Flight Delay Prediction

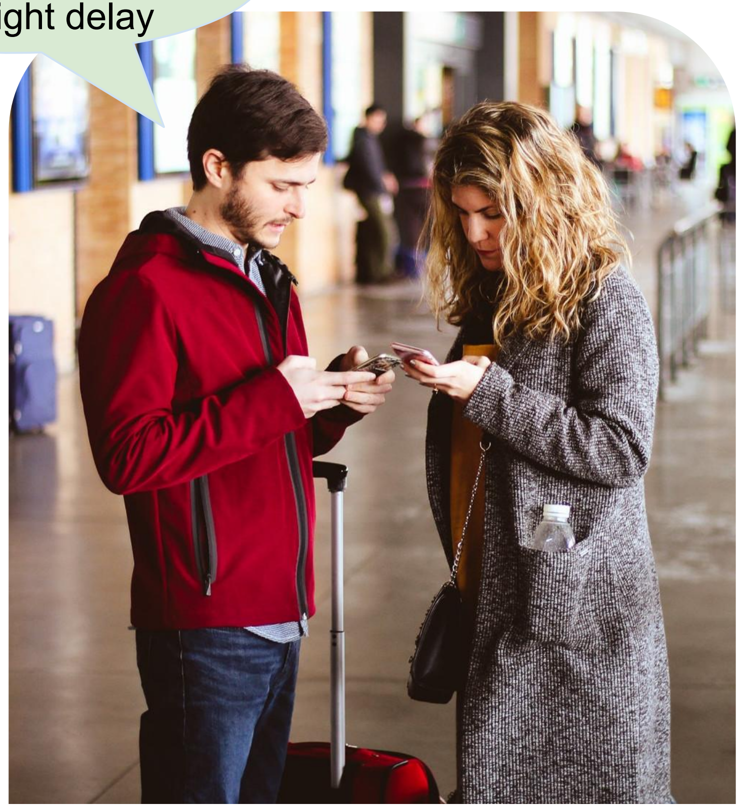
Team 3

Hongyuan Li
Na Yue
Fan Wu

Motivation & Overview

It would be great
if someone can
predict flight delay

- Build an app to predict flight delay
  - friendly simple input
  - include weather in prediction
- Build model on flight delay dataset
  - Join flight delay with weather data
  - Analyze improvement in the model
- Consider differences among airports
  - Clustering



Delayed Flight Data Set - Introduction

1. From the U.S. Department of Transportation's (DOT)
2. Period: 2015 - 2017
3. # Entries 17,111,358
4. 14 major airlines & 335 Airports

tab_info

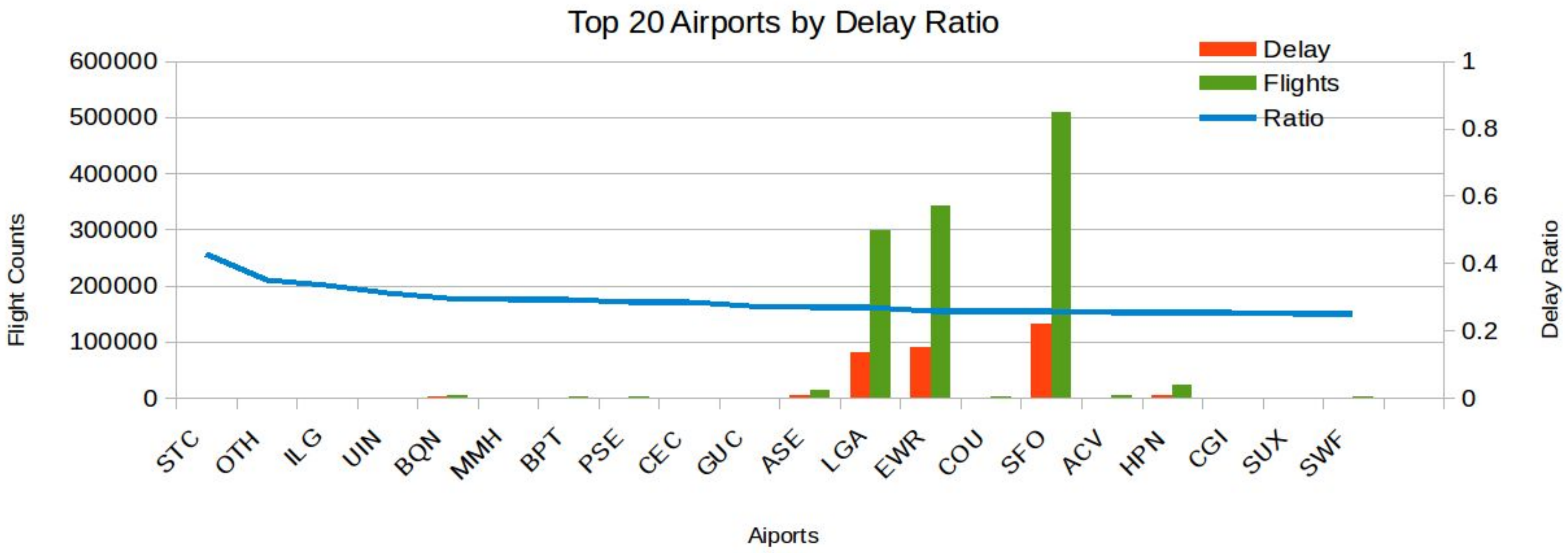
		QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	FL_DATE	UNIQUE_CARRIER	FL_NUM	ORIGIN	DEST	CRS_DEP_TIME	CRS_ARR_TIME	ARR_DEL15	CANCELLED	DIVERTED	DISTANCE
column type	**int64**	int64	int64	int64	int64	object	object	int64	object	object	int64	int64	float64	float64	float64	float64
null values	**0**	0	0	0	0	0	0	0	0	0	0	0	279795	0	0	0
null values (%)	**0**	0	0	0	0	0	0	0	0	0	0	0	1.6351419916525618	0	0	0

Delayed Flight Data Set - Preprocessing

Step 1. Label entries
Step 2. Cleaning dataset

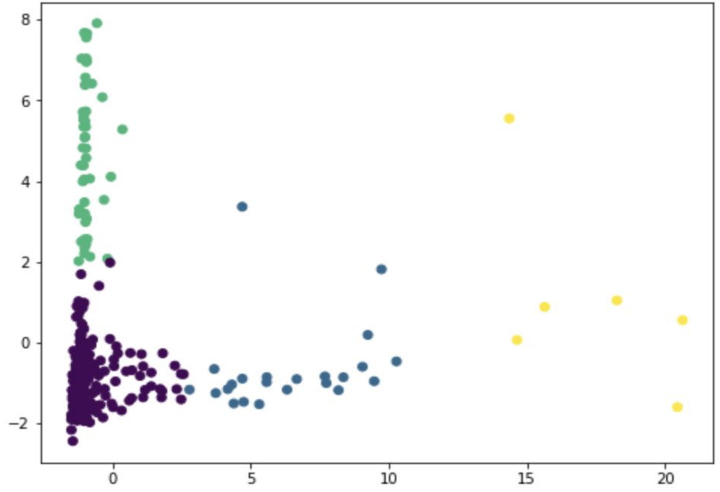
YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	FL_NUM	ORIGIN	DEST	CRS_DEP_TIME	CRS_ARR_TIME	AIR_TIME	DISTANCE	LABEL
2015	1	1	22	4	6876	1485	767165	776779	2050	2354	134	950	0
2015	1	1	22	4	6876	1503	678671	776779	1355	1609	102	757	0
2015	1	1	22	4	6876	1509	778380	776779	1112	1527	168	1310	0
2015	1	1	22	4	6876	1510	778380	776779	1918	2328	160	1310	0
2015	1	1	22	4	6876	1585	767165	776779	1829	2142	142	950	0
2015	1	1	22	4	6876	1669	658476	776779	1855	2026	56	404	0
2015	1	1	22	4	6876	1685	767165	776779	1404	1717	141	950	0

Airport and Flight Delay



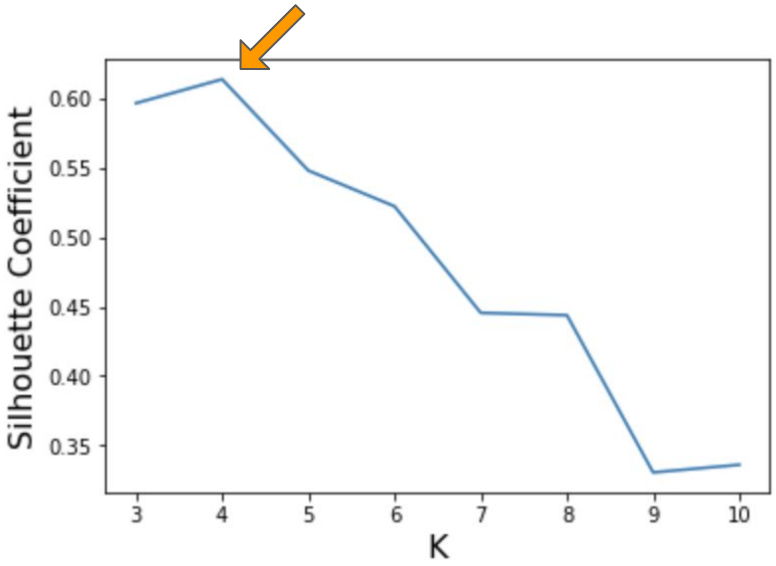
Large airports: LaGuardia Airport (LGA), Newark Airport(EWR) in New Jersey and San Francisco International Airport (SFO)

Airport Clustering



Dimension reduction for visualization

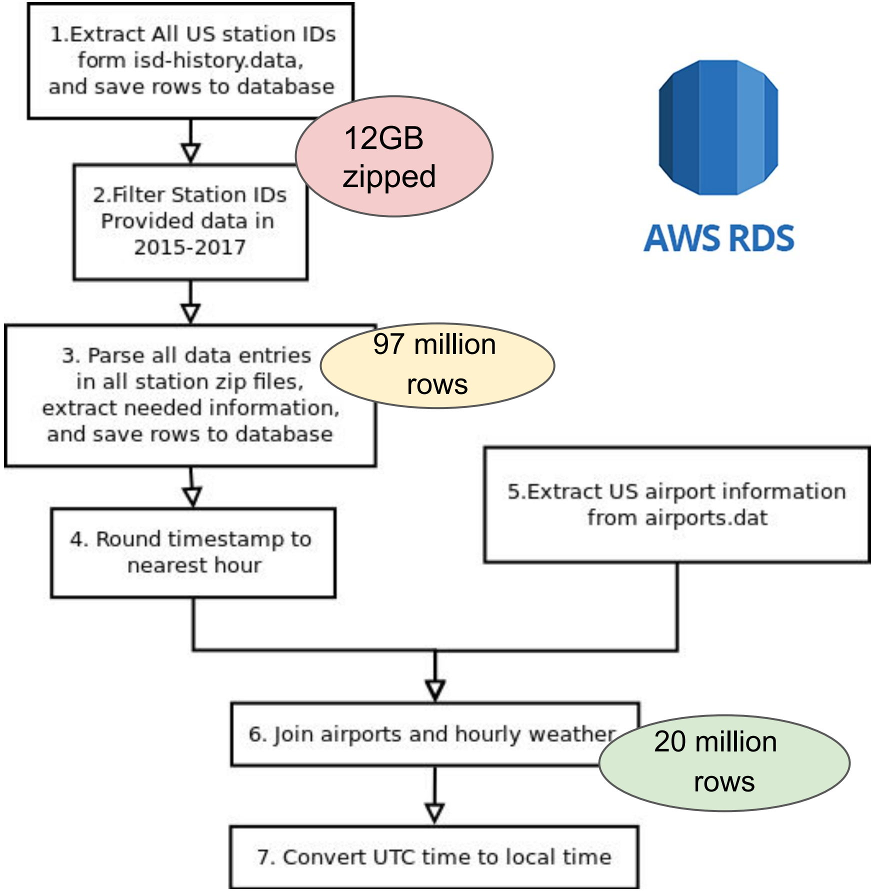
Best Value K



Weather Dataset - Preprocessing

National Oceanic and Atmospheric Administration (NOAA)

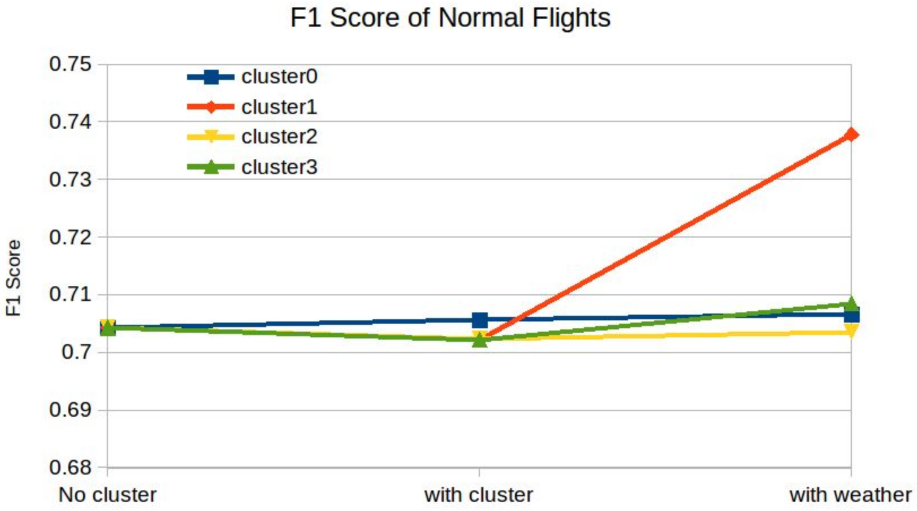
0248745090232442017112305004+37417-122050FM-15+0012KNUQ V0203601N0103122000019N0160931N1+01301-00601102391ADDGF100991999999999999999999KA1999M+01391KA2999N+01061MA1102371999999MD1610161+9999REMMET120METAR KNUQ 312356Z AUTO 36020KT 10SM CLR 13/M06 A3023 RMK AO2 PK WND 36027/2311 SLP239 T01281061 10139 20106 56016 TSNO;EQDQ01 993SCCGA1Q02 099SCCGD1



AWS RDS

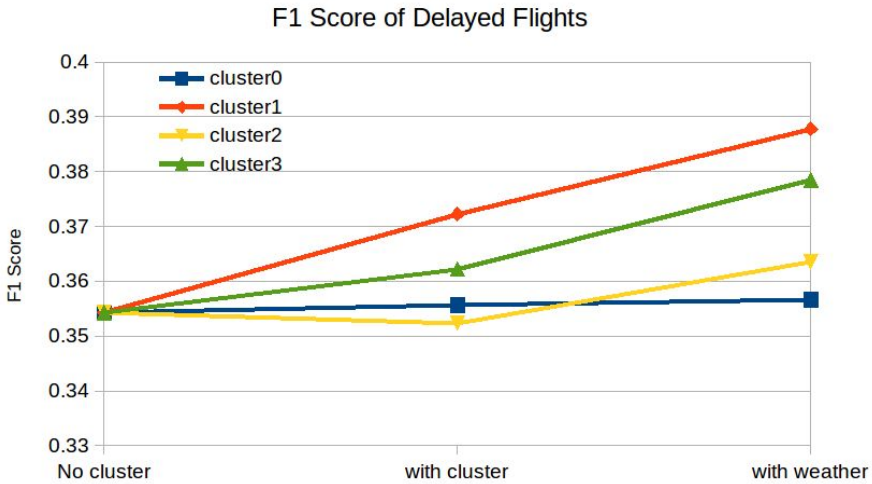
iata	ts	wind_angle	wind_speed	vis	temp	liqu_depth	snow_depth	local_ts	tz	code
AAF	2017-11-23 05:00:00	16.66666667	4.766666666	16,093	15.4	6.4	0	2017-11-23 00:00:00	America/New_York	656570

Strategy Analysis



Is including clustering and weather helping?

boosted the F1 scores for predicting delayed flights



limited effect on the f1 score of normal flights

Algorithm: Random Forest

Classification models with different algorithms

	F1 score for class “0”	F1 score for class “1”	average
cluster0 with decision tree	0.81638768	0.8164076	0.816425853882
cluster1 with KNN(n=3)	0.86564218	0.3068247	0.794709209948
cluster2 with SGD classifier	0.89086339	0.02819864	0.883121155655
cluster3 with SVC	0.89237248	0.10431948	0.876614764671

Try out different algorithms on different clusters as the delay percentage of all clusters are similar => clusters have similar distribution

Classification - XgBoost

1. Using cross validation to select models
2. Result comparison

	common params	gamma	learning rate	min child weight	max depth
cluster 0	base_score=0.5, booster='gbtree', colsample_bylevel=1, objective='binary:logistic',	5.71	0.32	42.50	27
cluster 1		5.68	0.24	1.58	26
cluster 2		5.91	0.35	45.19	28
cluster 3		5.60	0.35	5.32	11

	XgBoost with clustering				XgBoost without clustering
	cluster 0	cluster 1	cluster 2	cluster 3	entire dataset
F1 score(on time)	0.9	0.9	0.9	0.9	0.83
F1 score(delay)	0.26	0.28	0.33	0.12	0.44
F1 scoure（Total)	0.88	0.87	0.86	0.88	0.76
Count(delay)	956453	876288	1399597	80502	3312840
Count(total)	5058034	4402501	7120476	426316	17007327
Percentage(delay)	18.91%	19.90%	19.66%	18.88%	19.48%

Application Architecture