Airline Delay Prediction

Mohit Mohit
Abhishek Rai Sharma

Outline

1. Introduction
2. Goal
3. Dataset Description
4. Feature Transformation
5. Data Analysis
6. Data Cleaning
7. Class Imbalance
8. Data Sampling
9. Importing the data into WEKA
10. Data Wrangling
11. Attribute Selection
12. Classification Algorithms
13. Performance
14. Responsible Factors
15. Conclusion
16. Referencess

Introduction

- We got the data from Bureau of Transportation Statistics, U.S. Department of Transportation
- Data from Jan. 2018 to Dec. 2018 were selected to predict the delay.
- This Data consists of 7076405 tuples.
- We chose 30 important attributes.
- Out of these attributes we have selected ‘ARR_DELAY_NEW’ as our class attribute.

Goal

- Our goal is to predict whether a flight will be delayed or not.
- Factors causing the delay.

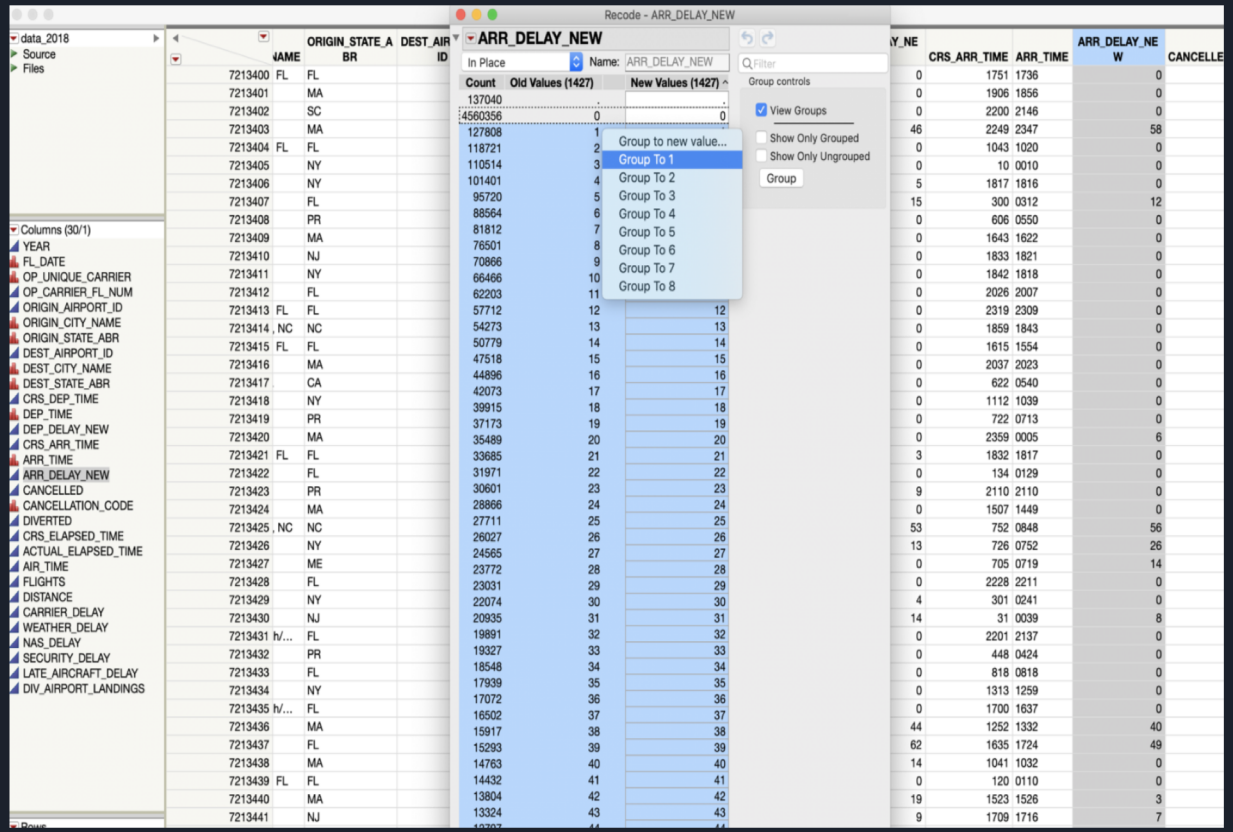
Dataset Description

These 30 attributes were selected while downloading the data as there were numerous irrelevant attributes considering our goal.

SYS_FIELD_NAME	FIELD_DESC
YEAR	Year
FL_DATE	Flight Date (yyyymmdd)
OP_UNIQUE_CARRIER	Unique Carrier Code.
OP_CARRIER_FL_NUM	Flight Number
ORIGIN_AIRPORT_ID	Origin Airport, Airport ID.
ORIGIN_CITY_NAME	Origin Airport, City Name
ORIGIN_STATE_ABR	Origin Airport, State Code
DEST_AIRPORT_ID	Destination Airport, Airport ID.
DEST_CITY_NAME	Destination Airport, City Name
DEST_STATE_ABR	Destination Airport, State Code
CRS_DEP_TIME	CRS Departure Time (local time: hhmm)
DEP_TIME	Actual Departure Time (local time: hhmm)
DEP_DELAY_NEW	Difference in minutes between scheduled and actual departure time. Early departures set to 0.
CRS_ARR_TIME	CRS Arrival Time (local time: hhmm)
ARR_TIME	Actual Arrival Time (local time: hhmm)
ARR_DELAY_NEW	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
CANCELLED	Cancelled Flight Indicator (1=Yes)
CANCELLATION_CODE	Specifies The Reason For Cancellation
DIVERTED	Diverted Flight Indicator (1=Yes)
CRS_ELAPSED_TIME	CRS Elapsed Time of Flight, in Minutes
ACTUAL_ELAPSED_TIME	Elapsed Time of Flight, in Minutes
AIR_TIME	Flight Time, in Minutes
FLIGHTS	Number of Flights
DISTANCE	Distance between airports (miles)
CARRIER_DELAY	Carrier Delay, in Minutes
WEATHER_DELAY	Weather Delay, in Minutes
NAS_DELAY	National Air System Delay, in Minutes
SECURITY_DELAY	Security Delay, in Minutes
LATE_AIRCRAFT_DELAY	Late Aircraft Delay, in Minutes
DIV_AIRPORT_LANDINGS	Number of Diverted Airport Landings

Feature Transformation

To predict the delay, we need to know whether the flight is delayed or not, we are not focusing on the magnitude of the delay. So, we have transformed our class attribute as ‘0’ for on-time flights and ‘1’ for delayed flights.



Data Analysis

In Column Viewer, we can check for missing values, mean, median, mode etc. under show summary.

Summary Statistics

31 Columns | Clear Select | Distribution

Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev
DEP_TIME	6962512	103273	1439	.	.	.	.
DEP_DELAY_NEW	6957831	107954	.	0	2482	13.282041774	44.172155295
CRS_ARR_TIME	7065785	0	.	1	2400	1488.8176675	515.63274887
ARR_TIME	6955393	110392	1440	.	.	.	.
ARR_DELAY_NEW	6937501	128284	.	0	2475	13.506217008	43.912921583
CANCELLED	7065785	0	.	0	1	0.0152094919	0.1223853154
CANCELLATION_CODE	107467	6958318	4	.	.	.	.
DIVERTED	7065785	0	.	0	1	0.0026131562	0.0510522086
CRS_ELAPSED_TIME	7065761	24	.	-99	704	141.48391164	73.38463324
ACTUAL_ELAPSED_TIME	6939855	125930	.	14	739	136.77425393	73.211201271
AIR_TIME	6939855	125930	.	7	696	111.65965096	71.224062508
FLIGHTS	7065785	0	.	1	1	1	0
DISTANCE	7065785	0	.	31	4983	798.92478529	597.08844194
CARRIER_DELAY	1339798	5725987	.	0	2109	19.536672692	59.857137517
WEATHER_DELAY	1339798	5725987	.	0	2475	3.685227176	30.392830167
NAS_DELAY	1339798	5725987	.	0	1848	15.567396727	34.709003121
SECURITY_DELAY	1339798	5725987	.	0	927	0.083306588	2.7971448789
LATE_AIRCRAFT_DELAY	1339798	5725987	.	0	2454	25.661979642	50.349601107
DIV_AIRPORT_LANDINGS	7065785	0	.	0	9	0.0036664008	0.1093153431
Column 31	0	7065785	.	.	.	.	.

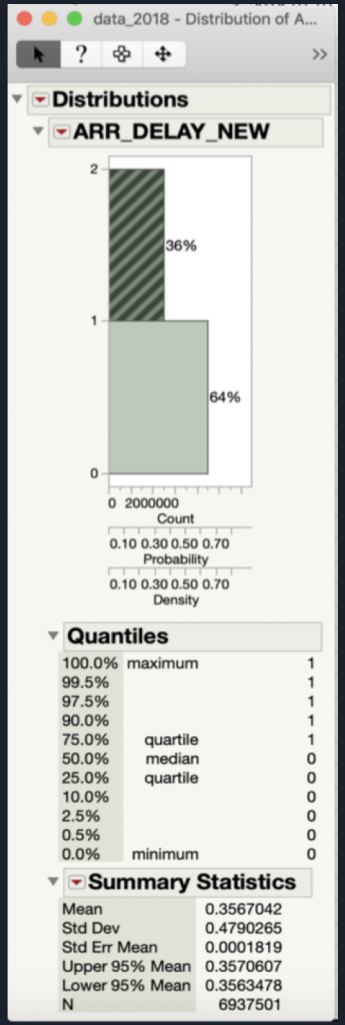
Data Cleaning

- On the basis of the missing values we have decided to delete the tuples where the class attribute is missing.
- After analyzing the dataset, we found that columns DEP_DELAY_NEW, CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, SECURITY_DELAY and LATE_AIRCRAFT_DELAY have a lot of missing values.
- For the missing values of DEP_DELAY_NEW, we will replace them by their mean because we can’t set them to be 0 or something else, as they represent the magnitude of the delay.
- For the remaining columns, the tuples with missing values had no delay. Hence, we will replace them with 0.

Class Imbalance

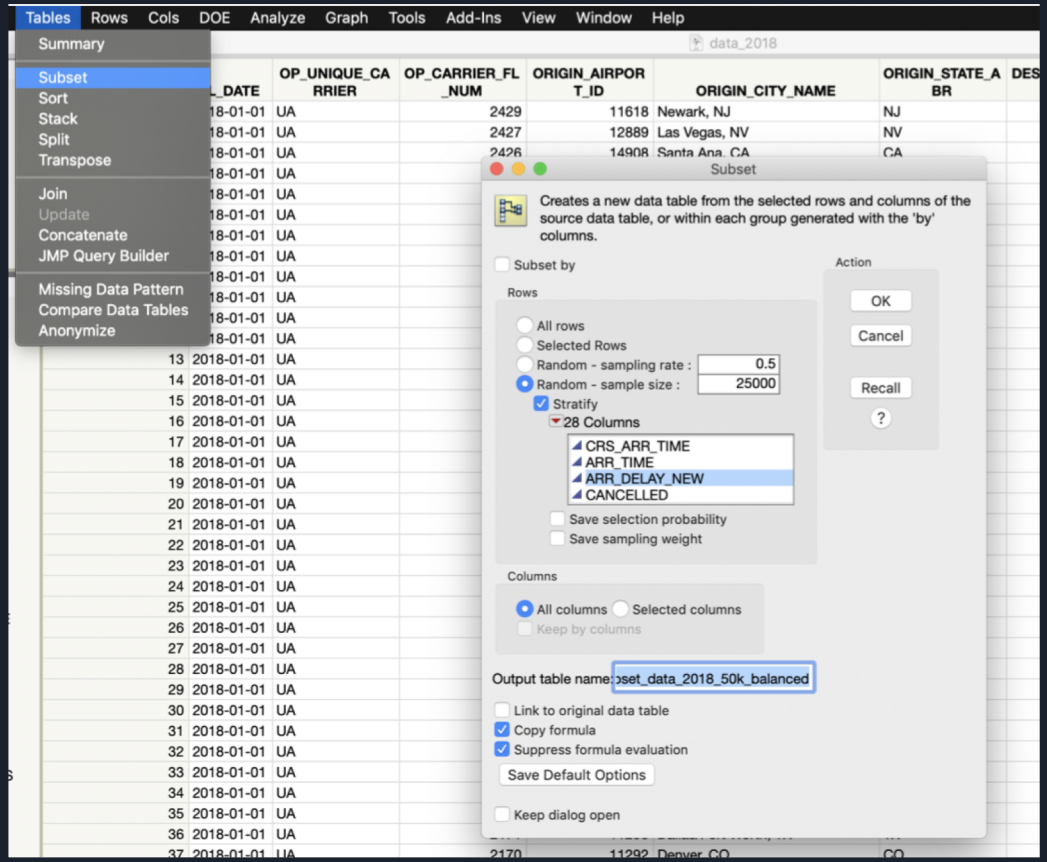
We have class imbalance in the data.

- 36% of the data represents delayed.
- 64% of the data represents on-time.



Data Sampling

To balance and reduce the number of tuples we decided to take a subset of the dataset in JMP Pro. Therefore, we decided to take a sample of 50,000 tuples from the data.

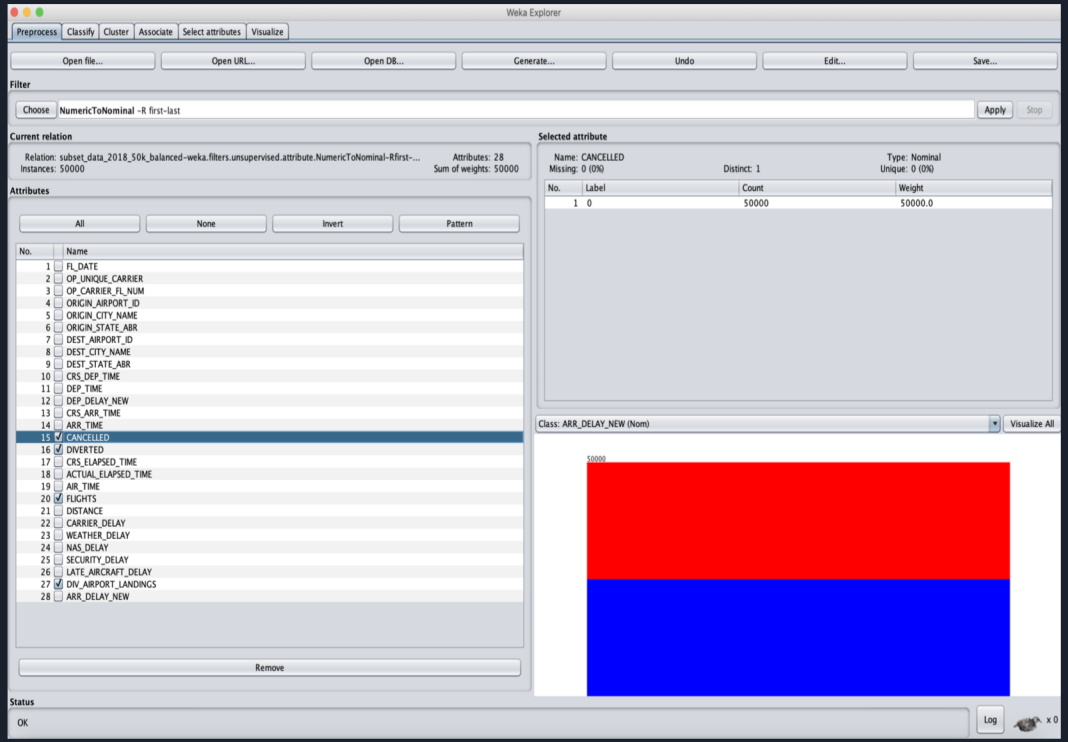


Importing the data into WEKA

- We exported the processed data from JMP Pro into WEKA.
- The csv file doesn’t contain class attribute, so we will make ARR_DELAY_NEW as a class attribute.
- Upon loading ARR_DELAY_NEW was of type numeric, so we converted it to nominal type for classification.
- We saved this file as .arff file, to make it compatible with WEKA.

Data Wrangling

Now after looking through all the attributes we found that the CANCELLED, DIVERTED, FLIGHTS and DIV_AIRPORT_LANDINGS have only single values either ‘0’ or ’1’. Hence, we deleted them.



Attribute Selection

- **OneR Attribute Evaluation**
  - It is a classification algorithm that generates one rule for each predictor in the data, then selects the rule with the smallest total error as its "one rule”.
- **Gain Ratio Attribute Evaluation**
  - It evaluates the worth of an attribute by measuring the gain ratio with respect to the class.
- **Classifier Attribute Evaluation**
  - Evaluates the worth of an attribute by using a user-specified classifier.
- **InfoGain Attribute Evaluation**
  - Attributes that contribute more information will have a higher information gain value and can be selected, whereas those that do not add much information will have a lower score and can be removed.
- **SymmetricalUncert Attribute Evaluation**
  - This algorithm overcomes the bias of information gain towards the features with the more values by normalizing its value to the range[0,1]

Classification Algorithms

- **Naïve Bayes Algorithm**
  - Naïve Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable
- **J-48**
  - J48 is an open source Java implementation of the C4.5 algorithm. C4.5 builds decision trees from a set of training data.
- **Adaboost M1 Algorithm**
  - AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.
- **Decision Stump Algorithm**
  - Decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rules.

Performance

Classification	Attribute	Accuracy %	TP Rate	FP Rate	ROC Area	F-Measure	Precision	RMS Error
Naïve Bayes	OneR	79.82	0.798	0.202	0.844	0.796	0.815	0.394
	Gain Ratio	79.80	0.798	0.202	0.841	0.795	0.817	0.394
	Classifier Attribute	61.98	0.619	0.381	0.678	0.619	0.620	0.488
	Info Gain	78.55	0.786	0.215	0.836	0.783	0.797	0.402
	Symmetrical Uncert	78.82	0.788	0.212	0.836	0.786	0.802	0.401
J48	OneR	81.724	0.817	0.183	0.853	0.815	0.833	0.363
	Gain Ratio	81.724	0.817	0.183	0.853	0.815	0.833	0.363
	Classifier Attribute	63.676	0.637	0.363	0.637	0.582	0.790	0.458
	Info Gain	81.724	0.817	0.183	0.853	0.815	0.833	0.363
	Symmetrical Uncert	81.724	0.817	0.183	0.853	0.815	0.833	0.363

Performance

Classification	Attribute	Accuracy %	TP Rate	FP Rate	ROC Area	F-Measure	Precision	RMS Error
Adaboost M1	OneR	80.894	0.809	0.191	0.836	0.809	0.811	0.384
	Gain Ratio	80.894	0.809	0.191	0.836	0.809	0.811	0.384
	Classifier Attribute	63.866	0.639	0.361	0.637	0.584	0.790	0.458
	Info Gain	80.894	0.809	0.191	0.836	0.809	0.811	0.384
	Symmetrical Uncert	80.894	0.809	0.191	0.836	0.809	0.811	0.384
Decision Stump	OneR	77.614	0.776	0.224	0.773	0.775	0.782	0.414
	Gain Ratio	77.614	0.776	0.224	0.773	0.775	0.782	0.414
	Classifier Attribute	64.004	0.640	0.360	0.637	0.586	0.791	0.457
	Info Gain	77.614	0.776	0.224	0.773	0.775	0.782	0.414
	Symmetrical Uncert	77.614	0.776	0.224	0.773	0.775	0.782	0.414

Responsible factors

21:28:46 - Ranker + OneRAttributeEva

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
78.335 +- 0.078      1   +- 0       12 DEP_DELAY_NEW
64.611 +- 0.037      2   +- 0       21 NAS_DELAY
63.845 +- 0.064      3   +- 0       23 LATE_AIRCRAFT_DELAY
62.796 +- 0.058      4   +- 0       19 CARRIER_DELAY
56.707 +- 0.185      5   +- 0        1 FL_DATE
55.957 +- 0.133      6   +- 0       14 ARR_TIME
55.031 +- 0.177      7   +- 0       11 DEP_TIME
54.098 +- 0.097      8.2 +- 0.4     10 CRS_DEP_TIME
53.775 +- 0.239      9.2 +- 0.75    13 CRS_ARR_TIME
53.669 +- 0.153      9.6 +- 0.49     2 OP_UNIQUE_CARRIER
53.182 +- 0.089     11   +- 0       16 ACTUAL_ELAPSED_TIME
53.018 +- 0.098     12.3 +- 0.46     6 ORIGIN_STATE_ABR
52.899 +- 0.124     12.7 +- 0.46     4 ORIGIN_AIRPORT_ID
52.796 +- 0.129     14   +- 0        5 ORIGIN_CITY_NAME
52.244 +- 0.081     15.8 +- 0.87     7 DEST_AIRPORT_ID
52.226 +- 0.075     16.2 +- 0.75     8 DEST_CITY_NAME
52.166 +- 0.198     16.3 +- 1.19    18 DISTANCE
51.953 +- 0.127     17.7 +- 0.46     9 DEST_STATE_ABR
51.588 +- 0.014     19.4 +- 0.49    20 WEATHER_DELAY
51.491 +- 0.226     19.6 +- 0.49     3 OP_CARRIER_FL_NUM
50.566 +- 0.292     21.4 +- 0.49    15 CRS_ELAPSED_TIME
50.301 +- 0.283     21.8 +- 0.75    17 AIR_TIME
50.046 +- 0.007     22.8 +- 0.4     22 SECURITY_DELAY
```

Conclusion

- After all the preprocessing and deploying the model in WEKA, there were several metrics evolving across the 20 models. The major metrics that should be taken into consideration are ROC area, Precision and Root mean square error.

- However, considering ROC area as a major parameter to select the best model, J48 with OneR Attribute selection gives the highest ROC area (0.853). The other models weren’t compared based on the class accuracy because sometimes there may be a case where the class is biased, or the TP rate may not have been accurate.

Reference

1. Naïve Bayes: Devin Soni, “Introduction to Naïve Bayes Classification” (source: https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54)

2. J48: Gaganjot Kaur, “Improved J48 Classification Algorithm for the Prediction of Diabetes” (source:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.678.9273&rep=rep1&type=pdf)

3. AdaBoost M1
(source: https://en.wikipedia.org/wiki/AdaBoost)

4. Decision Stump
(source: https://en.wikipedia.org/wiki/Decision_stump)

Thank You