Understanding Flight Delays

Abstract

There has been an significant increase in flight delay in recent times due to increase in the demand for air transportation. This delay has caused severe loss to the world in terms of resources and economy. In order to address this issue, we have come with this study, where we analyse different factors leading to flight delay and build a hybrid model to predict the same. The hybrid model is combination of a Random Forest classifier and XGBoost regression. This hybrid model gave a root mean square error of 9.2881 and outperformed regression models (Linear Regression, Random Forest, XGBoost). Our analysis and proposed approach will help predict flight delay beforehand and potentially mitigate the inconvenience and damage caused due to it.

1 Introduction

With the rapid development of the national economy, the demand for air transportation has skyrocketed. Flight delays are becoming increasingly severe, causing direct economic damage to passengers, airlines, and airports. According to data reported by carriers filing on-time performance, approximately 11.74 % of flights in the United States were delayed in November 2017. These delays cost the economy billions of dollars. The high cost of flight delays motivates the study and prediction of flight delays.

Given the need to analyze and predict air traffic delays, we have investigated various factors affecting flight delays, such as weather, maintenance, security, and so on. In addition, we have developed a hybrid prediction model to accurately predict delay time at a specific airport after a certain time period.

With the advent of state-of-art machine learning techniques, there has increased amount of research on flight delay prediction based on machine learning approaches.

(Baluch et al., 2017) studied various data mining techniques such as clustering, classification, and decision tree to provide answers to customer-related questions. For example, they responded to questions such as which airline, airport, and time is best to avoid delays, and so on. The study made use of a dataset containing 25,552,386 documented flights between July 2012 and July 2016. The only pre-processing required was the removal of commas from the city and state fields in the dataset they used.

(Ding, 2017) proposed a method for accurately forecasting the complex problem of flight delay using a multiple linear regression algorithm The results were compared to the Naive-Bayes and C4.5 approaches. They used 100,000 data records from November 3, 2015 to March 5, 2016. This dataset was further combined with secondary data sources to provide more information. This study considered not only predicting delay time but also categorizing whether the delay is above a certain threshold or not. The model outperformed the Naive-Bayes and C4.5 approaches.

(Rebollo and Balakrishnan, 2014) in their research used Random Forest algorithms to predict departure delays after a certain time frame in the future. They considered both spatial and temporal delay states as variables. Their primary goal is to predict global delays across the national air space at the time of prediction, rather than delays at a specific airport or airline. They gathered information from the Aviation System Performance Metrics (ASPM) database between January 2007 and December 2008. They also considered two types of prediction mechanisms, namely classification and prediction. The test error for classification of 100 links with a threshold of 60 minutes was 19 % for a 2hr time frame, and the median test error for predicting departure delays was 21 minutes.

(Ariyawansa and Aponso, 2016) reviewed various data mining techniques that can be applied

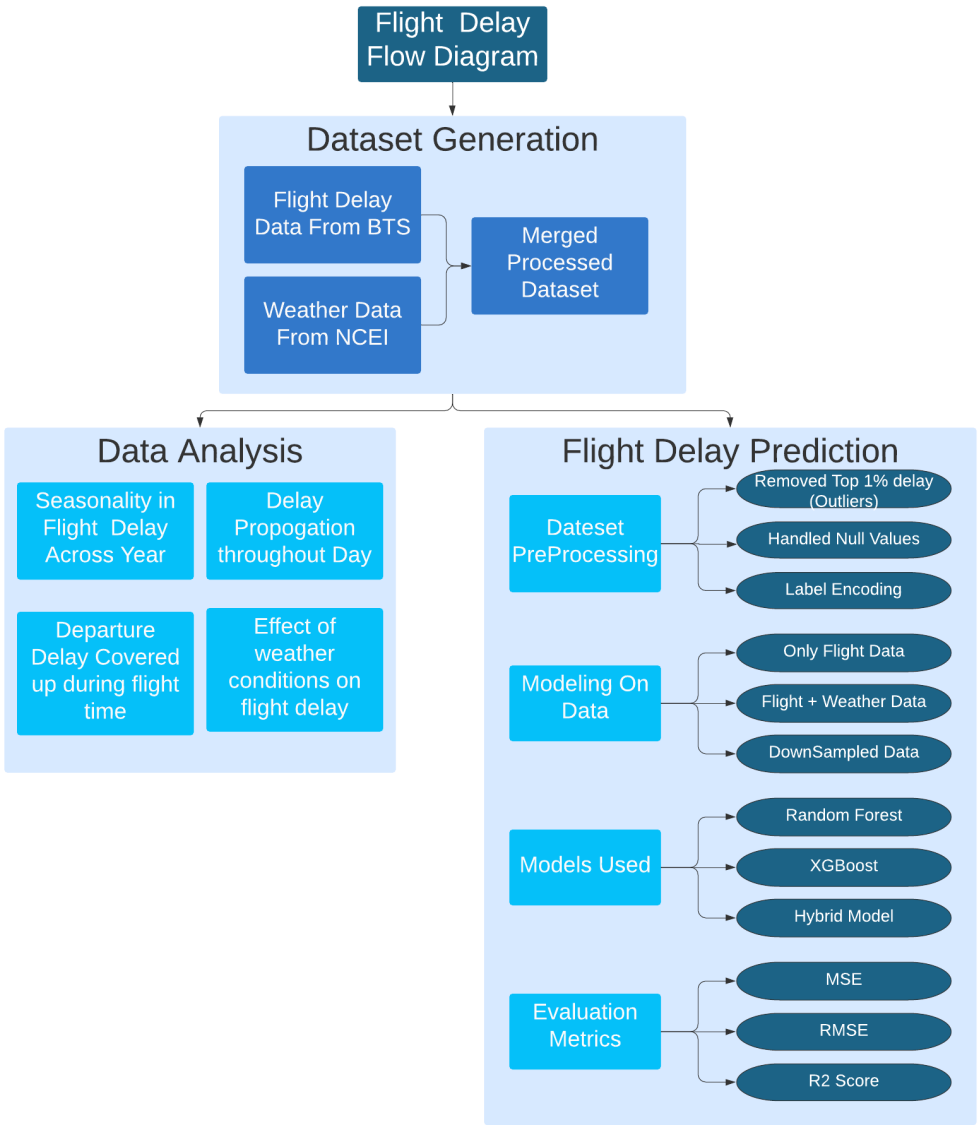

Figure 1: Flight Delay Flow Diagram

to improve airport systems. Flight delay prediction was also reviewed in this paper and various machine learning models were compared. Dataset for this study was collected from publicly available US department of transportation statistics. 80,000 records were considered for this study which was obtained by merging weather data. The study showed the Random Forest algorithm to outperform others in terms of accuracy.

(Sternberg et al., 2017) provided a thorough review of flight delay predictions from a data science standpoint. The manuscript clearly described the operations of a commercial flight as well as the problems with flight delay predictions. It went on to explain various studies on delay propagation, root delay, and cancellation domain problems in flight forecasting. The study looked at various approaches for predicting flight delays and discovered that machine learning approaches were the most popular in recent years. The most popular dataset for obtaining flight information came from the United States Department of Transportation, specifically the Federal Aviation Administration and the Bureau of Transportation Statistics databases.

Recently, (Yu et al., 2019) came up with a deep learning approach for flight delay predictions. The study analyzed high dimensional data from Beijing International Airport and focused on micro-influential factors that directly affect flight delay at the operational level. They used the DBN-SVR model, which consists of a deep belief network with a support vector regressor embedded in the developed model to perform supervised fine-tuning within the architecture. The dataset used was not

available publicly and the study was collaboration research with Beijing PEK airport. The results show DBN-SVR to outperform machine learning models like k-NN, SVM, and Linear Regressor in terms of RMSE and MAE.

(Shao et al., 2021) presented a novel end-to-end deep learning strategy that predicts flight delays using both spatial and temporal information. TrajCNN was utilized, and it had an error of 18 minutes in predicting aircraft delays at Los Angeles International Airport.

The rest of this paper is organized as follows: Section 2 summarizes the data generation process Section 3 presents various analysis performed in this study. Section 4 explains the methodology in detail and Section 5 discusses the results obtained in the study.

2 Dataset

The first primary task of finding relevant data for flights and weather is accomplished by taking flight delay data from United States Department of Transportation namely Bureau of Transportation Statistics and weather data from National Centers for Environmental Information.

Dataset Generation Process -

- Firstly, the month-wise data consisting of 69 features related to flight delay causes for the year 2018 and 2019 was collected from the Bureau of Transportation Statistics.
- The data from each month was then combined to create flight_delay_2018, which contains flight delay data for the year 2018 and flight_delay_2019, which contains flight delay data for the year 2019. Next, preprocessing of that data was done.
- Weather data for all US airports for the year 2018 and 2019 was collected from the National Centers for Environmental Information, which only had eight columns. Because the data was not provided in the proper format, considerable preprocessing was performed to turn the information into 21 features that could cause the aircraft to be delayed. Some features include snow, wind speed, fog, thunder, smoke, temperature, etc.
- Finally, we combined preprocessed flight delay and weather data to create our final dataset consisting of 4713650 rows and 58 columns. This dataset was used for analysis and prediction.

Dataset Preprocessing -

- To begin, the 30 busiest airports in the United States were identified through visual analysis, and only data on flight delays for these airports were included in the dataset. In addition, station-Ids for each of these airports were discovered from the NOAA website to compress weather data to just include information for these airports.
- The biggest problem with weather data was that it didn't have columns directly for TMAX, TMIN, Precipitation, Snowfall, Snow Depth, Wind Speed, Fog, Smoke or Haze, Thunder, and many other features. Instead, there were two columns along with station-Id (Airport) and date depicting all these weather conditions named element type (which includes feature name) and element value (which contains feature value). Following that, these two columns of element type and element name were transformed to 15 columns, with column name being element type and data in the rows being element value, resulting in unique rows of station-Id, year, month, and day containing all weather information.
- Finally, in the primary dataset, weather data for both departure airport and arrival airport was merged by first doing merge operation converting station-Id to Origin airport name and then doing merge operation converting station-Id to destination airport name.

3 Analysis

3.1 Is there any seasonality in flight delay across the year? Does flight delay increase during particular months?

As seen in Figure 2, there is an increase in flight delay time beginning in May and continuing through the end of August. This makes sense because air traffic increases over the summer months, often from Memorial Day to Labor Day. Many Americans use these months to spend time with their families, and many of us enjoy visiting the same places.

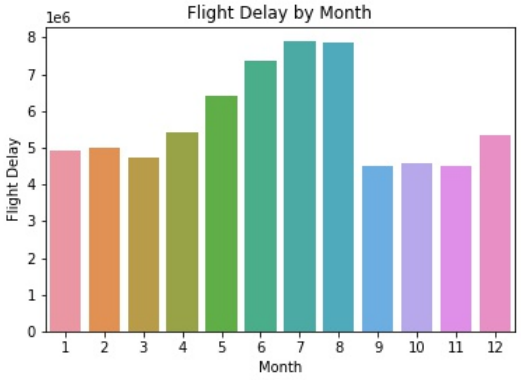

Figure 2: Monthly trend on flight delay. There is an sudden increase in flight delay from May and this trend continues till end of August.

3.2 Does flight delay propagate throughout the day and how does flight traffic vary across a day?

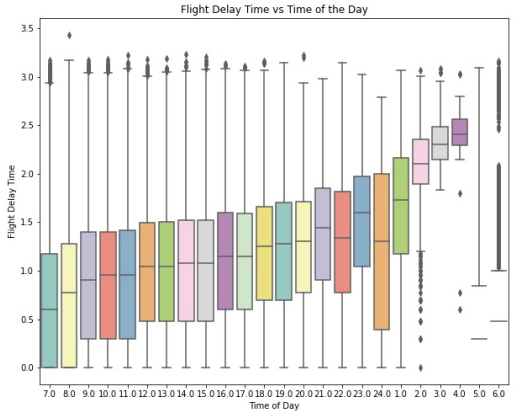

Figure 3: Delay propagation along the day. The delay occurs early in the morning at around 7 am, typically it accumulates and propagates forward, landing on the last flight of the day.

From Figure 3, we can see that the delay that usually starts early in the morning when the air traffic is high continues to propagate till the end of the day, even if the traffic reduces and finally the delay starts decreasing after 3-4 am. The reason for this could be that most airlines schedule relatively tightly in order to get the greatest use of aircraft and flight crew and once a delay is introduced it is very difficult for it to get mitigated. Figure 4 shows that the no. of flights is maximum early in the day and minimum late at night, therefore the delay propagated along the entire day starts to dissipate after mid night.

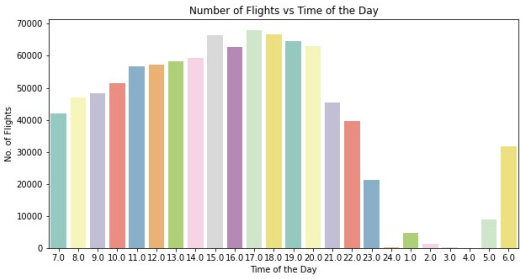

Figure 4: Air traffic along the day. Airlines are tightly scheduled early in the morning and the traffic decreases late after midnight.

Figure 5: Shows what percentage of delays at departure is propagated or reduced during fly time.

3.3 What percentage of Departure Delays gets covered up during fly time?

As shown in Figure 5, almost 52 % data has no delay either at the time of arrival or departure. 11 % of the flights could successfully mitigate the delay completely during the flight time and 15 % could do it partially. There are around 21 % of flights in 2018-19 where the delay has increased during the flight time or at the destination airport.

3.4 How different weather conditions affect flight delay?

Figure 6 explains how the bad weather condition could cause departure delays as well as arrival delays. In figure 6, plots-1 and plots-2 show how the mean flight delay minutes increase when there is a fog or thunder. Plot-3 between flight delay minutes and snow depicts how flight delay increases significantly when there is more than 50mm of snowfall on a single day. When the wind speed exceeds a specific threshold, as shown in plot-4, it can also cause flight delays.

4 Methodology

4.1 Data Preprocessing

As our dataset has 58 features and 4713650 rows consisting flight data of 2018 and 2019 year of 30 busiest airports of United States. It becomes nec-

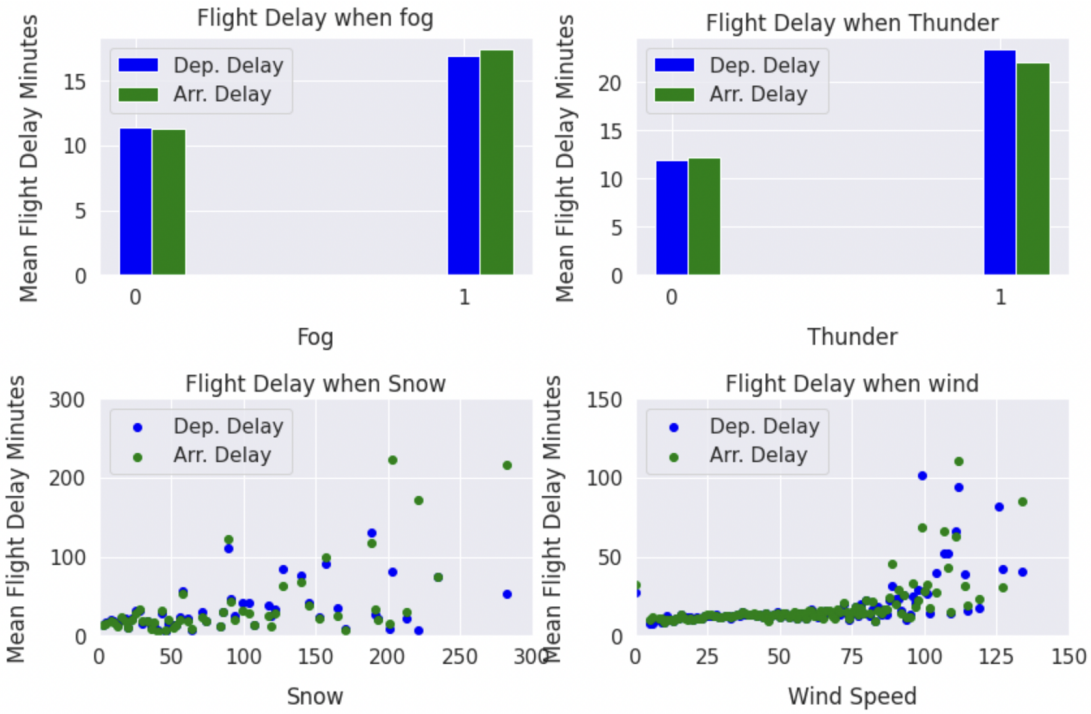

Figure 6: Effect of weather on flight delay

essary to extract important features affecting flight delay from that. First step was to remove top 1% of the highest delays as they would be outliers. When the data was analysed some flight delays were extreme so we decided to remove top 1% of the flight delays to get better predictions. Next step was to remove features which were of less significance to cause flight arrival delay. In this huge dataset, around 40000 rows were containing Null values of arrival delay and departure delay. So, we just removed that rows. There are categorical features like OP_UNIQUE_CARRIER, TAIL_NUM, ORIGIN_STATE_NM, DEST_STATE_NM and many other which are required to be handled to train our model. As we can't use categorical values in regression analysis we have converted categorical columns to numeric using a label encoder. Last step was to upsample our dataset because there was class imbalance in the dataset. More than 70% dataset has no delay so we upsampled our data to overcome the problem of class imbalance. For that, we used resample functionality of sklearn.utils library.

Both the approaches explained below uses this preprocessed data.

4.2 Baseline Model

The idea behind our baseline model is to see how well a simple linear regression model can predict the flight delay based only on general features without considering other factors like weather, security, maintenance, etc. Later, we plan to add features based on the factors mentioned above and check how well the model performs with respect to the baseline model.

Features used: Month, Day of month, Unique carrier code, Origin airport ID, Destination airport ID, Departure time, Arrival time, Distance.

Pre-processing: Normalizing features and Removing null values.

Model used: Linear Regression.

Evaluation: Root Mean Square Error (RMSE): 24.62

To predict flight delay more effectively we have come up with two approaches:

1. Predicting flight delay by training the pre-processed dataset on regression models like Linear Regression, XGBoost, Random Forest.

2. Applying an hybrid approach by training the pre-processed dataset on two models - a classifier (Random Forest) and a regression (XGBoost).

Both the approaches will be explained in detail in the next section.

4.3 Approach 1- Regression Models

In this approach we have used additional features based on the analysis as explained in the the analy-

Features

Factor	Features
Time Period	YEAR, MONTH, DAY_OF_MONTH, DEP_TIME, ARR_TIME
Airline	OP_UNIQUE_CARRIER, TAIL_NUM, OP_CARRIER_FL_NUM
Origin Airport	ORIGIN_AIRPORT_ID, ORIGIN_STATE_NM
Destination Airport	Destination_AIRPORT_ID, Destination_STATE_NM
Departure Performance	DEP_DELAY_NEW
Arrival Performance	ARR_DELAY_NEW
Security and Maintenance Indicator	TAXI_OUT, WHEELS_OFF, WHEELS_ON, TAXI_IN
General Flight Summary	AIR_TIME, DISTANCE
Weather at Origin Airport	TMAX, TMIN, SNOW, AVG_WIND_SPEED, FOG, THUNDER, SMOKE
Weather at Destination Airport	TMAX_Dest, TMIN_Dest, SNOW_Dest, AVG_WIND_SPEED_Dest, FOG_Dest, THUNDER_Dest, SMOKE_Dest

Figure 7: Features

sis section on top of the features used in the baseline model.

The regression models used in this approach are:

1. Linear Regression
2. Random Forest
3. XGBoost

After training the model we performed hyperparameter tuning to further decrease the error.

Basically, Hyperparameters are specific values or weights that determine the learning process of an algorithm. As we know, XGBoost has a vast number of hyperparameters to choose from. By fine-tuning XGBoost's hyperparameters we can increase the model's performance.

In our model, we have used the parameters - objective, colsample_bytree, learning_rate, max_depth, alpha, n_estimators. GridSearchCV function comes in Scikit learn python library. It helps us in finding the best values of our parameters and fit our model on training set. Hence, after using it we can choose the best parameters from the listed hyperparameters to train and evaluate the model perfomance. Here, we have tuned both n_estimators and max_depth parameters using GridSearchCV and we got the Best score of 0.984179 Using parameters 'max_depth': 5, 'n_estimators': 300

Here best pair of hyperparameters gives us the best fit model and lower error values i.e. evaluation metrics.

4.4 Approach 2- Hybrid Model

Though the regression models used in approach 1 performed quite well as compared to the baseline model but the error can be reduced further using this hybrid approach. The main drawback of the regression models was that even if there is no delay i.e. the actual delay value = 0, the model would predict some small value of delay for that sample. Therefore, all the samples which does not have delay used to increase the error of the regression model. To overcome this drawback and further improve the model performance we use a hybrid of classifier and regression model.

The hybrid approach is as follows:

1. Training:

a. Classifier: We trained a random forest classifier on the dataset, where the class is 0 if there is no delay and 1 if there is a delay. We got an F1-Score of 98.3 % for this classifier.

b. Regression: We trained XGBoost model on a custom dataset. This custom dataset has only those

samples where there is delay and all the samples where the delay was 0 originally, were removed.

2. Testing:

a. First we pass the sample to be predicted to the classifier.

b. If the classifier predicts 1 (i.e. delay is present), we pass the sample to the regression model to predict the amount of flight delay.

c. Else if the classifier predicts 0 (i.e. delay is not present) we simply return 0 as the delay value.

5 Results and Discussion

As shown in the Figure 8, both the approaches have performed significantly well as compared to the baseline model. Hence, we can say that the additional features added helped the regression models to learn better. Performing upsampling and removing outliers have also made the model more robust and give more accurate results.

Though adding weather data increased the error but the increase might be due to significant increase in dimension of the dataset. The hybrid model outperfomed the regression models. The combination of Random Forest Classifier and XGBoost gave the best results.

From the regression models used, XGBoost outperformed Random Forest and Linear Regression in terms of all the evaluation metrics used as shown in the table. The main reason to use XGBoost for our study is because of its performance and speed of execution. Performing

The use of hybrid models significantly reduced the error of the regression models when the actual delay value is 0. The hybrid model is a simple but effective approach to predict flight delays. The classification task for this dataset is quite easy for the model to learn, hence we use this advantage to improve the performance of the regression task.

References

Chamath Malinda Ariyawansa and Achala Chathuranga Aponso. 2016. Review on state of art data mining and machine learning techniques for intelligent airport systems. In *2016 2nd International Conference on Information Management (ICIM)*, pages 134–138. IEEE.

Megan Baluch, Tristan Bergstra, and Mohamad El-Hajj. 2017. Complex analysis of united states flight data using a data mining approach. In *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1–6. IEEE.

Yi Ding. 2017. Predicting flight delay based on multiple linear regression. In *IOP conference series: Earth and environmental science*, volume 81, page 012198. IOP Publishing.

Juan Jose Rebollo and Hamsa Balakrishnan. 2014. Characterization and prediction of air traffic delays. *Transportation research part C: Emerging technologies*, 44:231–241.

Wei Shao, Arian Prabowo, Sichen Zhao, Piotr Koniusz, and Flora D Salim. 2021. Predicting flight delay with spatio-temporal trajectory convolutional network and airport situational awareness map. *arXiv preprint arXiv:2105.08969*.

Alice Sternberg, Jorge Soares, Diego Carvalho, and Eduardo Ogasawara. 2017. A review on flight delay prediction. *arXiv preprint arXiv:1703.06118*.

Bin Yu, Zhen Guo, Sobhan Asian, Huaizhu Wang, and Gang Chen. 2019. Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, 125:203–221.

Dataset	Models	Hyperparameters	Evaluation Metrics		
			MSE	RMSE	R2 Score
Flight	Linear Regression		1974.2321	44.4323	0.0835
	Random Forest	n_estimators=100 random_state=42 criterion=gini	570.2411	23.9215	0.7145
	XGBoost	n_estimators = 100	555.5961	23.5711	0.7421
		colsample_bytree = 0.4 learning_rate=0.08 max_depth=5 alpha=10 n_estimators=150	151.6794	12.3158	0.9295
		colsample_bytree = 0.4 learning_rate=0.1 max_depth=5 alpha=10 n_estimators=300	87.1964	9.3379	0.9595
Flight + Weather	Linear Regression		1926.3699	43.8904	0.1057
	Random Forest	n_estimators=100 criterion=gini	560.7382	23.6799	0.7182
	XGBoost	n_estimators = 100	547.4925	23.3985	0.7458
		colsample_bytree = 0.4 learning_rate=0.08 max_depth=5 alpha=10 n_estimators=150	167.0778	12.9258	0.9224
		colsample_bytree = 0.4 learning_rate=0.1 max_depth=5 alpha=10 n_estimators=300	99.6193	9.9809	0.9537
	(Random Forest Classifier + XGBoost)	Random Forest: XGBoost: colsample_bytree = 0.4 learning_rate=0.1 max_depth=5 alpha=10 n_estimators=300	86.2696	9.2881	0.9612

Figure 8: Result Table