PREDICTING ON-TIME FLIGHTS

Rowan Rollman

DATASET

- Data comes from the U.S. Department of Transportation - Bureau of Transportation Statistics
- The website provided allows you to create your own table
- 537,902 records
- 11 columns

- https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr

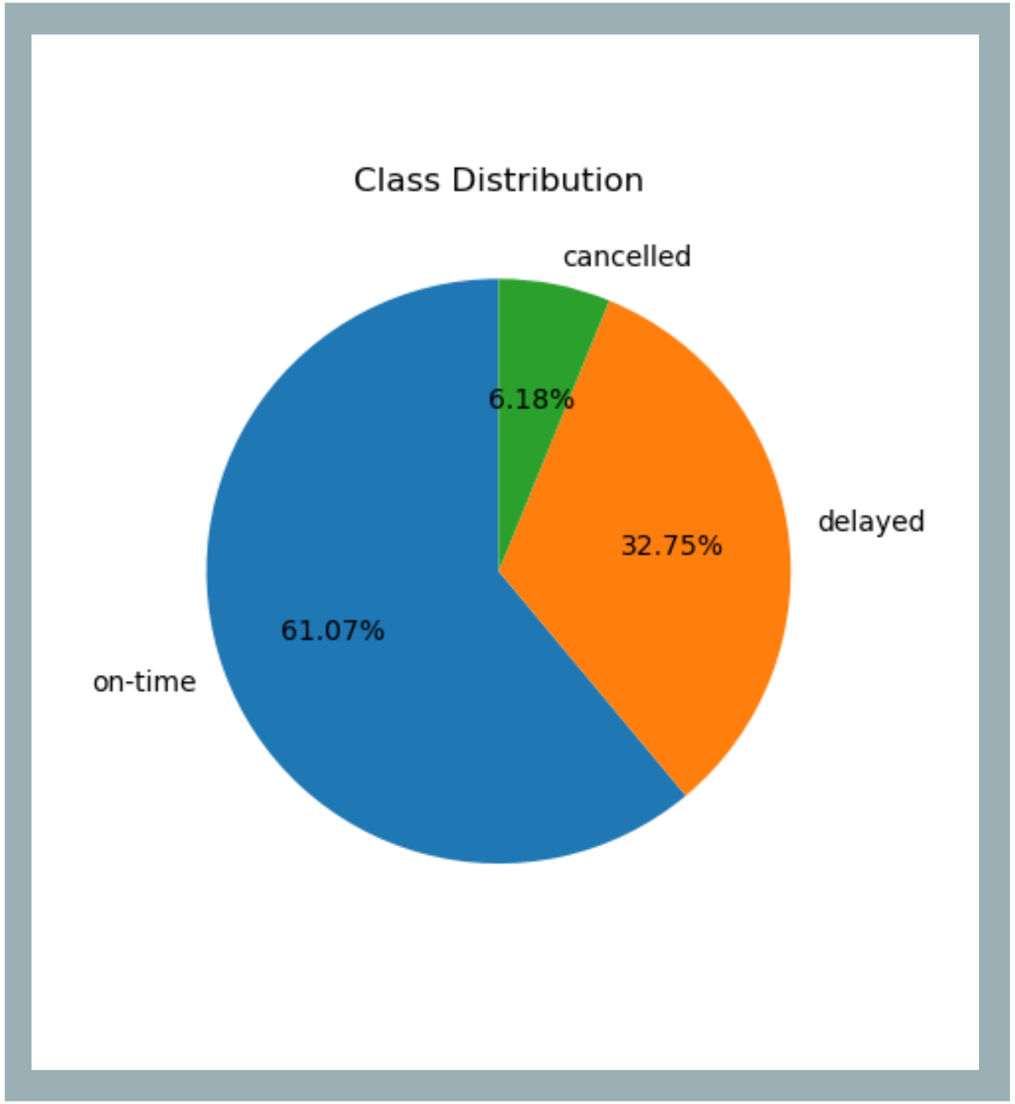
EXPLORATORY DATA ANALYSIS

VARIABLES

Variable	Description
OP_UNIQUE_CARRIER	Unique carrier code, specific to a certain airline
ORIGIN	Code of the origin airport, where the flight departed from
DEST	Code of the destination airport
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.
CANCELLED	Cancelled Flight Indicator (1=Yes)
DAY_OF_THE_MONTH	Numerical representation of the day of the month
DAY_OF_THE_WEEK	Numerical representation of the day of the week

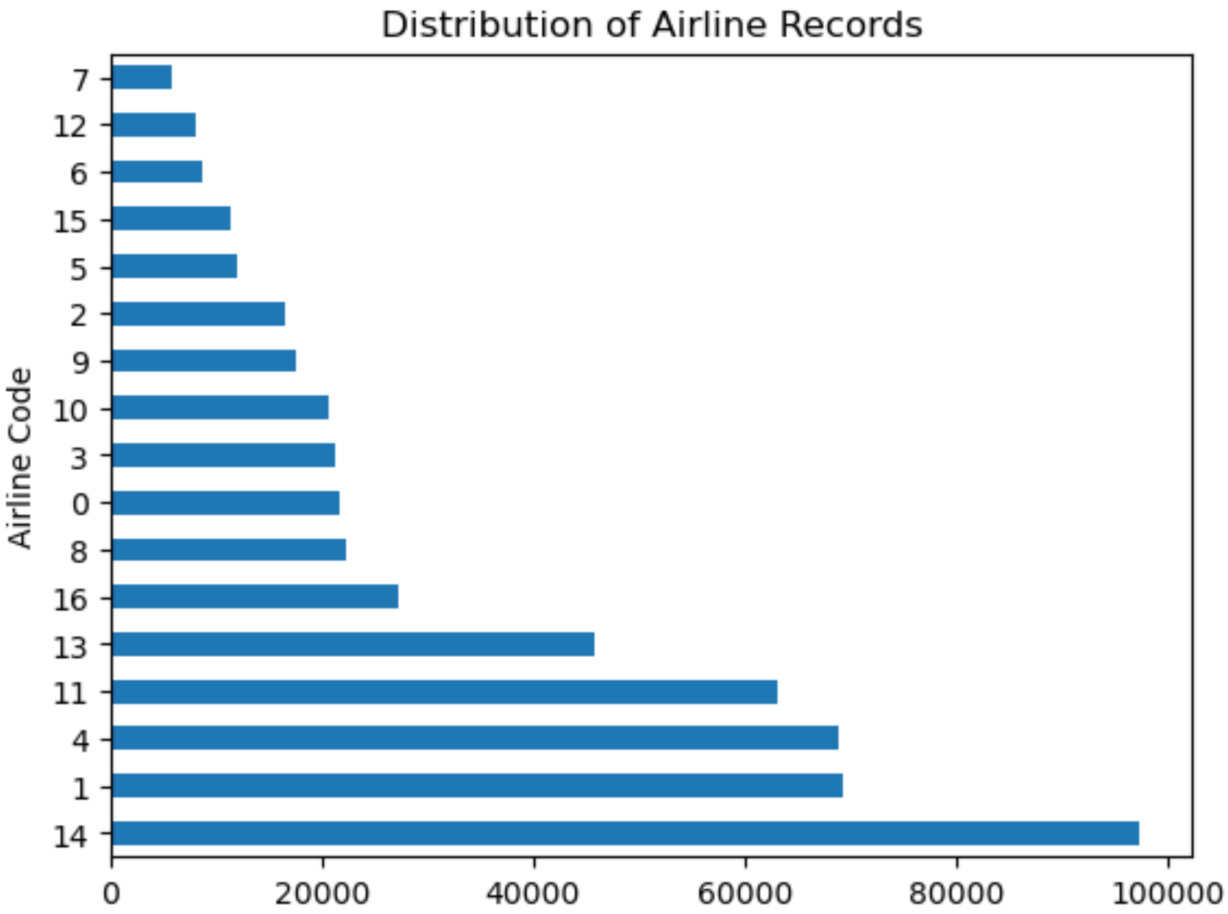
DATA

- 17 unique airlines
- Create indicator for on-time, delayed, or cancelled
- Replace empty cells with 0
- Double check for unexplained missing values

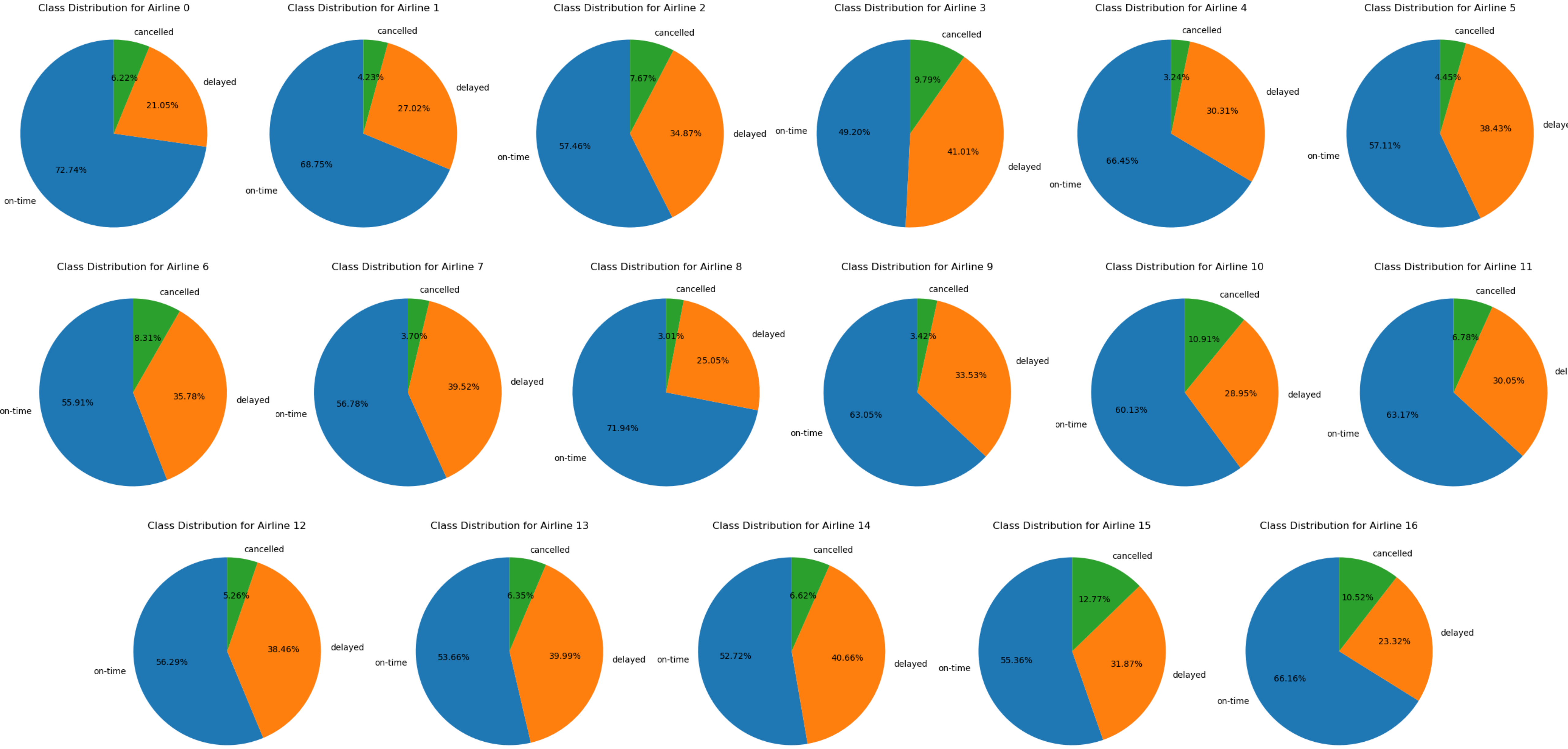


Beginning distribution of classes is showed in the pie-chart

- 328470 on-time
- 176176 delayed
- 33256 cancelled



Distribution of Airline Records
Airline Code
7
12
6
15
5
2
9
10
3
0
8
16
13
11
4
1
14
0
20000
40000
60000
80000
100000



Class Distribution for Airline 0

Class Distribution for Airline 1

Class Distribution for Airline 2

Class Distribution for Airline 3

Class Distribution for Airline 4

Class Distribution for Airline 5

Class Distribution for Airline 6

Class Distribution for Airline 7

Class Distribution for Airline 8

Class Distribution for Airline 9

Class Distribution for Airline 10

Class Distribution for Airline 11

Class Distribution for Airline 12

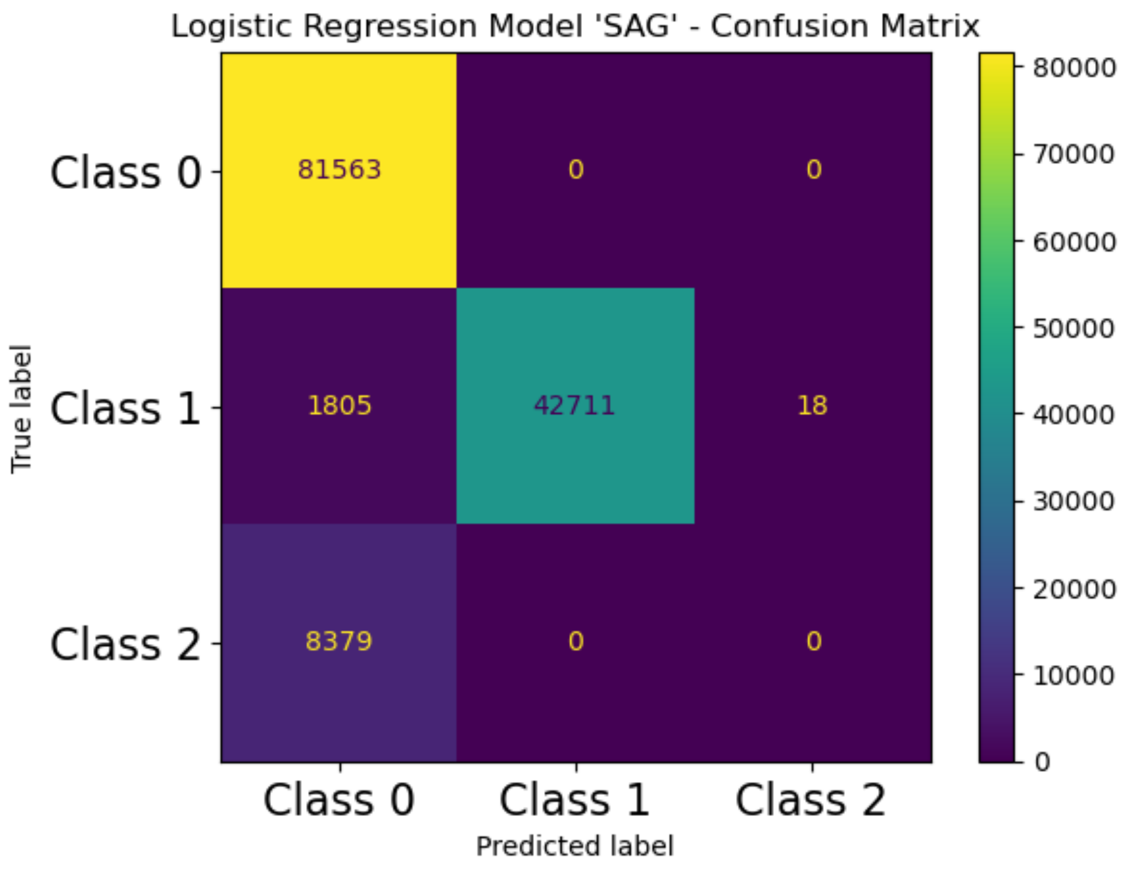
Class Distribution for Airline 13

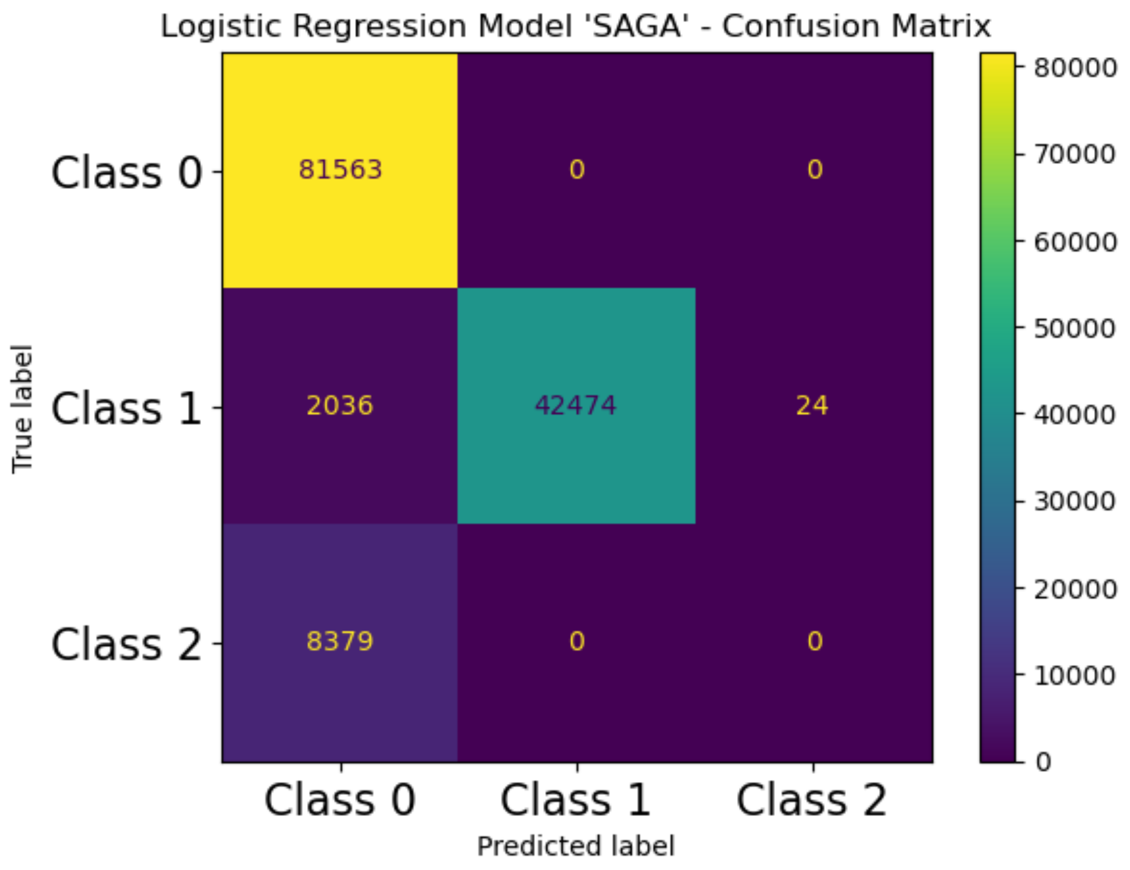
Class Distribution for Airline 14

Class Distribution for Airline 15

Class Distribution for Airline 16

LOGISTIC REGRESSION





Model score: 92.413%
Cross validation scores:
[0.92064299 0.9212617 0.92081552 0.92079073
0.92007188]
Mean of cross validation scores: 0.920716564

Model score: 92.237%
Cross validation scores:
[0.92020921 0.92090227 0.920233 0.92029497
0.91974964]
Mean of cross validation scores: 0.920277816

KNN CLASSIFIER

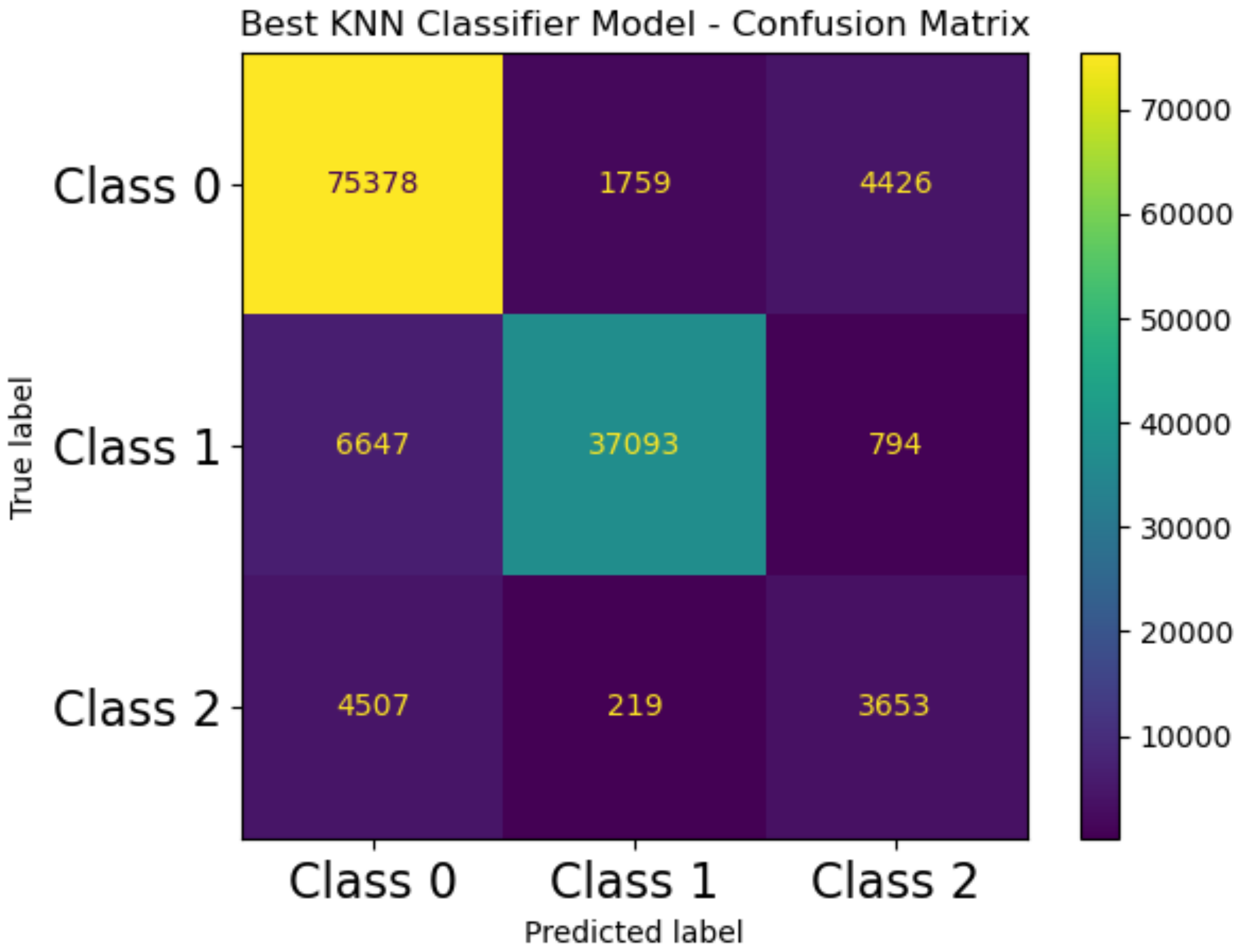
KNN SCORES FROM 1-29

K-Value	Train Score	Test Score
1	0.977529956	0.863529552
2	0.913952001	0.861826646
3	0.914001576	0.858093637
4	0.888095958	0.849846813
5	0.887706791	0.851081234
6	0.871966606	0.843563164
:	:	:
:	:	:

- Best Test Score: 0.8635295517415748
- From K Value 1

BEST K VALUE MODEL

Using the best K value from the previous slide, I created another KNN Model, with a score of 86.35%



Best KNN Classifier Model - Confusion Matrix

True label \ Predicted label	Class 0	Class 1	Class 2
Class 0	75378	1759	4426
Class 1	6647	37093	794
Class 2	4507	219	3653

CROSS VALIDATION ON KNN

The array of scores is as shown

Max Score:.85.646%

[0.8564594217462498, 0.8560603395076043,
0.8529147836781534, 0.845498301159704,
0.8465542596698821, 0.8389320428721128,
0.8402978445699109, 0.8342223797424044,
0.835134566693872, 0.8299465022918342,
0.8306380800376992, 0.8262209123230466,
0.8268777868434677, 0.8229538948051658,
0.8235388851437371, 0.8200810015448632,
0.8206387235098248, 0.8175700076006391,
0.8179195137458823, 0.8154333076793925,
0.8153787742500693, 0.8133858450273384,
0.8133833648693967, 0.8113309453729615,
0.8112045269937571, 0.8093479300847621,
0.8092785244326273, 0.8075880034227968,
0.8072979866603409]

XGBCLASSIFIER

Used multiple learning rates to find the optimal one,

learning_rates = [0.05, 0.1, 0.25, 0.5, 0.75, 1]

xgb.XGBClassifier(max_depth=10, n_estimators=100, learning_rate=lr)

Learning Rate	Model Score
0.05	0.9474999256372885
0.1	0.9483179154644695
0.25	0.9496638805437402
0.5	0.9521773401945328
0.75	0.9523483744311253
1.0	0.9521029774829709

Best learning rate is 0.75 with a score of 0.9523483744311253

LIGHT GBM

```
lgbm.LGBMClassifier(num_leaves=100,n_estimators=100,max_depth=10,learning_rate=lr,
bagging_fraction=.8, bagging_freq=5)
```

Learning Rate	Model Score
0.05	0.9518798893482852
0.1	0.9533373984948987
0.25	0.9535233052738035
0.5	0.882767185222642
0.75	0.5450414943930515
1.0	0.8444629524970999

Best learning rate is 0.25 with a score of 0.9535233052738035

HISTGRADIENTBOOSTINGCLASSIFIER

HistGradientBoostingClassifier(learning_rate=lr)

Learning Rate	Model Score
0.05	0.9496936256283649
0.1	0.9515006395193194
0.25	0.5756640590142479
0.5	0.9450905737826825
0.75	0.9329248341711532
1.0	0.8978553793985544

Best learning rate is 0.1 with a score of 0.9515006395193194

LAST TWO MODELS

SGDClassifier

- Running on all processors
- Using log_loss
- Score 0.8893259763824028

ADABoostClassifier

- N_estimators 100
- Random_state = 42
- Score 0.937341979237931