Airline Review Analysis

Brandon Liu, Max Li, Aidan Frantz, Yuchen Cai

Survey

Any unpleasant flight experiences?

Purpose / Objective

Accurately predict whether an airline is recommended based on various features extracted from airline reviews.

Correspondingly, our model can help airlines understand what factors they can work on to improve their services.

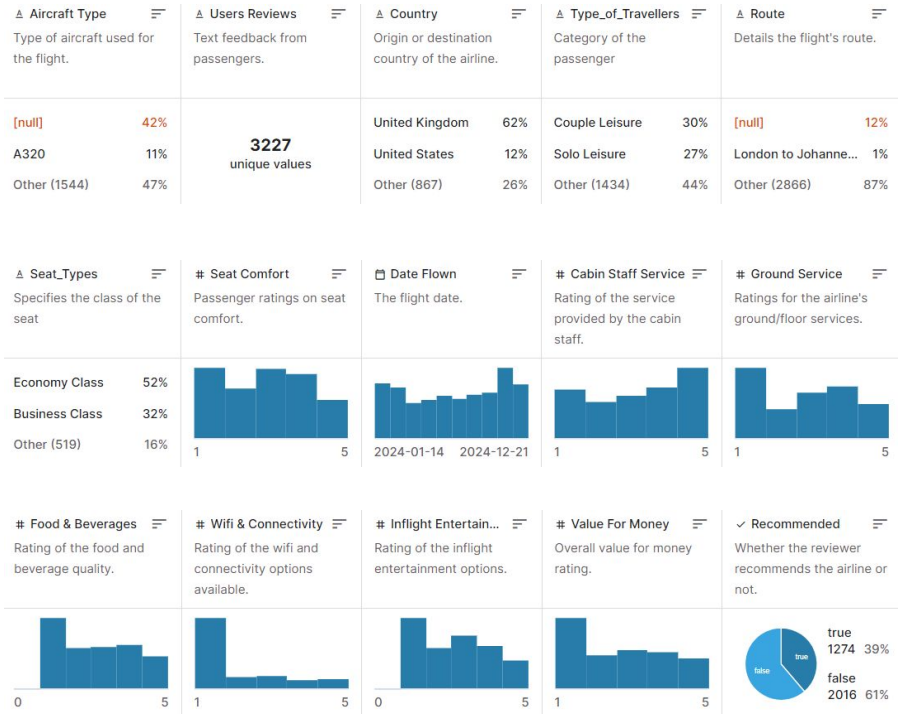
What does our dataset include

3290 records × 14 features

3227 unique users reviews

Categorical data and numerical rating

Predict whether reviewer recommends



https://www.kaggle.com/datasets/anandshaw2001/airlines-reviews-and-rating

Data Preprocessing (not including reviews)

Fillna (11 features allow empty entry)

Drop “users reviews” or perform word processing on “users reviews”

One-hot encoding after dropping “users reviews”

1972 features after one-hot encoding rest 13 features

Algorithms we used

KNN

Linear Regressor

Logistic Regression (L1, L2)

Decision Tree / Random Forest Classifier

Support Vector Machine (SVM)

Word Processing

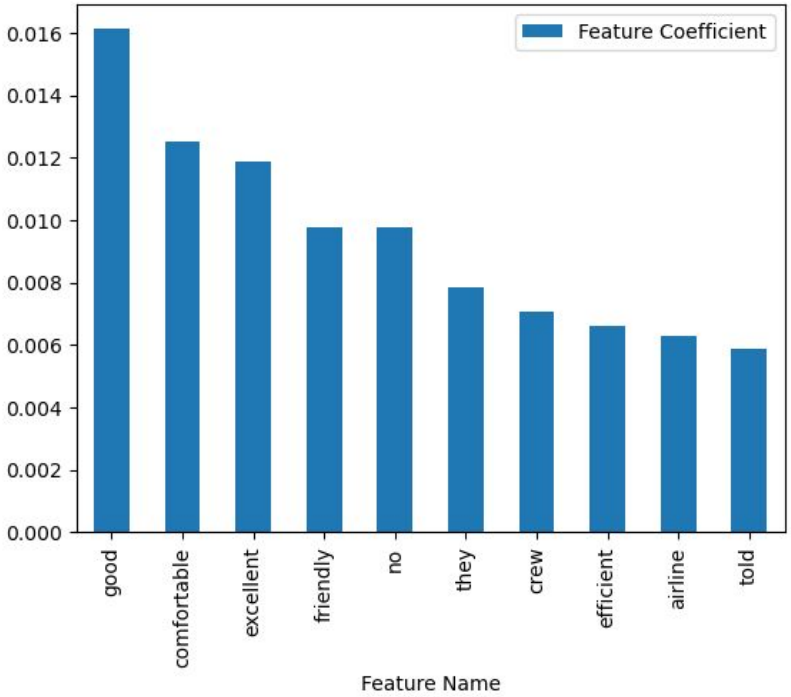
- We experimented using the words in the reviews to improve accuracy
- Each word in the reviews was made a feature, and then was set to true or false depending on whether or not the original review had that word
- We trained a **Random Forest classifier** with default hyperparameters and an 80% train set to make predictions based solely on the words in the reviews
- It achieved 84.5% accuracy

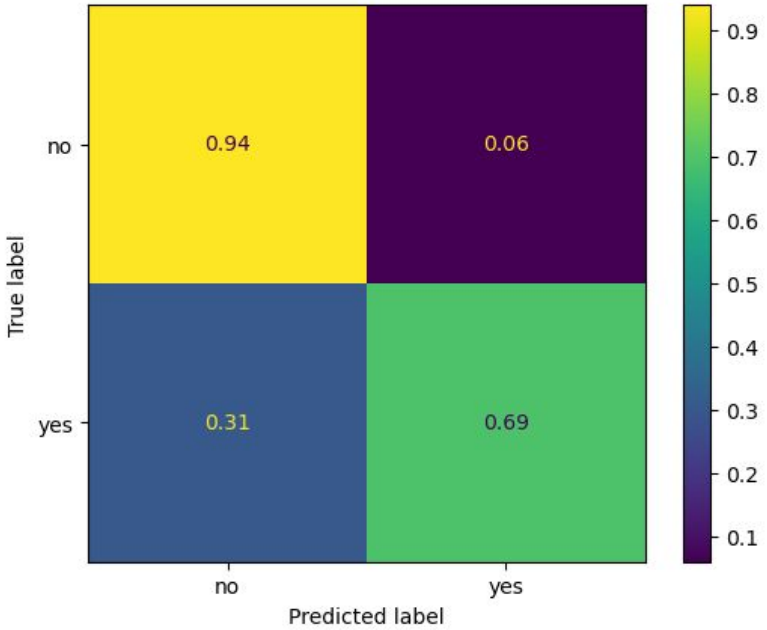
Other Preprocessing

- For some of our work, we preprocessed the Route feature
- It involved repairing broken data entries, and then splitting the Route into a list of countries by splitting on “to”
- The lists of countries were then turned into additional features

```
[11]: Aircraft Type            category
      Users Reviews              object
      Country                  category
      Type_of_Travellers       category
      Seat_Types               category
      Seat Comfort                Int64
      Date Flown               category
      Cabin Staff Service         Int64
      Ground Service              Int64
      Food & Beverages            Int64
      Wifi & Connectivity         Int64
      Inflight Entertainment      Int64
      Value For Money             int64
      Recommended              category
      country_0                category
      country_1                category
      country_2                category
      country_3                category
      dtype: object
```

Word Processing Random Forest Results



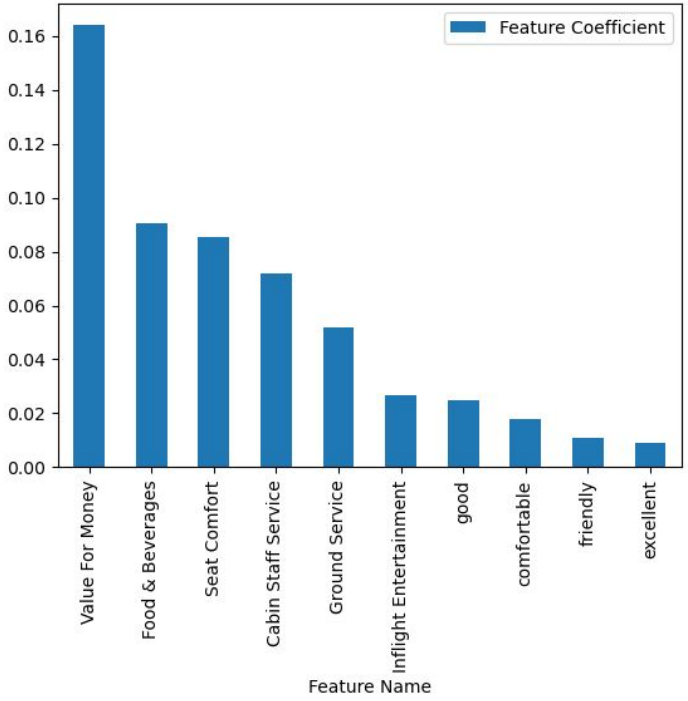


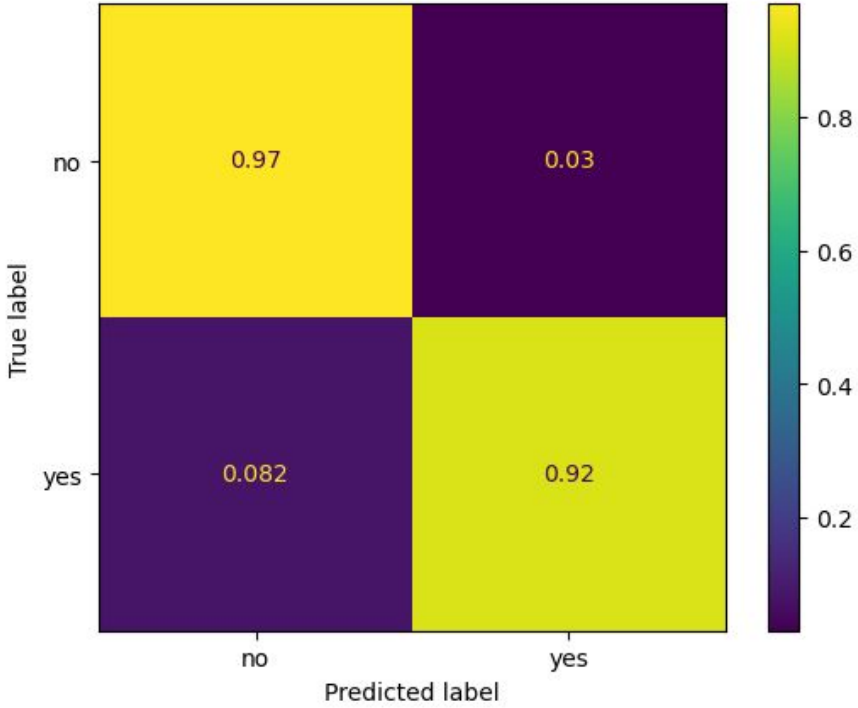
Full Dataset (Word processing and others)

- Used Random Forest, Boosting Gradient, and Logistic Regression classifiers to establish baselines
- Used L1 logistic regression to remove roughly 98% of features
- Used L2 logistic regression, a Random Forest, Gradient Boosting, a SVC, and a KNN classifier to classify the reduced feature set
- Hyperparameters for each are included above their results
- We used PCA to further reduce over ⅔ of the features for SVC and KNN.
- The following slides show the results of a single, random run of the algorithms.

Random Forest (Feature set with words)

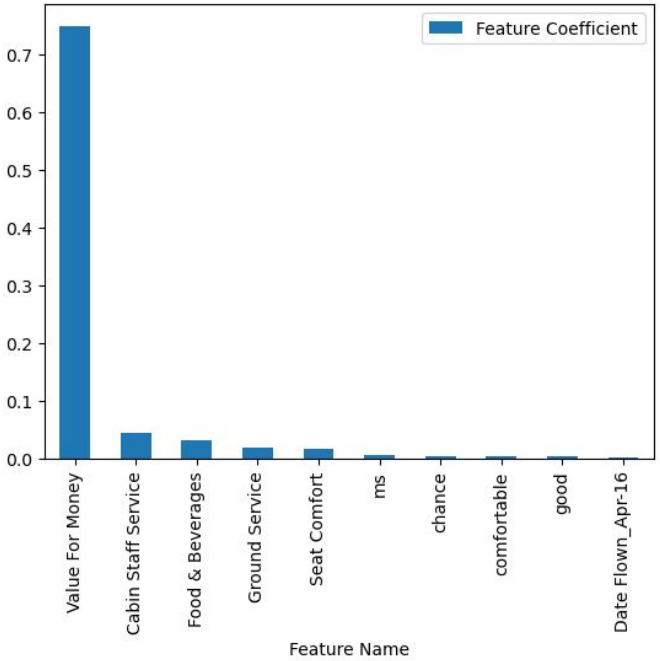
- Using 525 estimators

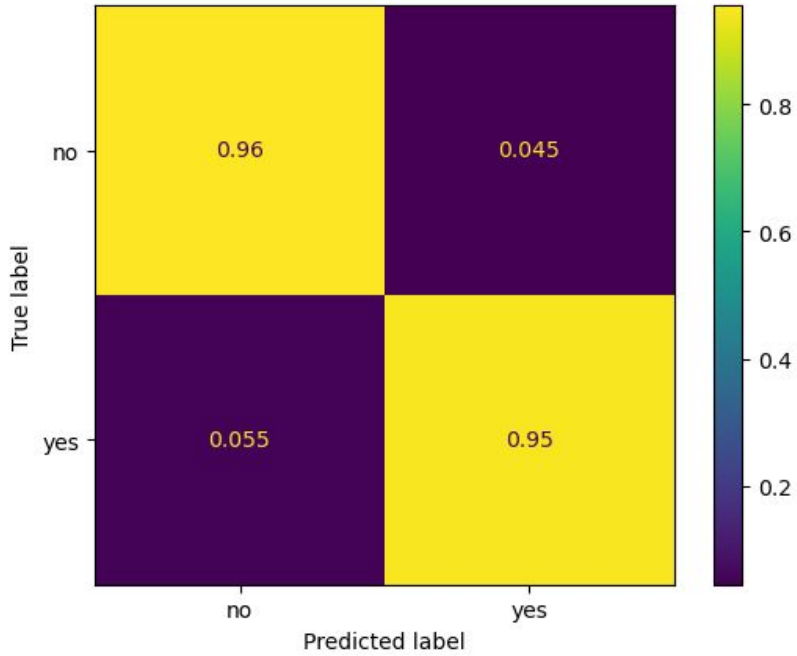




Gradient Boosting (Feature Set with Words)

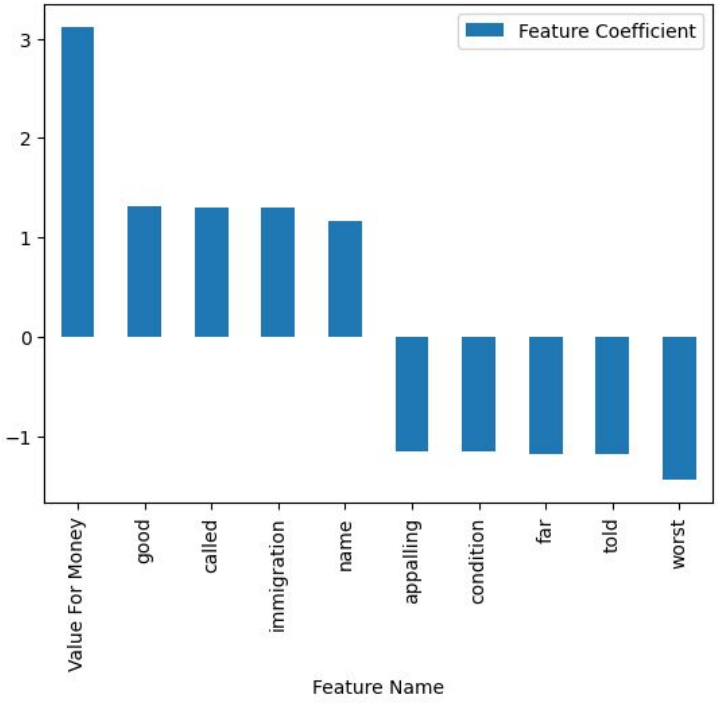
- Used ‘log_loss’, learning_rate=0.78, and n_estimators=935

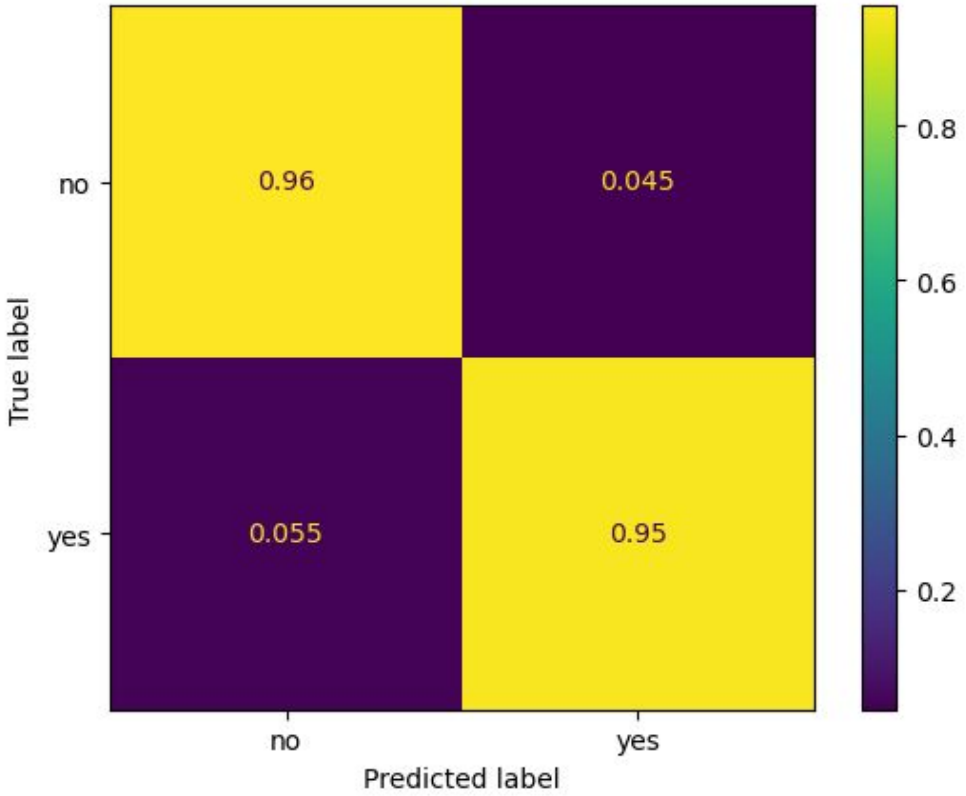




Logistic Regression

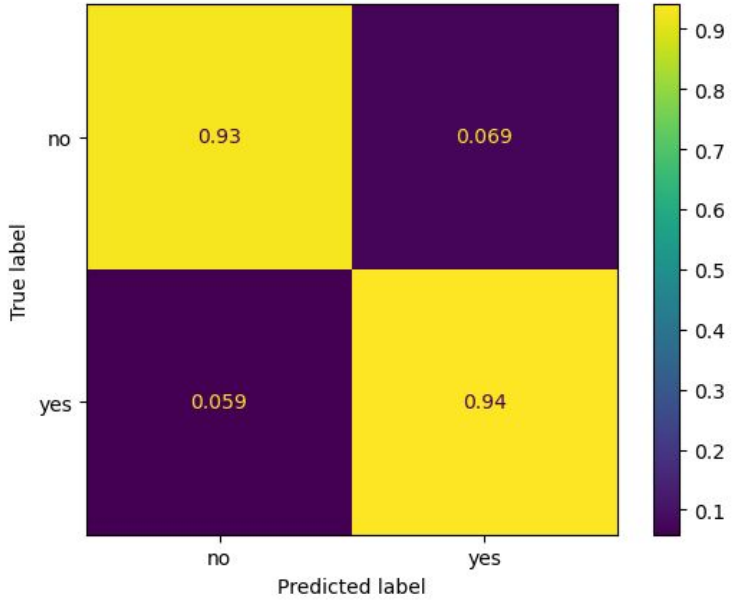
- With L2 tuning

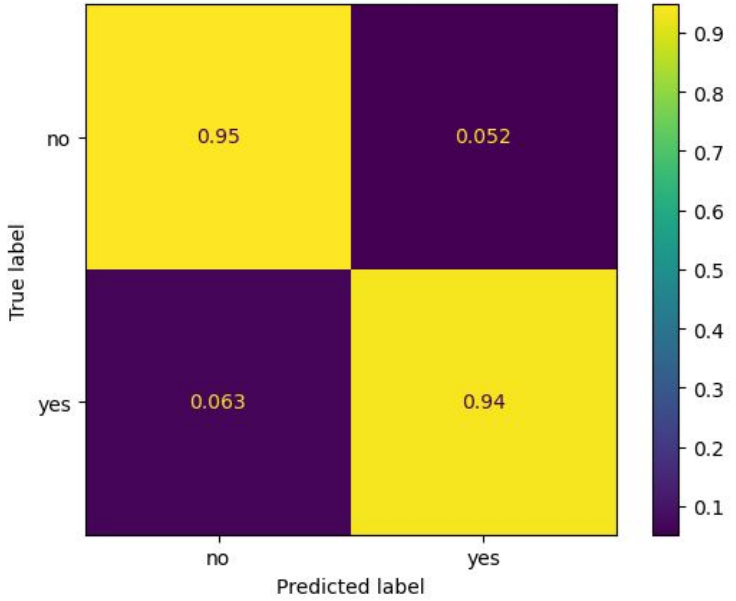




KNN and SVC Classifiers

- KNN used 14 neighbors with weights based on euclidean distance (p=2)
- SVC used gamma=1 / (n_features * X.var()), kernel=rbf, and C=2.

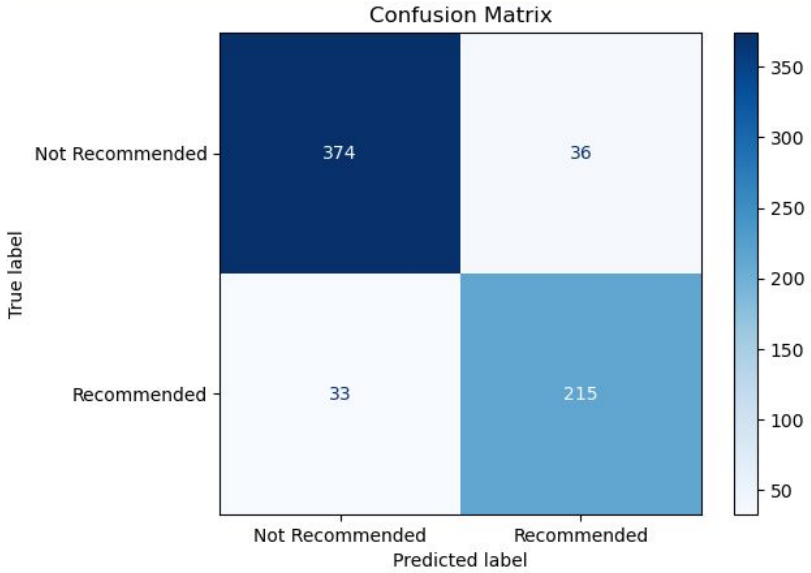




Algorithms without Word processing

KNN Algorithm: 92%

Decision Tree Algorithm: 90%

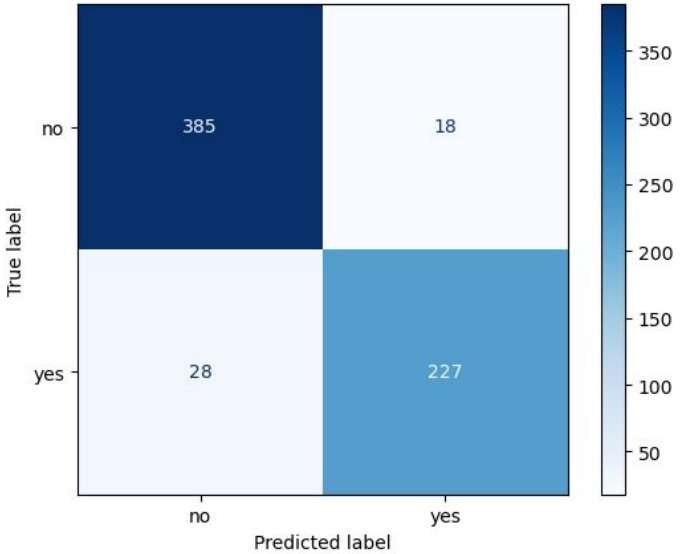


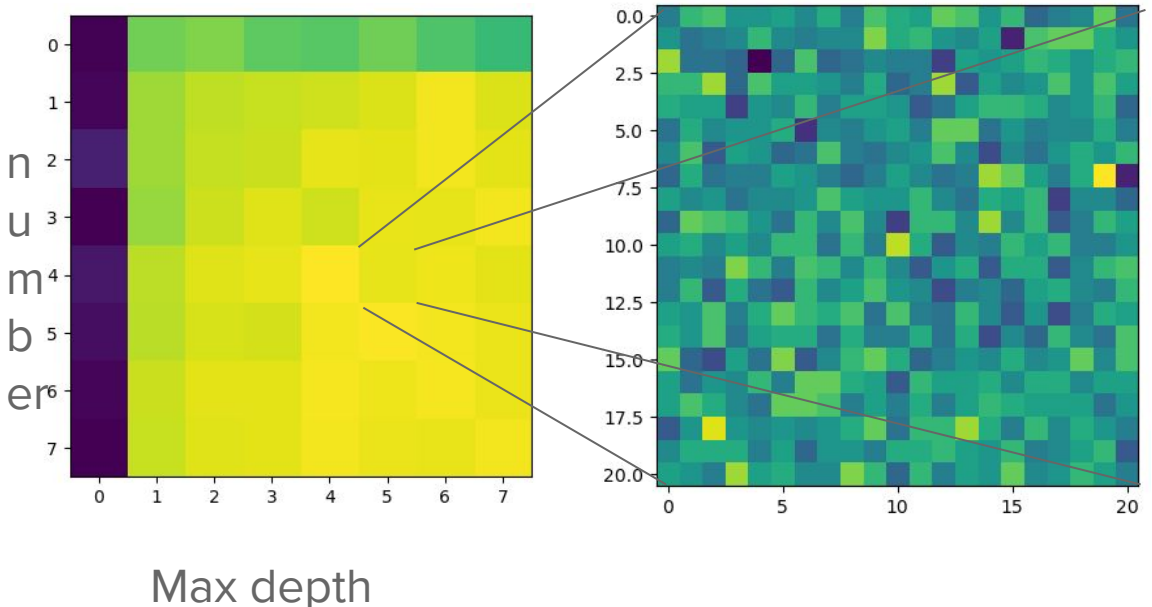
Algorithms without Word processing

Random forest, all one-hot encoded features without user reviews

38 estimators, 195 max depth

Best Accuracy=94.07%





Algorithms without Word processing

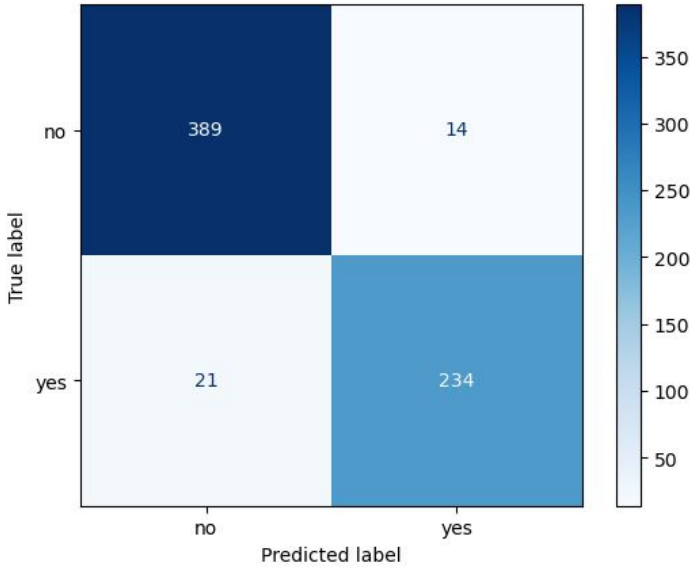
SVM , all one-hot encoded features without user reviews

C=1, gamma=0.24

Best Accuracy=93.92%

With only numerical value, fill empty with 0

Accuracy=94.68%



Algorithms without Word processing

SVM, fillna with 0, only with numerical features, incorrectly predicted entries.

Seat Comfort	Date Flown	Cabin Staff Service	Ground Service	Food & Beverages	Wifi & Connectivity	Inflight Entertainment	Value For Money	Recommended
4.0	Jan-16	4.0	1.0	4.0	0.0	0.0	4	no
4.0	Oct-18	3.0	3.0	2.0	1.0	2.0	3	yes
2.0	Jun-15	3.0	3.0	2.0	0.0	3.0	3	yes

Best Accuracy Results

- Between 93% and 95% accuracy on runs using the following algorithms:
  - Random Forest Classifier
  - Gradient Boosting Classifier
  - L2 Logistic Regression

Demo

- False prediction review examples (from L2 Logistic Regression with words)

	Aircraft Type	Users Reviews	Country	Type_of_Travellers	Seat_Types	Seat Comfort	Date Flown	Cabin Staff Service	Ground Service	Food & Beverages	Wifi & Connectivity	Inflight Entertainment	Value For Money	Recommended
1404	A320	✅Verified Review| Flew Leeds Bradford to Gen...	New Zealand	Solo Leisure	Economy Class	4	Oct-17	4	4	2	<NA>	<NA>	2	yes
2625	Boeing 777-300	We paid for World Traveller Plus (Premium Econ...	United States	Couple Leisure	Premium Economy	3	Sep-15	4	3	4	3	3	2	no
882	NaN	✅Trip Verified| London to Seville. Following ...	United Kingdom	Couple Leisure	Economy Class	<NA>	Mar-19	<NA>	4	<NA>	<NA>	<NA>	4	yes

3 Lessons/Obstacles

Long running time with large feature amount.

- “Liblinear” algorithm removed significant runtime

Word processing

- Process similar to One-Hot-Encoding

Simple features might yield better result

- L1 logistic regression and PCA

Q&A

Questions?