Can Airlines Predict the Satisfaction of their Customers?

— Peter Kryspin, Shijie Li, Ke Wu, Yu Luan, Viona Xu

Motivation

- “Commercial aviation drives 5% of U.S. GDP—the equivalent of $1.25 trillion in 2022” - (https://www.airlines.org/impact)
- Training the models on our data can be used to predict future cases of customer satisfaction
- Predicting satisfaction through modeling allows companies to cater their resources towards the specific areas that improve customer satisfaction the most.

Data Cleaning

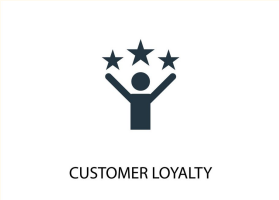
Gender: Gender of the passengers
Customer Type: The customer type (based on loyalty)
Age: The actual age of the passengers
Type of Travel: Purpose of the flight of the passengers
Class: Travel class in the plane of the passengers
Flight distance: The flight distance of this journey
Inflight wifi service: Satisfaction level of the inflight WIFI service
Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
Ease of Online booking: Satisfaction level of online booking
Gate location: Satisfaction level of Gate location
Food and drink: Satisfaction level of Food and drink
Online boarding: Satisfaction level of online boarding
Seat comfort: Satisfaction level of Seat comfort
Inflight entertainment: Satisfaction level of inflight entertainment
On-board service: Satisfaction level of On-board service
Leg room service: Satisfaction level of Leg room service
Baggage handling: Satisfaction level of baggage handling
Check-in service: Satisfaction level of Check-in service
Inflight service: Satisfaction level of inflight service
Cleanliness: Satisfaction level of Cleanliness
Departure Delay in Minutes: Minutes delayed when departure
Arrival Delay in Minutes: Minutes delayed when Arrival
Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

- One-Hot encoding method for categorical level variables

Features in the Models

- **Gender**

  What is the gender of the traveler, male or female?



- **Customer Type**

  Whether this customer is loyal or not?

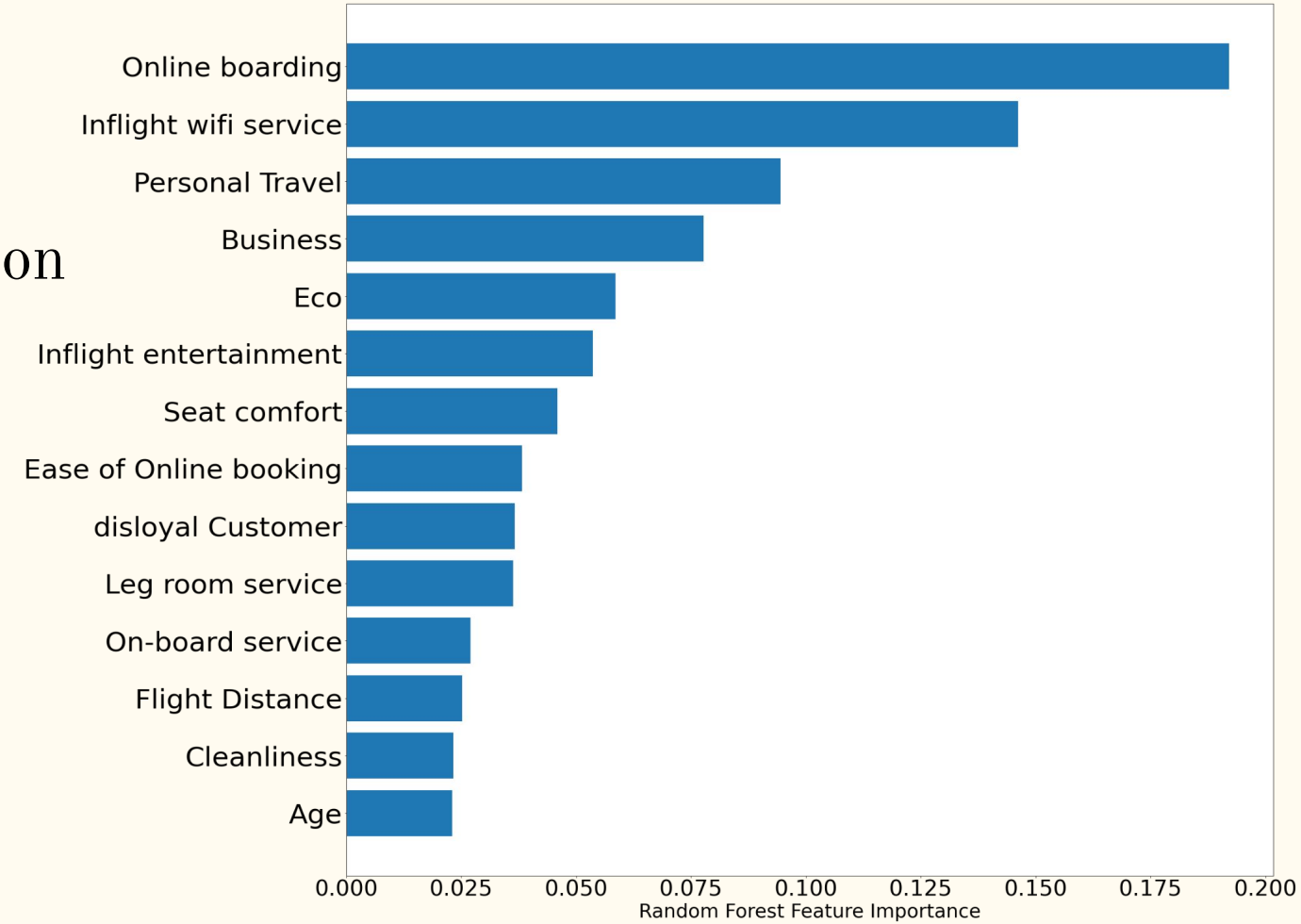
- **Type of Travel**

  A personal travel or business travel.

- **Class**

  What is the class of the flight? Business/Eco/Eco plus

Feature Selection



Building data set

- **Source:**https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction contains two data sets **train.csv** and **test.csv.**

  1.There are 100K rows in train.csv and 26K rows in test.csv.

  2.Based on the response variable. There are 40% satisfaction and 60% neutral or unsatisfied in both train and test data set. So they are not imbalanced.

- **Data Transformation:**

  Change categorical data in features Gender/Customer type/

  Type of Travel/Class/Satisfaction become the dummy variables by get_dummies method

```
def transformData (feature):
    df_ = pd.get_dummies(feature, drop_first=False)
    df_name = df_.columns[0]
    return df .drop([df name], axis=1)
```

Response variable

- **Satisfaction**

  Whether this traveler is satisfied or not (coded as a 1 for satisfied or a 0 for neutral/dissatisfied).

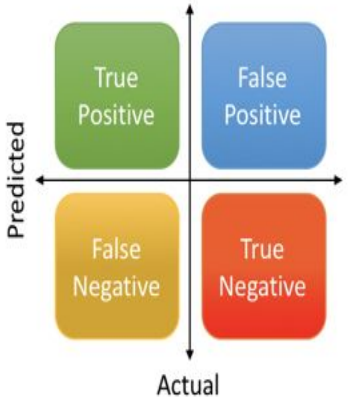
Our Models

- Logistic regression
- KNN classification
- Decision tree
- Stacking model

Precision = $\frac{\text{True Positive}}{\text{Actual Results}}$ or $\frac{\text{True Positive}}{\text{True Positive + False Positive}}$

Recall = $\frac{\text{True Positive}}{\text{Predicted Results}}$ or $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$

Accuracy = $\frac{\text{True Positive + True Negative}}{\text{Total}}$



Model: LASSO Logistic Regression

- **Parameters:**

  C=1

  Penalty: L1 Penalty

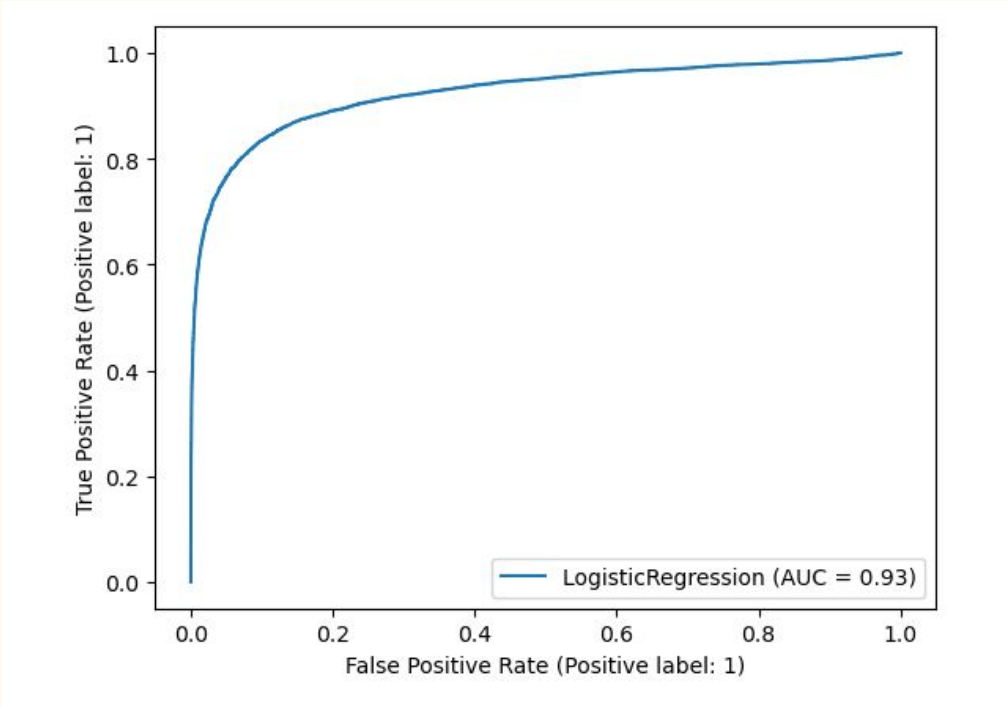
- **Model Performance:**

  Accuracy =0.872

  Recall=0.833

  Precision=0.869

  AUC=0.926



ROC CURVE for Logistic Regression Model

Model: Decision Tree

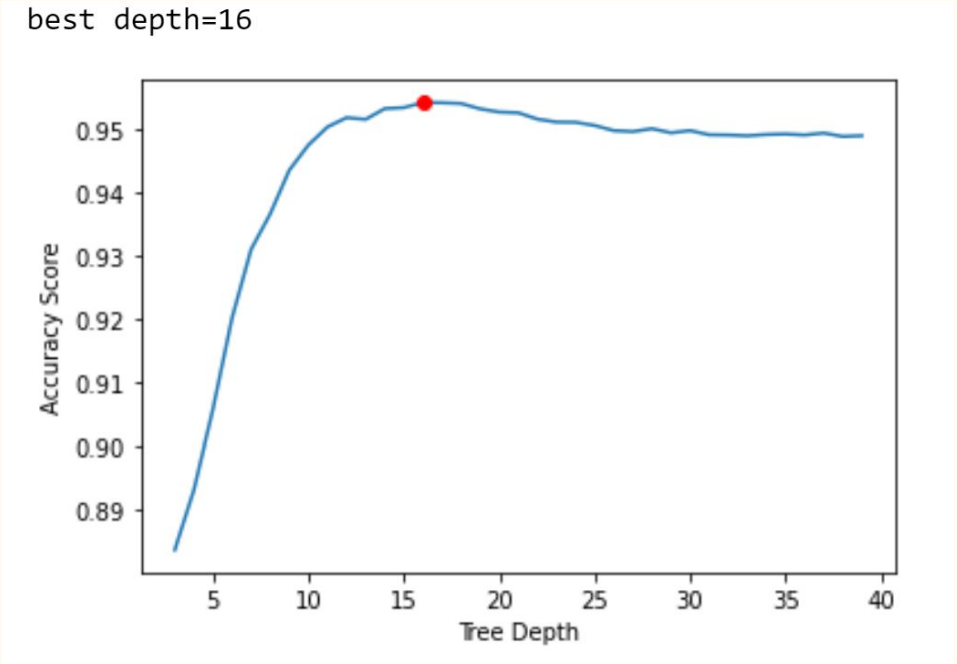
- **Hyperparameter tuning:**

1. Find best max_depth=16

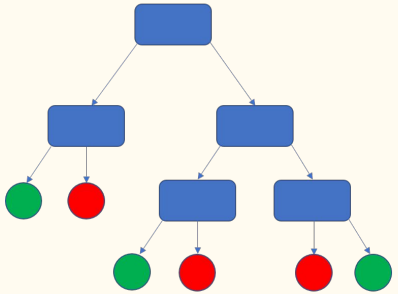
2. Grid search by comparing

criterion ["gini","entropy","log_loss"]

3.Best criterion is “Entropy”



Model: Decision Tree



- **Parameters:**

  criterion='entropy'

  max_depth=16

  random_state=0

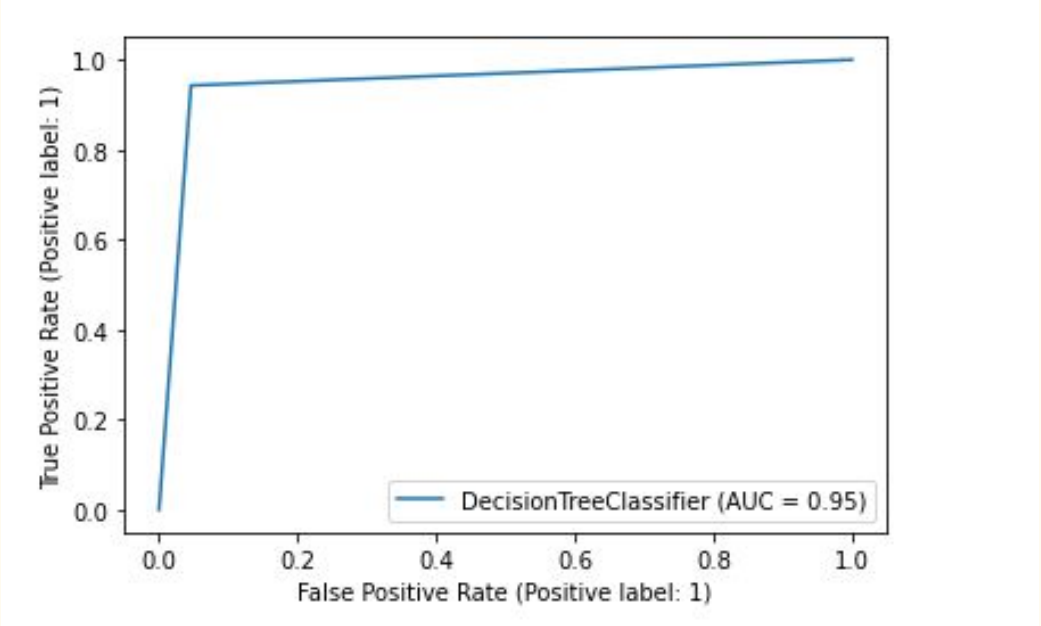
- **Model Performance:**

  Accuracy =0.949

  Recall=0.943

  Precision=0.941

  AUC=0.948



ROC CURVE for Decision Tree Model

Model: Knn

- **Parameters:**

  n_neighbors=9

  metric='manhattan'

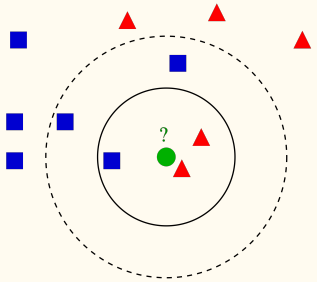
- **Model Performance:**

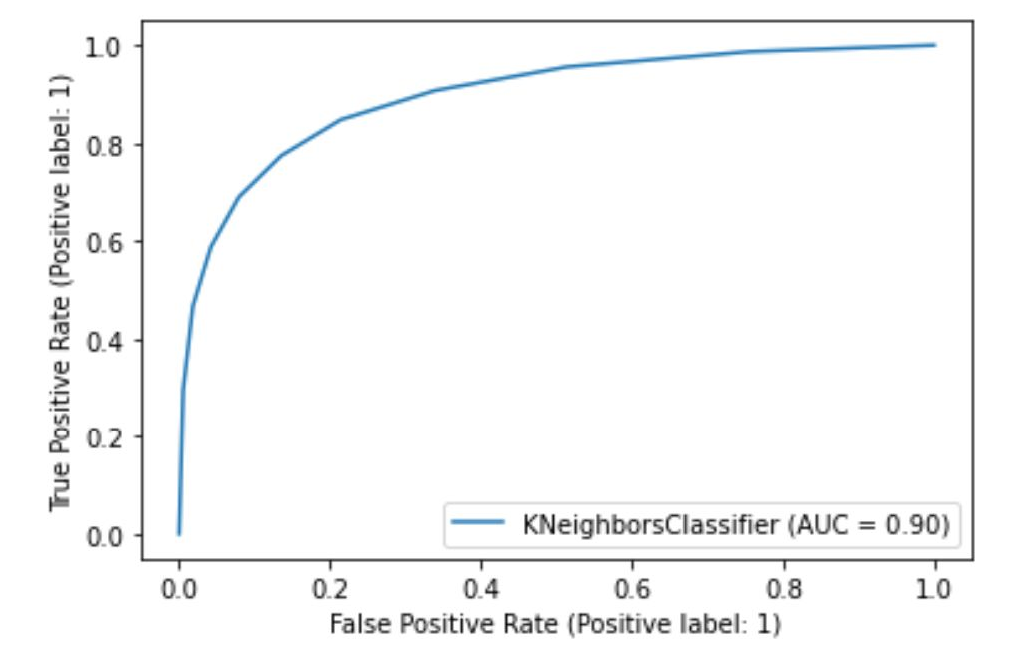
  Accuracy = 0.825

  Precision = 0.817

  Recall = 0.774

  AUC = 0.897





ROC CURVE for Knn Model

Models: Knn and Decision Tree Stacked

- **Parameters:**

  Knn: n_neighbors=9

  metric='manhattan'

  Decision Tree: criterion='entropy'

  max_depth=16

  random_state=0

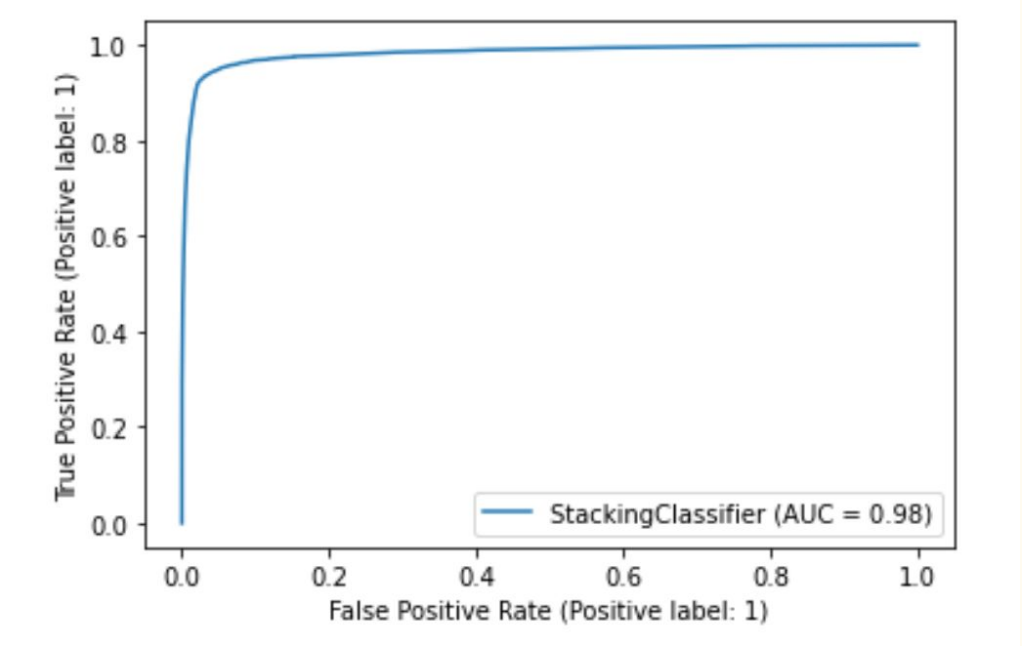
- **Model Performance:**

  Accuracy = 0.953

  Precision is = 0.954

  Recall = 0.939

  AUC = 0.983



ROC CURVE for Stacked Decision Tree & Knn Model

Conclusion

	Accuracy	Recall	Precision	AUC
Logistic Regression	0.872	0.833	0.869	0.926
KNN	0.825	0.817	0.774	0.897
Decision Tree*	0.949	0.943	0.941	0.948
Stacking Model**	0.953	0.954	0.939	0.983

Feature Importance & Possible Issues

Features	Importance
Online boarding	0.19
Inflight wifi service	0.14
Type of Travel	0.09
Class	0.07

- **Unknown Correlation**
- **Single & Unspecified Standard of Satisfaction**