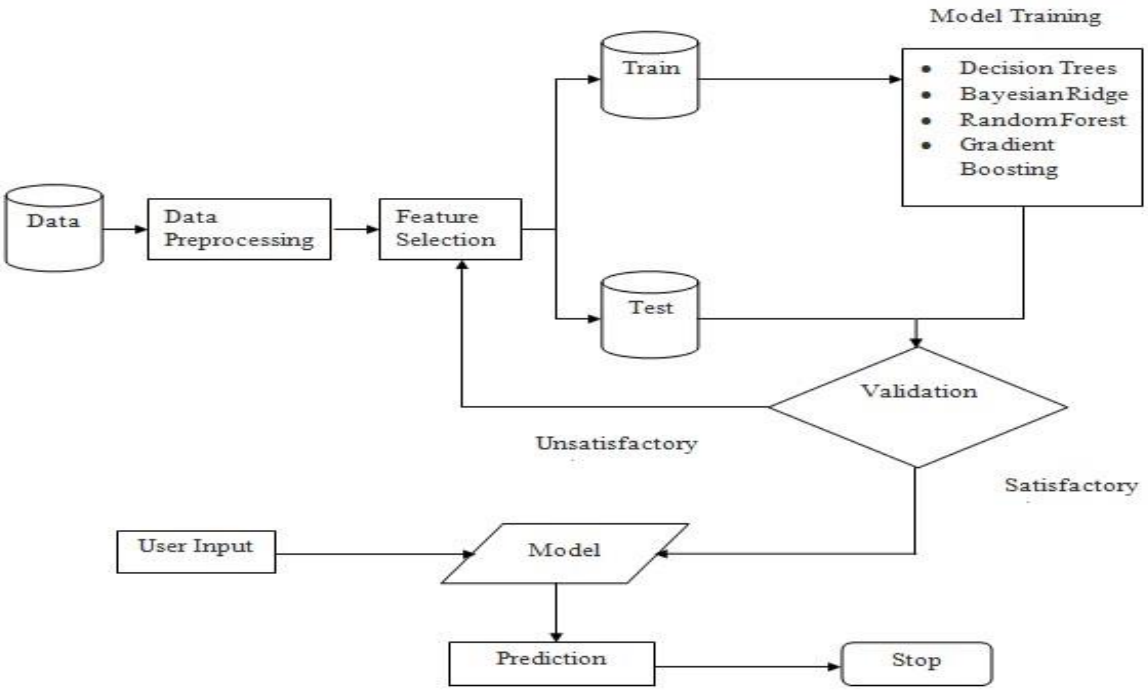
Accurately predicting flight delays is fundamental to building a more efficient airline business. An airline's core business is customer satisfaction. Existing methods are expensive to implement and require highly skilled personnel and manual prediction feature selection. In this post, we use machine learning models such as decision

tree regression, Bayesian Ridge, random forest regression, and gradient boosting regression to predict whether a particular flight will arrive late.

ADVANTAGES:

- Higher Accuracy.
- Low variance.
- Does not require highly trained staff to operate.
- Cheaper to implement.

SYSTEM ARCHITECTURE:



IMPLEMENTATION

MODULES

- **System**
- **User**

1. System:

1.1 Takes Dataset:

The system retrieves the .csv data uploaded by the user.

1.2 Preprocessing:

Data preprocessing is a technique used to transform raw data into a clean dataset. Cleaning data means removing null values,filling null values with meaningful values**,** removing duplicate values, removing outliers, and removing unwanted attributes. If your data set contains categorical data sets, this means converting these categorical variables to numeric values.

Here we remove rows with null values and select features.

1.3 Encoding Dataset:

The dataset is divided into predictor variables and target variables**.** Perform ordinal encoding on the predictors and scale with StandardScaler.

1.4 Model Training:

When the user selects a model**,** the data is split into two parts, a test data set and a training data set.

The models:

- Decision Tree:

Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to build a model that predicts the value of the target variable by learning simple decision rules derived from data properties.

- Bayesian Ridge Regression:

By formulating linear regression using probability distributions instead of point estimates, Bayesian regression allows the natural mechanism to tolerate bad or improperly distributed data. One of the most useful types of Bayesian regression is Bayesian ridge regression, which estimates probabilistic models for regression problems.

- Random Forest Regression:

Random Forest or Random Decision Forest performs classification, regression, and other tasks that work by building different decision trees during training and outputting bins or mean/median prediction modes. It's targeted. Ensemble learning method of individual trees. The random decision forest fixes the habit of the decision tree to overfill the training set. Random forests are generally better than decision trees, but not as accurate as gradient-enhanced trees. However, data properties can affect performance.

- Gradient Boosting Regression:

Gradient boosting is a machine learning technique for regression and classification problems that generates a predictive model in the form of a set of weak predictive models. As with other boosting methods, we build the model step by step and generalize the model by allowing us to optimize the differentiable loss function.

1.5 Tabulation:

The system takes input from the user about a target variable and displays all five-scoring metrics for all algorithms for that target variable.

1.6 Visualization:

The system takes a specific evaluation metric as user input and generates an aggregated bar chart of all different models with two different target variables.

2. User

2.1 Upload Data:

A user uploads a .csv record from a web application containing FAA flight records.

2.2 View Data:

The user checks the data in the web application after the data is cleaned.

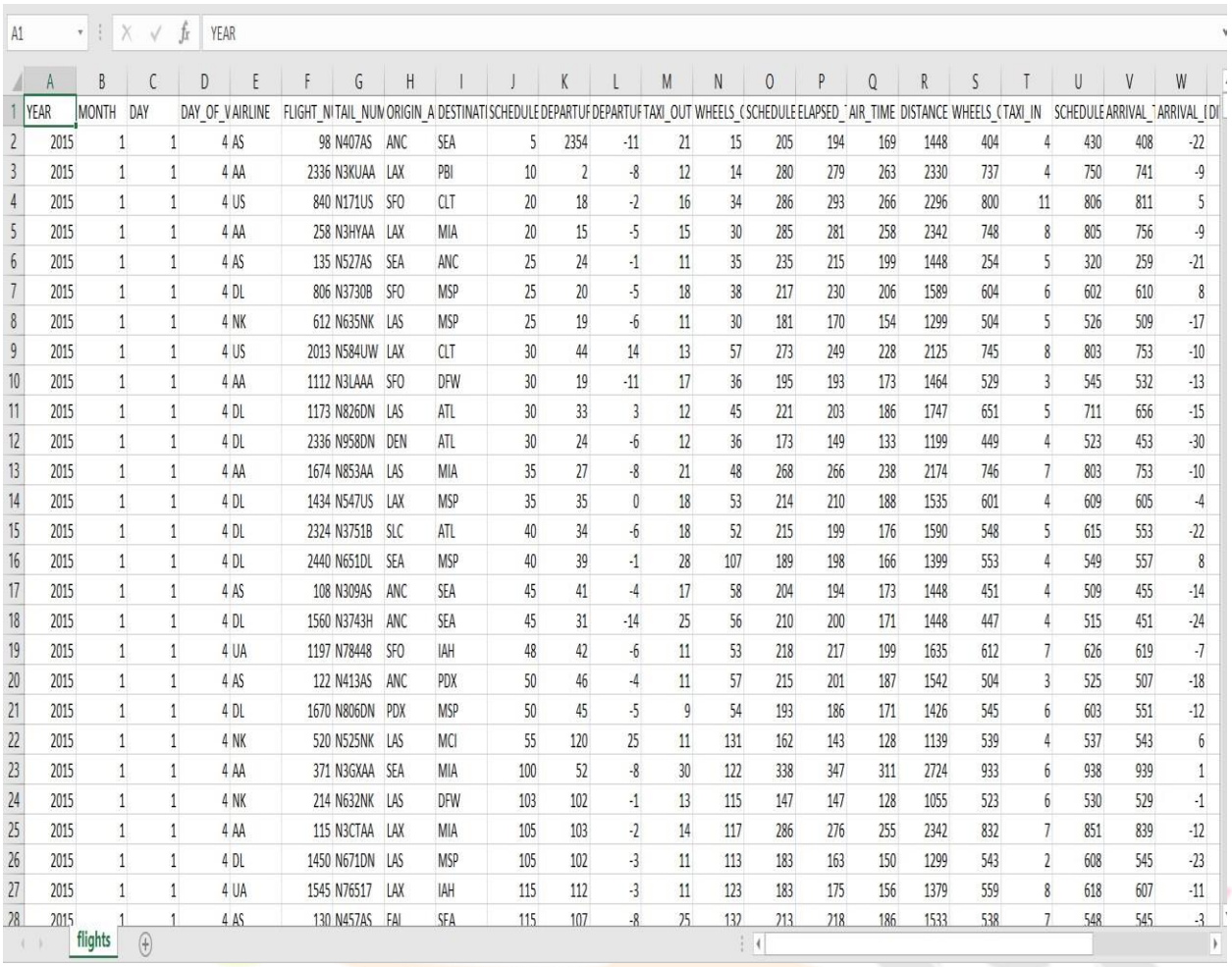
2.3 Model Testing:

The user tests all possible models trained by the system using the test data set and displays all five-evaluation metrics in a table.

2.4 Visualization:

Users can select a specific metric to view a comparison chart of all algorithms in a clustered bar chart using both target variables.

INPUT:



YEAR	MONTH	DAY	DAY_OF_W	AIRLINE	FLIGHT_NI	TAIL_NUM	ORIGIN_A	DESTINATI	SCHEDULE	DEPARTUF	DEPARTUF	TAXI_OUT	WHEELS_(	SCHEDULE	ELAPSED_	AIR_TIME	DISTANCE	WHEELS_(	TAXI_IN	SCHEDULE	ARRIVAL_	ARRIVAL_[DI
2015	1	1	4	AS	98	N407AS	ANC	SEA	5	2354	-11	21	15	205	194	169	1448	404	4	430	408	-22
2015	1	1	4	AA	2336	N3KUAA	LAX	PBI	10	2	-8	12	14	280	279	263	2330	737	4	750	741	-9
2015	1	1	4	US	840	N171US	SFO	CLT	20	18	-2	16	34	286	293	266	2296	800	11	806	811	5
2015	1	1	4	AA	258	N3HYAA	LAX	MIA	20	15	-5	15	30	285	281	258	2342	748	8	805	756	-9
2015	1	1	4	AS	135	N527AS	SEA	ANC	25	24	-1	11	35	235	215	199	1448	254	5	320	259	-21
2015	1	1	4	DL	806	N3730B	SFO	MSP	25	20	-5	18	38	217	230	206	1589	604	6	602	610	8
2015	1	1	4	NK	612	N635NK	LAS	MSP	25	19	-6	11	30	181	170	154	1299	504	5	526	509	-17
2015	1	1	4	US	2013	N584UW	LAX	CLT	30	44	14	13	57	273	249	228	2125	745	8	803	753	-10
2015	1	1	4	AA	1112	N3LAAA	SFO	DFW	30	19	-11	17	36	195	193	173	1464	529	3	545	532	-13
2015	1	1	4	DL	1173	N826DN	LAS	ATL	30	33	3	12	45	221	203	186	1747	651	5	711	656	-15
2015	1	1	4	DL	2336	N958DN	DEN	ATL	30	24	-6	12	36	173	149	133	1199	449	4	523	453	-30
2015	1	1	4	AA	1674	N853AA	LAS	MIA	35	27	-8	21	48	268	266	238	2174	746	7	803	753	-10
2015	1	1	4	DL	1434	N547US	LAX	MSP	35	35	0	18	53	214	210	188	1535	601	4	609	605	-4
2015	1	1	4	DL	2324	N3751B	SLC	ATL	40	34	-6	18	52	215	199	176	1590	548	5	615	553	-22
2015	1	1	4	DL	2440	N651DL	SEA	MSP	40	39	-1	28	107	189	198	166	1399	553	4	549	557	8
2015	1	1	4	AS	108	N309AS	ANC	SEA	45	41	-4	17	58	204	194	173	1448	451	4	509	455	-14
2015	1	1	4	DL	1560	N3743H	ANC	SEA	45	31	-14	25	56	210	200	171	1448	447	4	515	451	-24
2015	1	1	4	UA	1197	N78448	SFO	IAH	48	42	-6	11	53	218	217	199	1635	612	7	626	619	-7
2015	1	1	4	AS	122	N413AS	ANC	PDX	50	46	-4	11	57	215	201	187	1542	504	3	525	507	-18
2015	1	1	4	DL	1670	N806DN	PDX	MSP	50	45	-5	9	54	193	186	171	1426	545	6	603	551	-12
2015	1	1	4	NK	520	N525NK	LAS	MCI	55	120	25	11	131	162	143	128	1139	539	4	537	543	6
2015	1	1	4	AA	371	N3GXAA	SEA	MIA	100	52	-8	30	122	338	347	311	2724	933	6	938	939	1
2015	1	1	4	NK	214	N632NK	LAS	DFW	103	102	-1	13	115	147	147	128	1055	523	6	530	529	-1
2015	1	1	4	AA	115	N3CTAA	LAX	MIA	105	103	-2	14	117	286	276	255	2342	832	7	851	839	-12
2015	1	1	4	DL	1450	N671DN	LAS	MSP	105	102	-3	11	113	183	163	150	1299	543	2	608	545	-23
2015	1	1	4	UA	1545	N76517	LAX	IAH	115	112	-3	11	123	183	175	156	1379	559	8	618	607	-11
2015	1	1	4	AS	130	N457AS	FAI	SEA	115	107	-8	25	132	213	218	186	1533	538	7	548	545	-3

OUTPUT:

The performance for all the models are shown below:

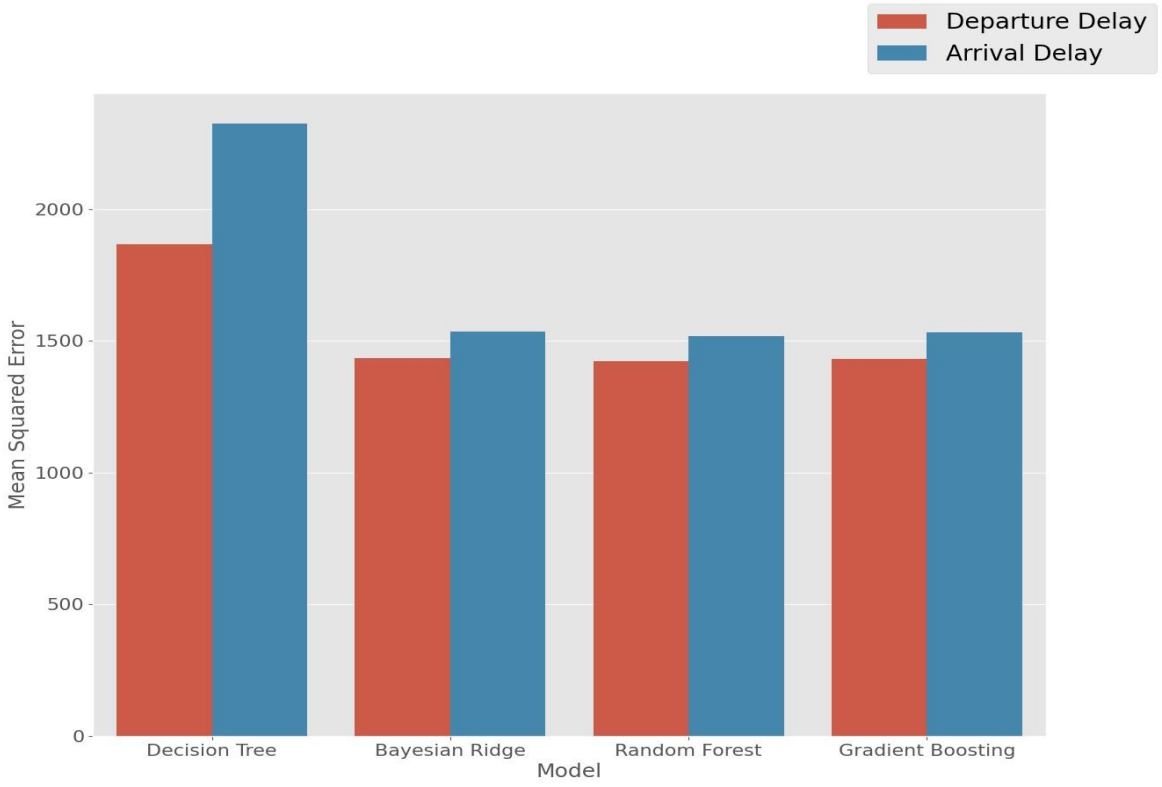
For Departure Delay as Target variable:

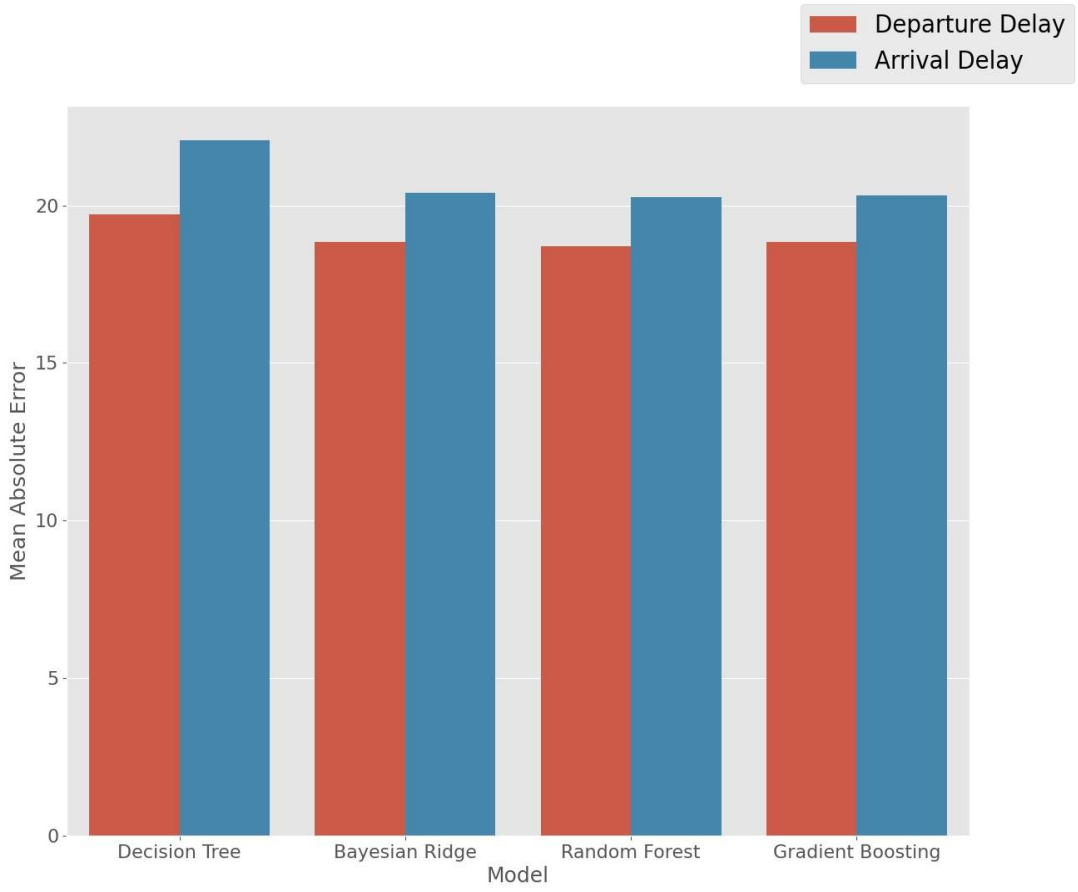
Model	Mean Squared Error	Mean Absolute Error	Explained Variance Score	Median Absolute Error	R2 Score
Decision Tree	1625.961592640178	19.390195853784427	-0.19551877082259206	11.88268156424581	-0.1955942753654134
Bayesian Ridge	1352.0079303838831	18.63101770739971	0.005849323425994757	12.53529895889582	0.005848016870513506
Random Forest	1342.596315162799	18.49396774481932	0.012770013316034001	12.034925708112553	0.012768520608857 47
Gradient Boosting	1348.5314250934398	18.613158798459324	0.008404906668856 96	12.552908284299095	0.008404344057050928

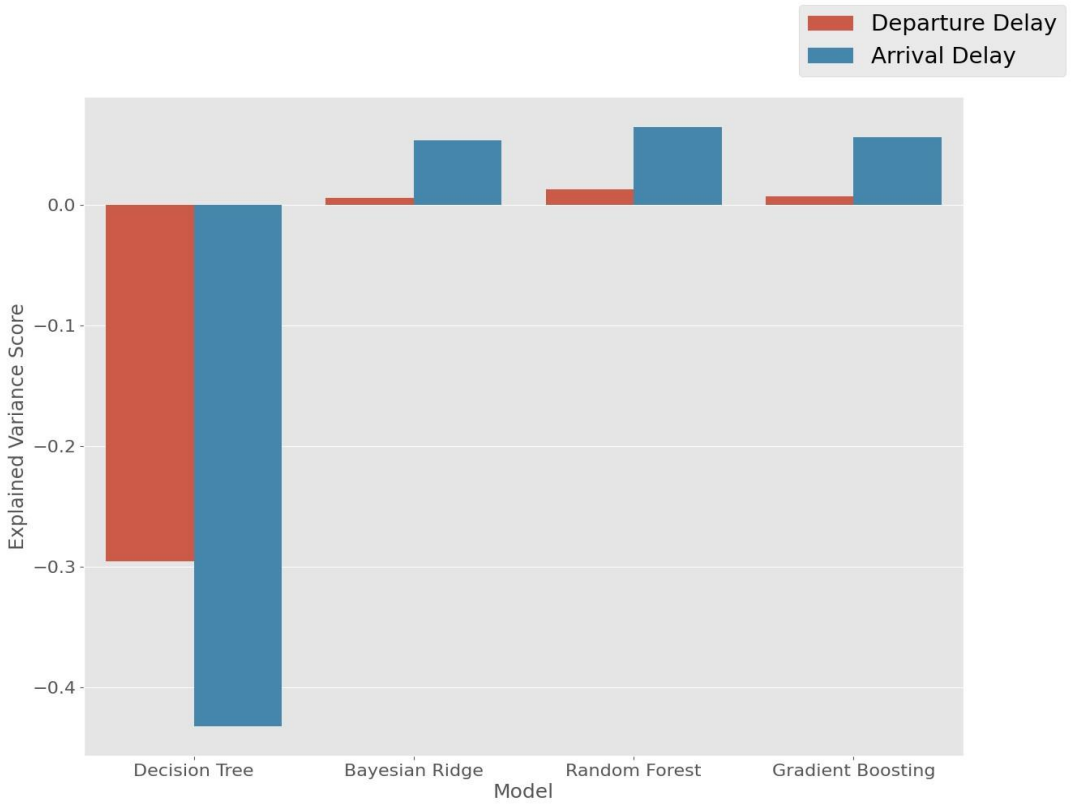
For Arrival Delay as Target variable:

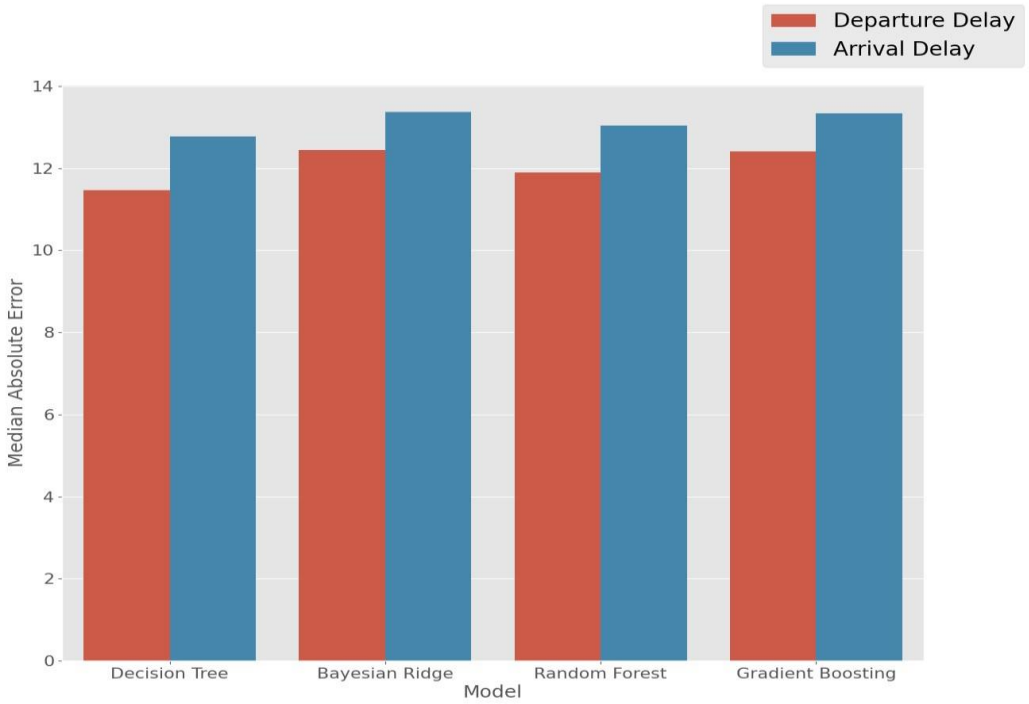
Model	Mean Squared Error	Mean Absolute Error	Explained Variance Score	Median Absolute Error	R2 Score
Decision Tree	1934.1489762920787	21.691321102954923	-0.2587278048772328	12.80952380952381	-0.25882654172691844
Bayesian Ridge	1458.8621297986604	20.1261724366214	0.05052009582036843	13.217917760299745	0.050510383522035074
Random Forest	1442.8551111286981	19.99601108204287	0.0609303963709088	13.04923119647298	0.06092843311524487
Gradient Boosting	1453.832986535228	20.046047618960916	0.05379177922589573	13.25701919388393	0.05378356418172381

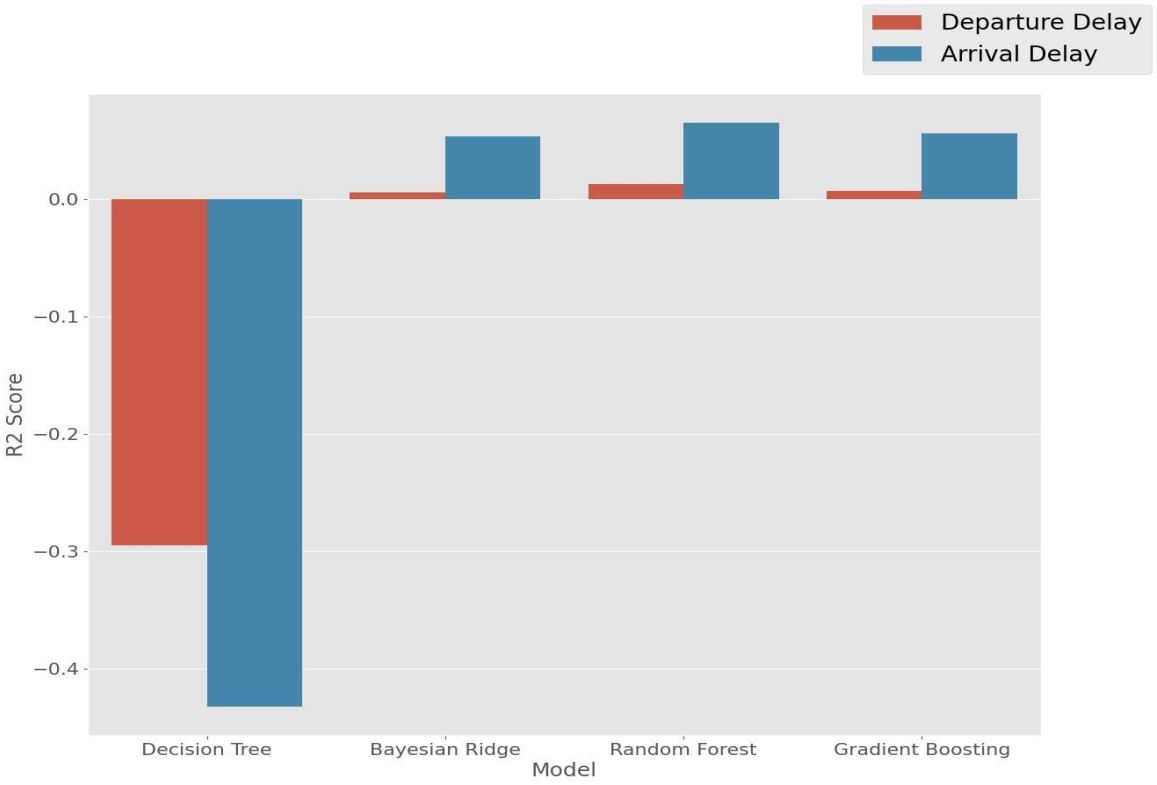
VISUALIZATION:




Departure Delay
Arrival Delay
Mean Absolute Error
0
5
10
15
20
Decision Tree
Bayesian Ridge
Random Forest
Gradient Boosting
Model


Departure Delay
Arrival Delay
Explained Variance Score
0.0
−0.1
−0.2
−0.3
−0.4
Decision Tree
Bayesian Ridge
Random Forest
Gradient Boosting
Model





CONCLUSION

In this application, we removed null values, coded all variables, and preprocessed the data. We also scaled all predictors. We used decision trees**,** Bayesian ridge, random forest**,** and gradient-boosted regression. The best

model was a random forest model with hyperparameter tuning (low margin). The dataset used was the 2015 FAA Flight dataset.

REFERENCE

Chakravarti, Navoniru. (2019). A data mining approach for predicting American Airlines flight arrival delays.

[1] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning," IEEE Transactions on Vehicular Technology, vol. 69, No. 1, pp. 140-150, Jan. 2020.

[2] Sharma, Himani & Kumar, Sunil. (2016). Investigating classification decision tree algorithms in data mining. International Journal of Scientific Research (IJSR). Five.

[3] Friedman, Jerome. (2002). Stochastic gradient amplification. Computational statistics and data analysis. 38 367-378. 10.1016/S0167-9473(01)00065-2.

[4] N. G. Rupp, "Further Investigation into Causes of Flight Delays," Department of Economics, East Carolina University, 2007.

[5] Navoneel, et al., Chakrabarty, Flight Arrival Delay Prediction Using Gradient Boosting Classifier", in Emerging Technologies in Data Mining and Information Security, Singapur, 2019.

[6] A. M. Kalliguddi, Area K., Leboulluec, "Predictive modeling of aircraft flight delays," Universal Journal of Management, S. 485–491, 2017.

[7] Noriko Eya, "Development of on-time arrival combat prediction model for passenger aircraft by discovering correlation between combat data and meteorological data"', 2019.

[8] Chakravarti, Navoniru. "A Data Mining Approach for Predicting Flight Arrival Delays for American Airlines." 2019 9th Annual Information Technology, Electromechanical and Microelectronics (IEMECON) Conference (2019): 102-107.

[9] Sternberg, Alice & Soares, Jorge & Carvalho, Diego & Ogasawara, Eduardo. (2017). Review flight delay predictions.

[10] V. Venkatesh、A. Arya、P. Agarwal、S. Lakshmi、S. Balana, "Iterative Machine and Deep Learning Approaches for Aviation Delay Prediction," 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, 2017, S. 562-567,doi:10.1109 /UPCON .2017.8251111.

[11] Yogita Borse, Dhruvin Jain, Shreyash Sharma, Viral Vora, Aakash Zaveri, 2020, Flight Delay Prediction System, International Journal of Engineering Research & Technology (IJERT) Band 09, Ausgabe 03 (März 2020).