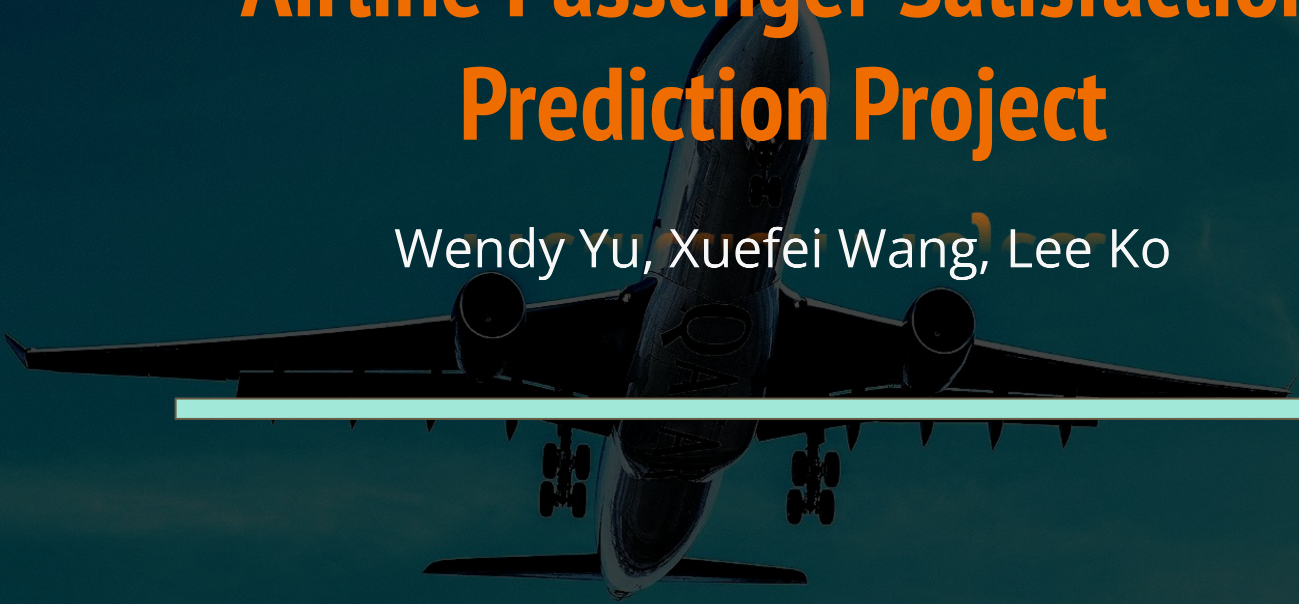
Airline Passenger Satisfaction Prediction Project

Wendy Yu, Xuefei Wang, Lee Ko



Project Introduction

- **Question**
  - What are major factors of Airline passenger satisfaction?
  - Can we predict passenger satisfaction?
- **Task**
  - Train a binary classification model to predict customer satisfaction of airline passengers based on airline customer survey
- **Provided information**
  - Satisfaction level for each survey (Satisfied: 1, Neutral or dissatisfied: 0)
  - Information from survey (22 columns): Gender, Customer type, Age, Type of travel, Class, Flight distance, Inflight wifi service, etc.

Data Overview

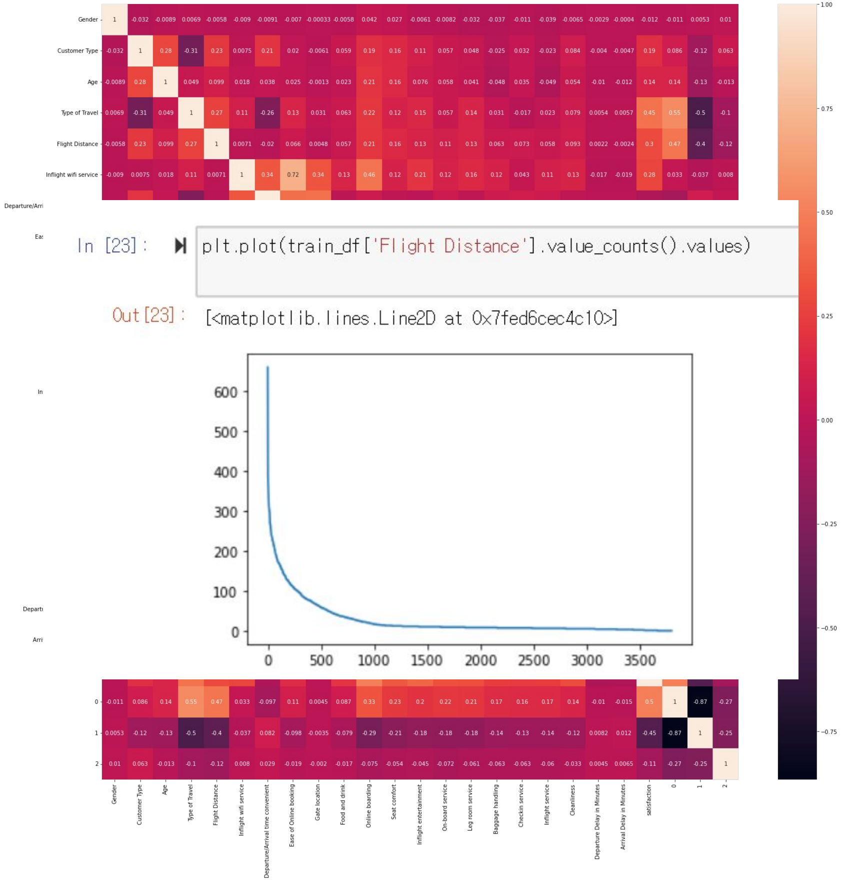
- Kaggle dataset Airline Passenger Satisfacti
  - 80% train data, 20% test data (103904 + 25976 =
  - 25 columns
    - 22 distinct features: 4 categorical, 18 numer
    - 3 columns to be drop/split: index, id, satisfaction (target column)
  - Class distribution
    - Binary classification
    - 43% “satisfied” (11403)
    - 57% “neutral or dissatisfied” (14573)
  - Competition data
    - Cleaned up to some extent, more focus on transformation/feature engineering

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	Inflight entertainment	On-board service	Leg room service
0	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	...	5	4	3
1	1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	...	1	1	5
2	2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	...	5	4	3
3	3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	...	2	2	5
4	4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	...	3	3	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
25971	25971	78463	Male	disloyal Customer	34	Business travel	Business	526	3	3	...	4	3	2
25972	25972	71167	Male	Loyal Customer	23	Business travel	Business	646	4	4	...	4	4	5
25973	25973	37675	Female	Loyal Customer	17	Personal Travel	Eco	828	2	5	...	2	4	3
25974	25974	90086	Male	Loyal Customer	14	Business travel	Business	1127	3	3	...	4	3	2
25975	25975	34799	Female	Loyal Customer	42	Personal Travel	Eco	264	2	5	...	1	1	2

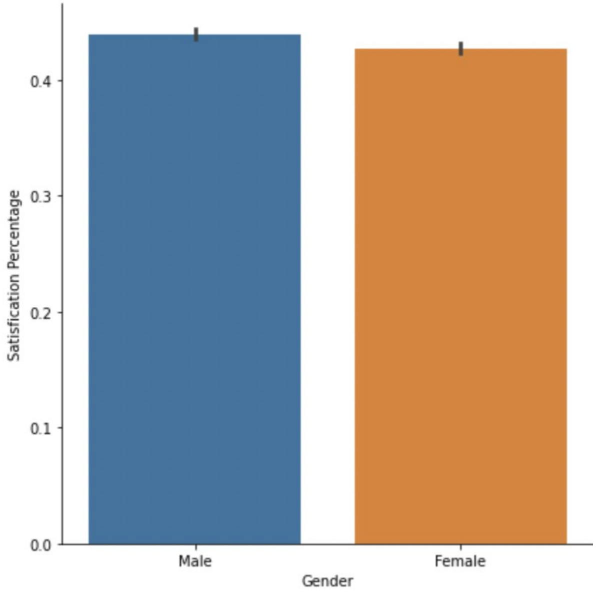
129487 rows × 25 columns

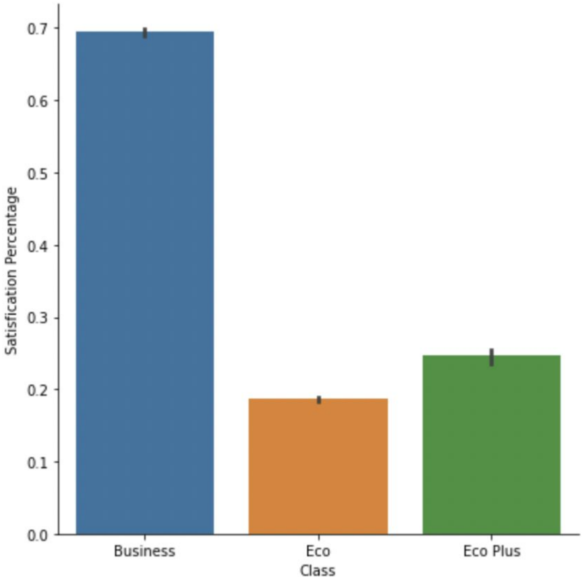
Data Processing

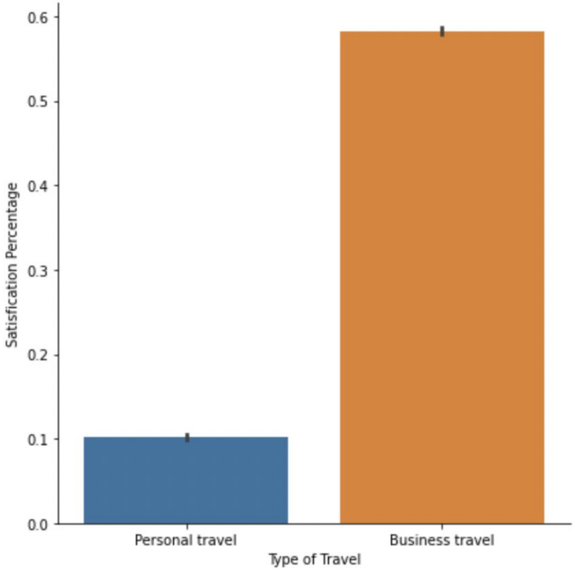
- Preprocessing
  - Dropping unused columns (‘Unnamed’, ‘id’)
  - Data rearranging
    - OneHotEncoder
  - Filling in missing values
    - Numerical (mean)
    - categorical (re-categorize)
  - Outlier detection
    - Set upper threshold (Q3 + 1.5*IQR)
- Exploration
  - Distribution of values
  - Correlation check

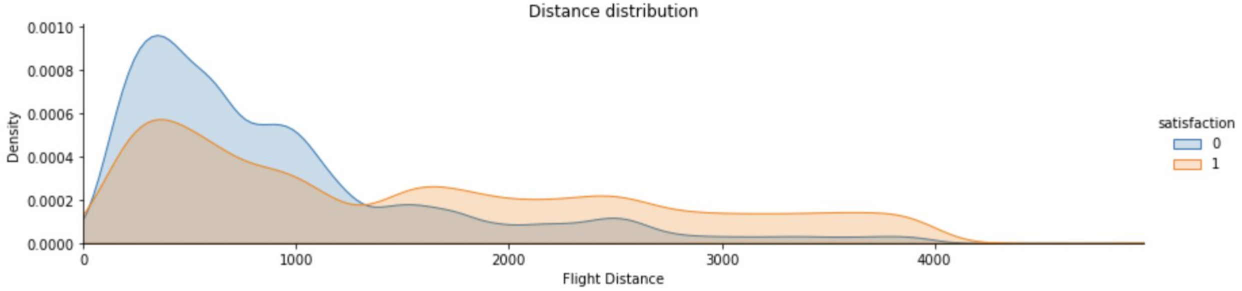


Data Exploration









Model

- Models of choice
    1. Decision Tree (DT)
    2. Random Forest (RF)
    3. Linear Regression (LR)

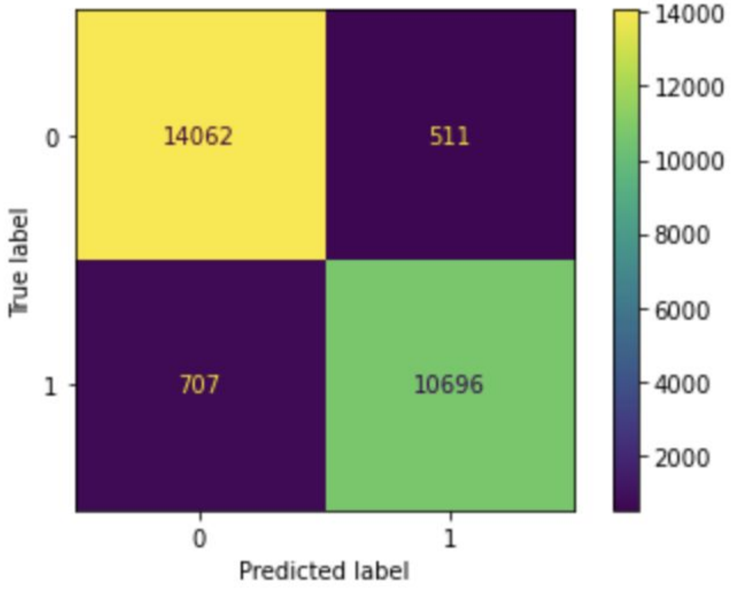
Model - 1. Decision Tree

- Parameter tuning
  - Random state, max_depth, and min_sample_
  - Model accuracy score went up
  - Model performance score after parameter tu
- Result Analysis
  - Online boarding, inflight wifi, type of travel
- Error Analysis
    - Confusion matrix
    - 1218 instances

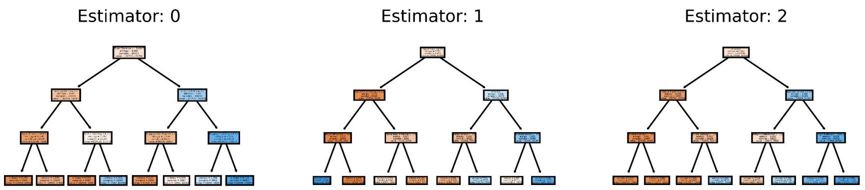
Feature

Depai

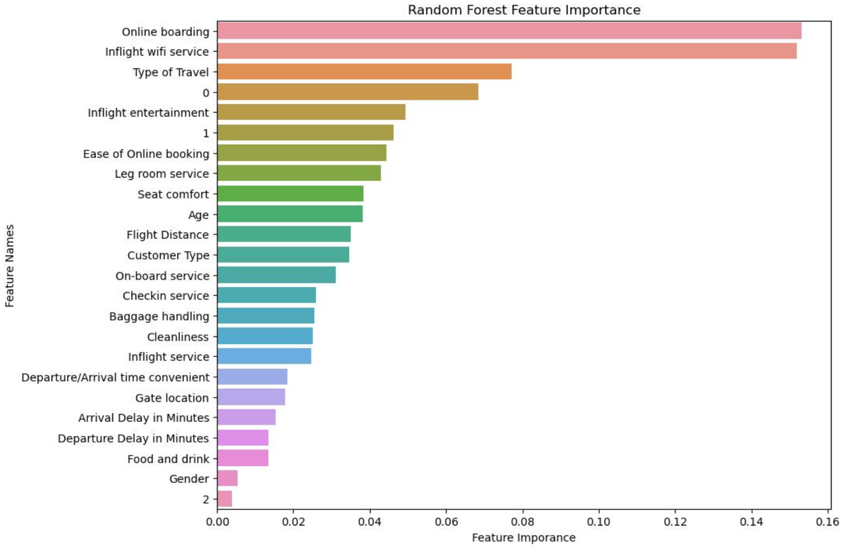
```
Confusion matrix

 [[14062   511]
 [  707 10696]]
```



Model - 2. Random Forest



- Parameter tuning
  - Experiment with n_estimators, max_depth, min_samples_split
    - n_estimators ↑, max_depth ↑, min_samples_split ↓: Higher score
  - Trade-offs with computing time
  - Highest model score: 0.9641

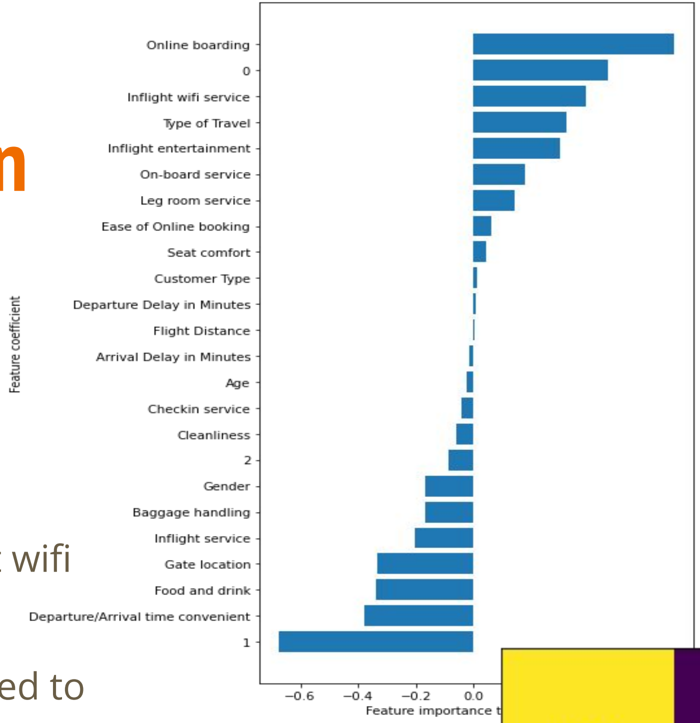


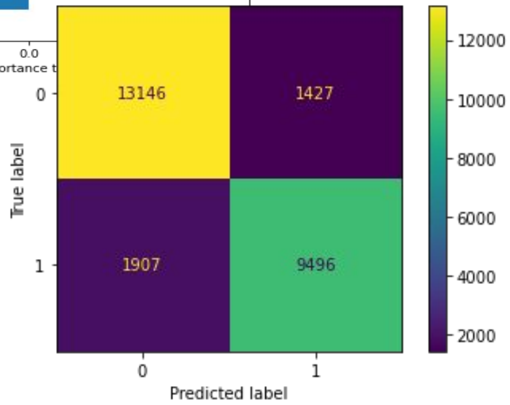
- Results

- Classification errors
  - Corrected in RF compared with DT
  - Coefficient of Online boarding service
    - DT: 0.318974
    - RF: 0.153231 (Robust, reasonable)
  - Giving more weights to specific feature could be useful to make prediction simple, but it results in more classification errors

Model - 3. Logistic Regression

- Parameter tuning
  - C, max_iter
  - class_weight
  - Highest model score: 0.8717
- Result Analysis
  - Online boarding, business class, inflight wifi service
  - Error Analysis
    - More classification errors compared to RF and DT
    - Simple algorism





Model Comparison

	1. Decision Tree	2. Random Forest	3. Logistic Regression
Parameter tuning	Random state = 42 max_depth min_sample_split	n_estimators (# of trees) = 100 max_depth min_samples_split	C = 10 Max_iter = 10000 Class_weight random_state
Strength		Robust coefficient compared to DT	Simple to apply; fast speed
Limitation : Reasoning for error samples	· Limited robustness compared to RF. Too much weight given to one feature, which was corrected in RF	· Slow speed when generating multiple trees with no limit on depths	· Assumption of no collinearity between input variables
Performance score	0.9531	0.9622	0.8717
roc_auc_score	0.9472	0.9598	0.8673