Group 7A
Sahir Doshi, Anna Vasconcelos, Yonghee Jeon

QUESTROM AIR

PREDICTING PASSENGER SATISFACTION

LET'S GET STARTED

AGENDA

1. Introduction and Dataset
2. Exploratory Data Analysis
3. Predictive Modeling
4. Challenges
5. Conclusions

Problem and Audience

Who We Are:

We are Boston's newest airline: Questrom Air

Customer Satisfaction is our #1 Priority

Goals:

1) Deduce the most important features impacting customer satisfaction so we can optimize our airline's experience

2) Construct a model that predicts satisfaction accuractely

DATASET DESCRIPTION

Overview & Features

Dataset Description and Cleaning

25,893 Rows

18 Numerical Predictor Variables:

Age
Flight Distance
Inflight Wifi Service*
Departure/Arrival Time Convenience*
Ease of Online Booking*
Gate Location*
Food and Drink*
Online Boarding*
Seat Comfort*
Inflight Entertainment*
On-Board Service*
Leg Room Service*
Baggage Handling*
Check-In Service*
Inflight Service*
Cleanliness*
Departure Delay in Minutes
Arrival Delay in Minutes

*Customer Satisfaction Rating (0 = N/A, 1 – 5)

6 Categorical Predictor Variables:

Gender
Customer Type
Type of Travel
Economy Class (dummy coded)
Economy Plus Class (dummy coded)
Business Class (dummy coded)

Target Variable:
Guest Satisfaction

Target Variable Outcomes:
Neutral/Dissatisfied
Satisfied

Dataset originally from Kaggle

Dataset Description and Cleaning

Dummy Coding Catagorical Variables

Gender (1 = Male / 0 = Female)
Customer Type (1 = Loyal Customer / 0 = Disloyal Customer)
Type of Travel (1 = Business Travel / 0 = Personal Travel)
Satisfaction (1 = Satisfied / 0 = Neutral or Dissatisfied)

Class_Business (1 = True / 0 = False)
Class_Eco (1 = True / 0 = False)
Class_Eco Plus (1 = True / 0 = False)

Dropped:

ID

Unnamed 0

EXPLORATORY DATA ANALYSIS

Initial Observations & Patterns

Correlation Heat Map

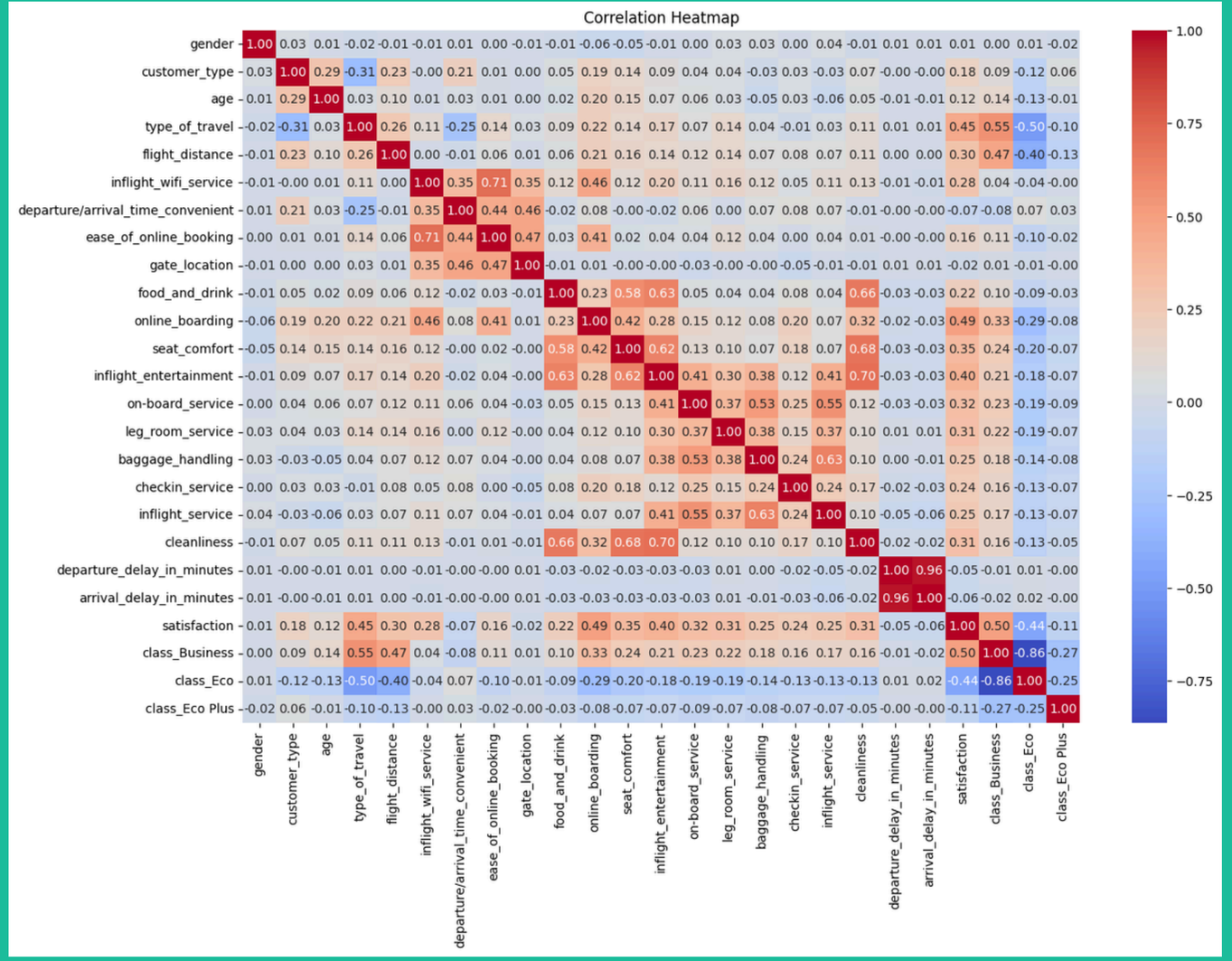
Key Takeaways

1) Positive Correlations with Satisfaction

2) Negative Correlations with Delays

3) Multicollinearity

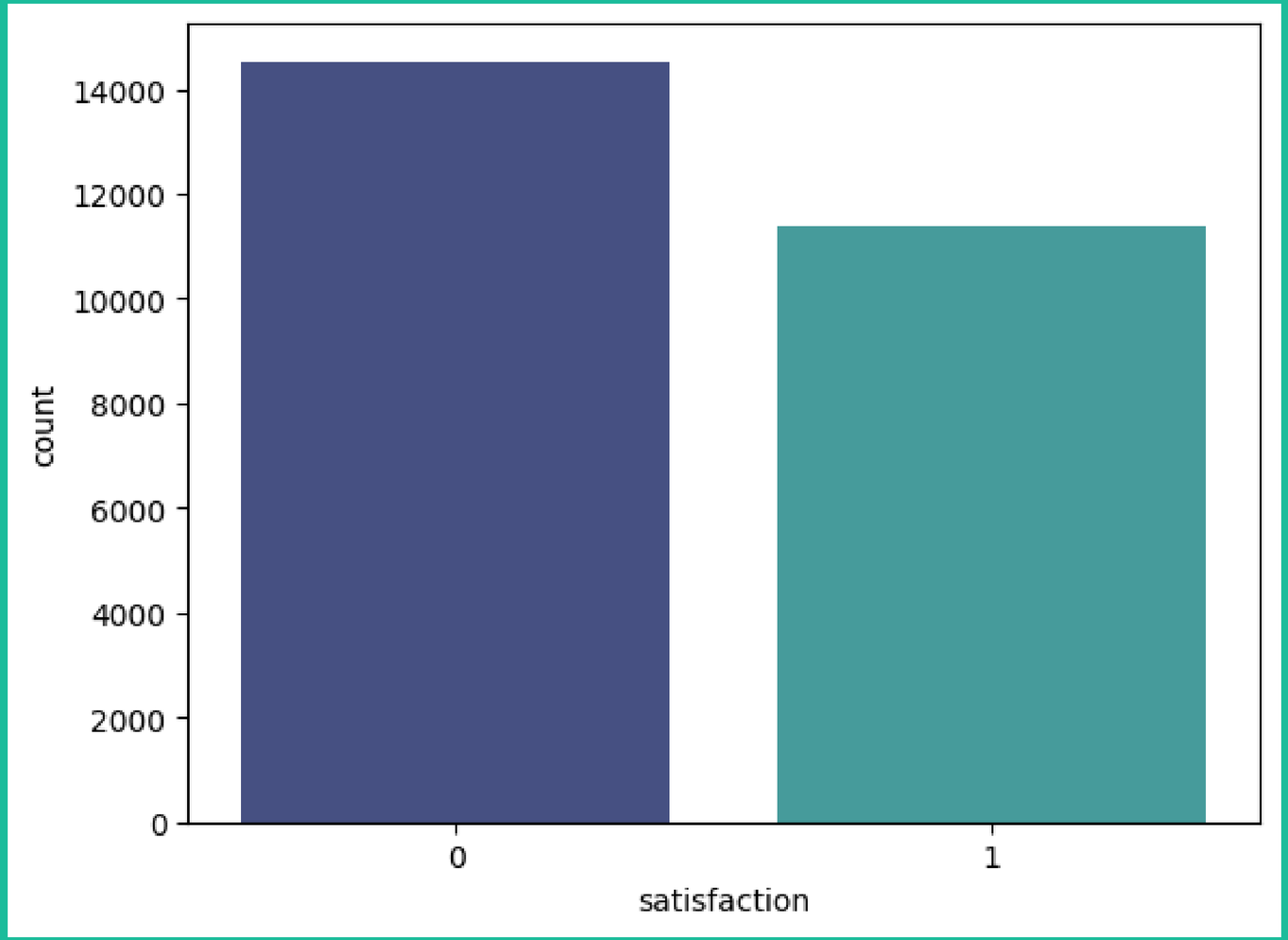
4) Weak Correlations for Demographic Variables



Target Variable Breakdown

Key Takeaways

On average, customers are <u>not</u> satisfied with their airline experience



PREDICTIVE MODELING

<u>Classification Models:</u>

How well can our models predict whether a customer is satisfied?
What factors influence satisfaction?
How well can our models generalize?

Overview of All Models Tested

Naive Rule - Classify all as majority class - 56% accuracy

K - Nearest Neighbors

Logistic Regression

- Default
- Ridge with GridSearch CV
- Lasso with GridSearch CV

Decision Trees

- Default
- Tuned with Cost Complexity Alpha
- Tuned with Cost Complexity Alpha and Bagging, 10 Fold CV
- Tuned with random hyperparameters, 10 Fold CV
- Tuned with GridSearchCV hyperparameters, 10 Fold CV

Random Forest

- Default
- Tuned with GridSearch CV, 10 Fold CV

Neural Network

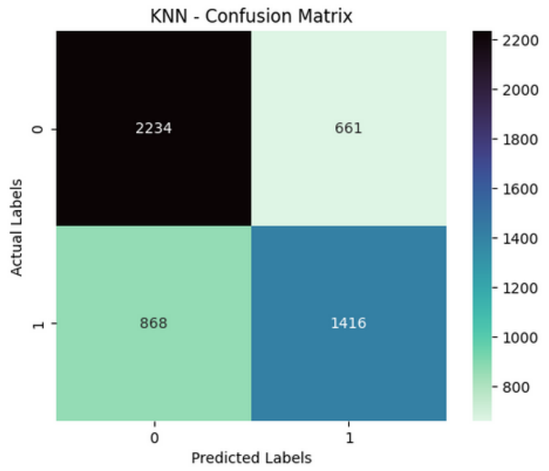
Stacking Classifier

- combines Random Forest with KNN and DT

K-Nearest Neighbors/Logistic Regression

K-Nearest Neighbors

Best k: k = 9



Training Accuracy: 0.77

Testing Accuracy: 0.70

Logistic Default

Training Accuracy: 0.81

Testing Accuracy: 0.81

Logistic Ridge Regression GridSearchCV

Training Accuracy: 0.87

Testing Accuracy: 0.87

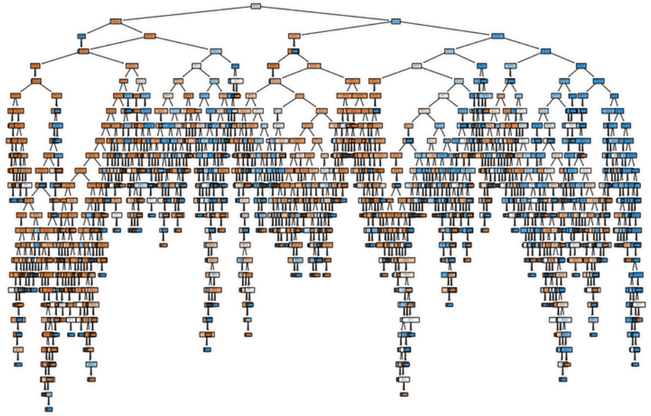
Logistic Lasso Regression GridSearchCV

Training Accuracy: 0.87

Testing Accuracy: 0.87

Decision Trees

Full Decision Tree Unpruned

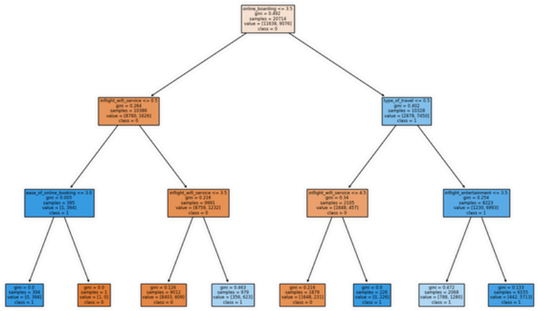


Training Accuracy:
1.0

Testing Accuracy:
0.93

***Overfits!**

Pruned DT Random Hyperparameters 10 Fold CV

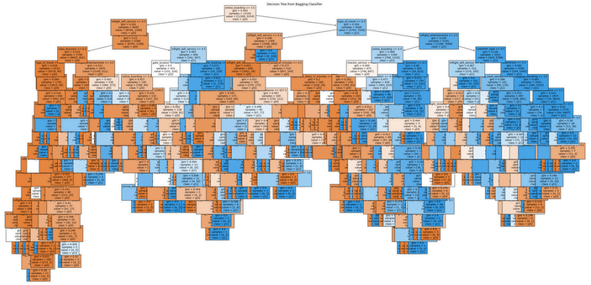


Training Accuracy:
0.88

Testing Accuracy:
0.88

Pruned DT CCP Alpha Pruning

Alpha = 0.0001

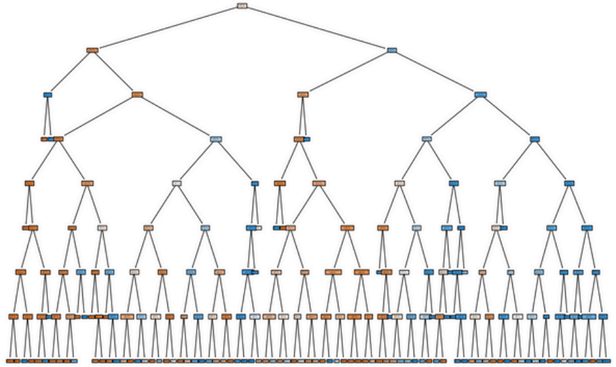


Training Accuracy:
0.97

Testing Accuracy:
0.94

Pruned DT GridSearch CV Hyperparameters 10 Fold CV

Best Parameters: {'max_depth': 8, 'min_samples_split': 8}



Training Accuracy:
0.94

Testing Accuracy:
0.93

Bagging Classifier

CCP Alpha Pruned Decision Tree as the base

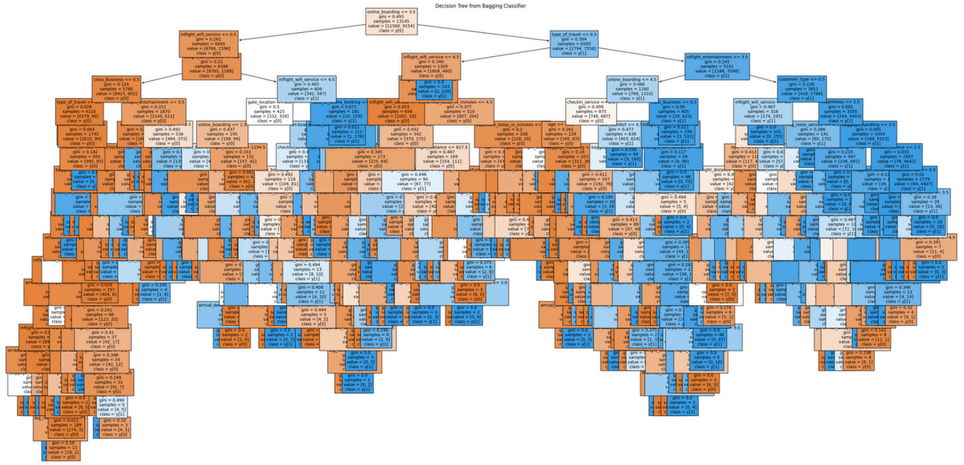
- **Added Cross Validation of 10 folds**

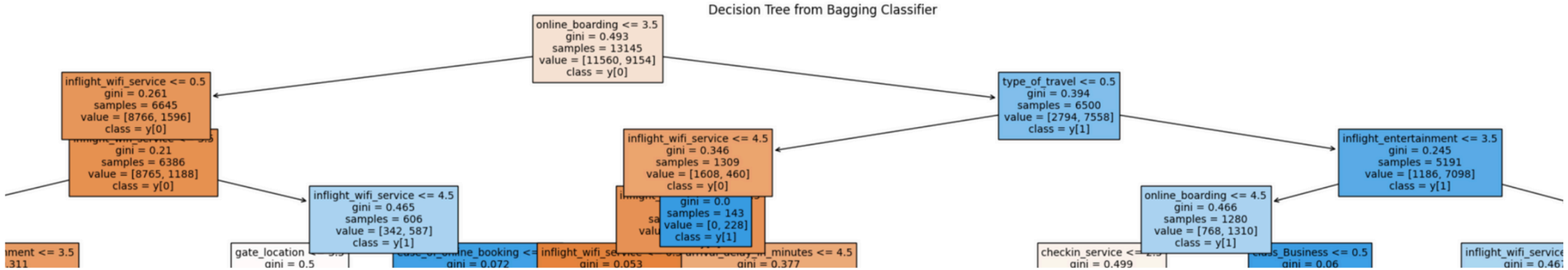
Training Accuracy:
0.98

Testing Accuracy:
0.95

Most Important Features:

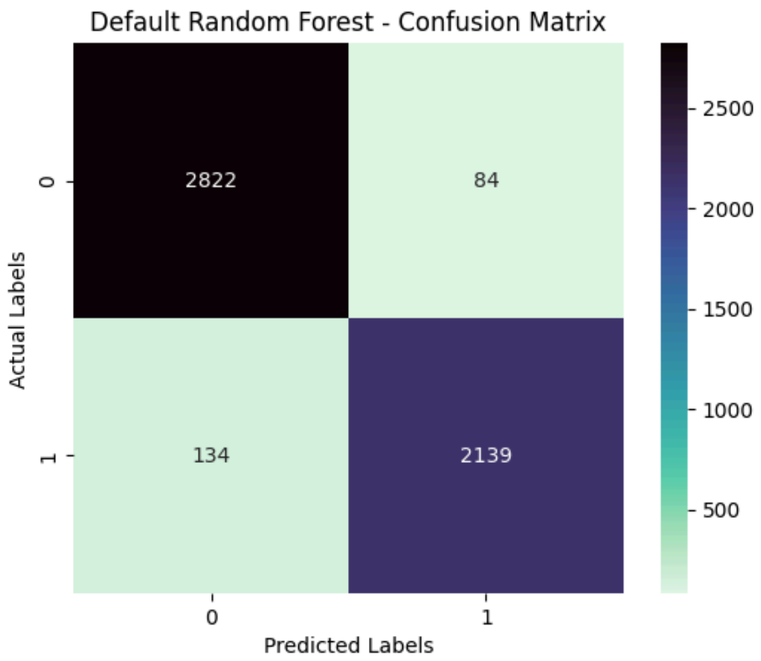
- **online boarding**
- **inflight wifi**
- **type of travel**
- **inflight entertainment**





Random Forests

Random Forest Default



Training Accuracy: 1.0

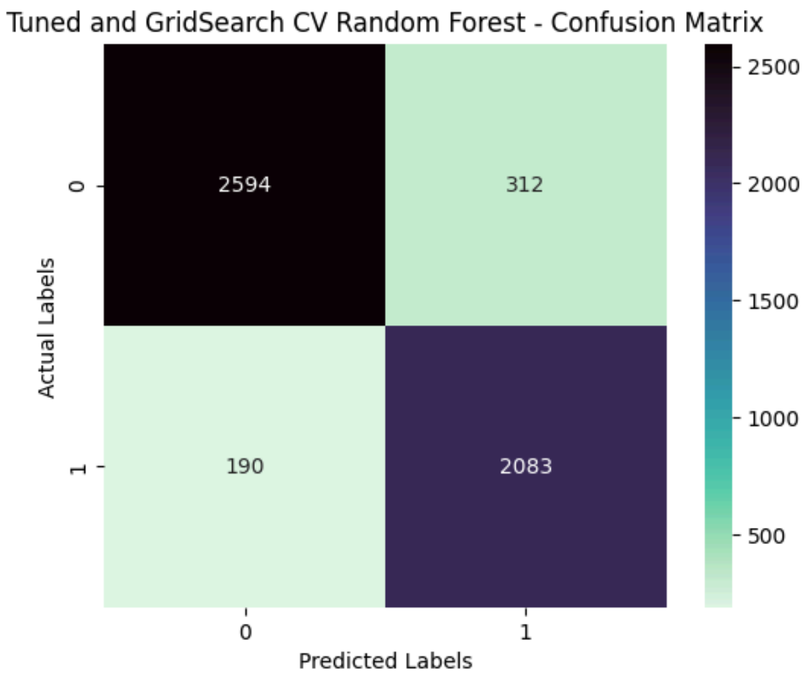
Testing Accuracy: 0.96

***Overfits!**

Tuned Random Forest GridSearch CV

{'max_depth': 6, 'max_features': None, 'max_leaf_nodes': 9, 'n_estimators': 100}

10 Fold CV in Grid

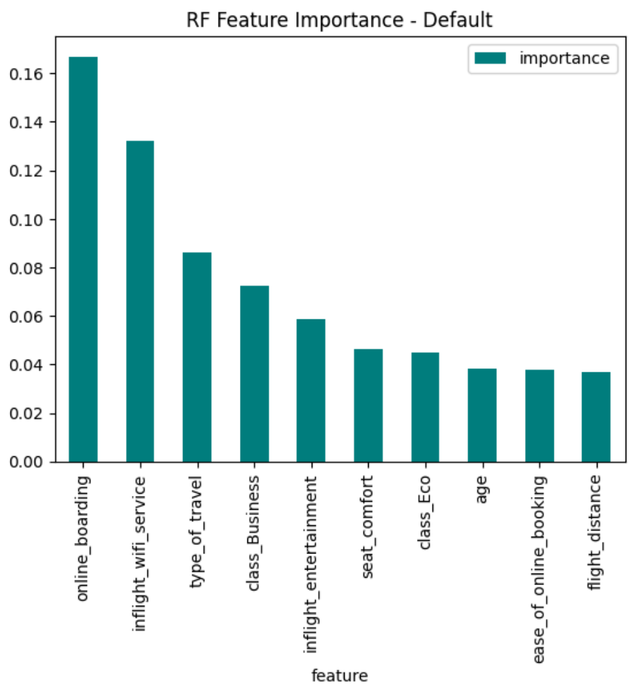


Training Accuracy: 0.90

Testing Accuracy: 0.90

Random Forests Feature Importances

Random Forest
Default



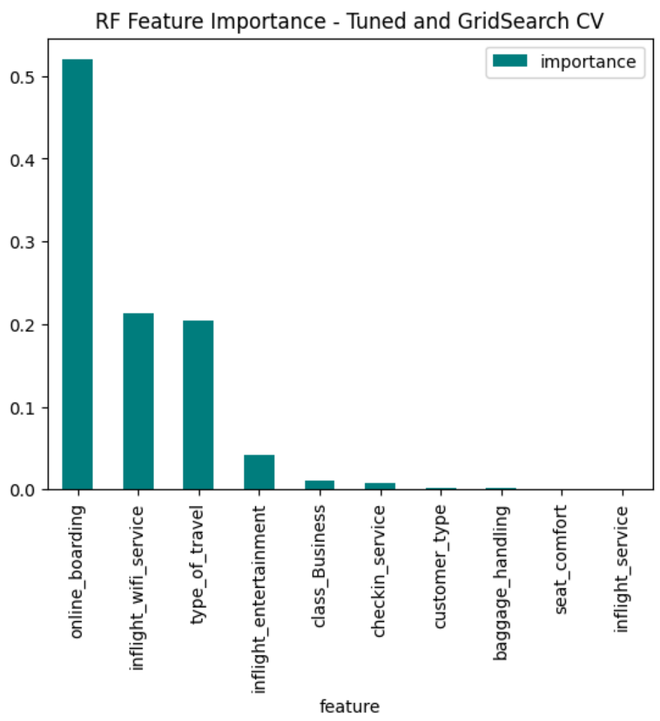
Most Important Features:

- online boarding
- inflight wifi
- type of travel
- business class

Training Accuracy: 1.0

Testing Accuracy: 0.96

Tuned Random Forest
GridSearch CV
10 Fold CV in Grid



Most Important Features:

- online boarding
- inflight wifi
- type of travel
- inflight entertainment

Training Accuracy: 0.90

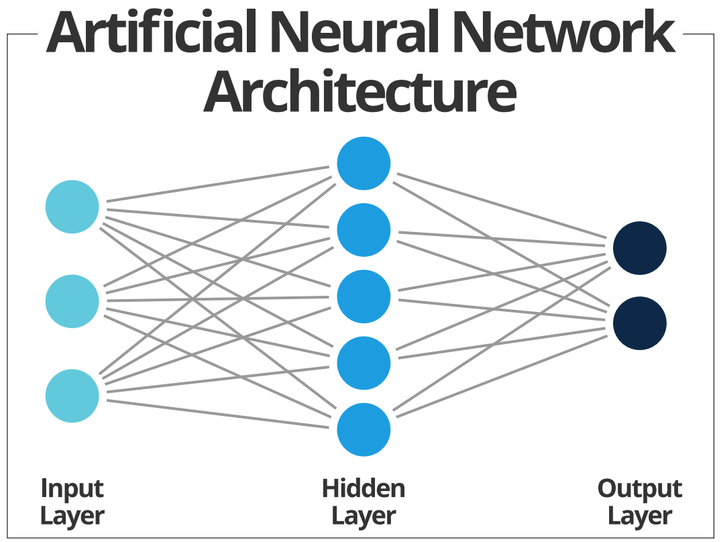
Testing Accuracy: 0.90

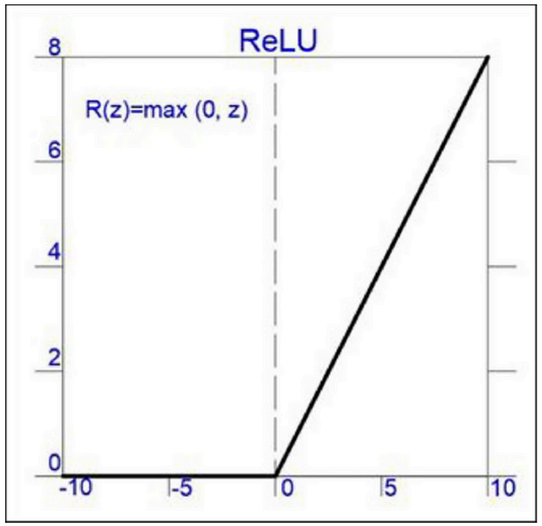
Neural Network

Loop through hidden layer sizes 5, 10, 15

Activation Function: 'ReLu'

Best Hidden Layer Size: 15



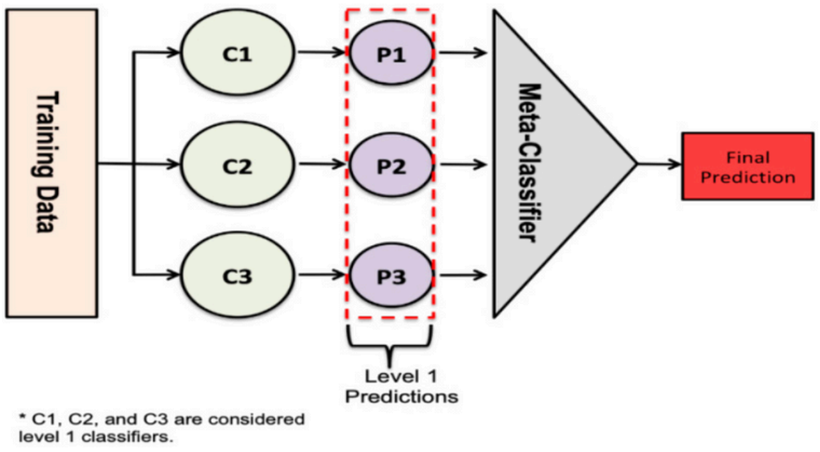


Training Accuracy:
0.87

Testing Accuracy:
0.87

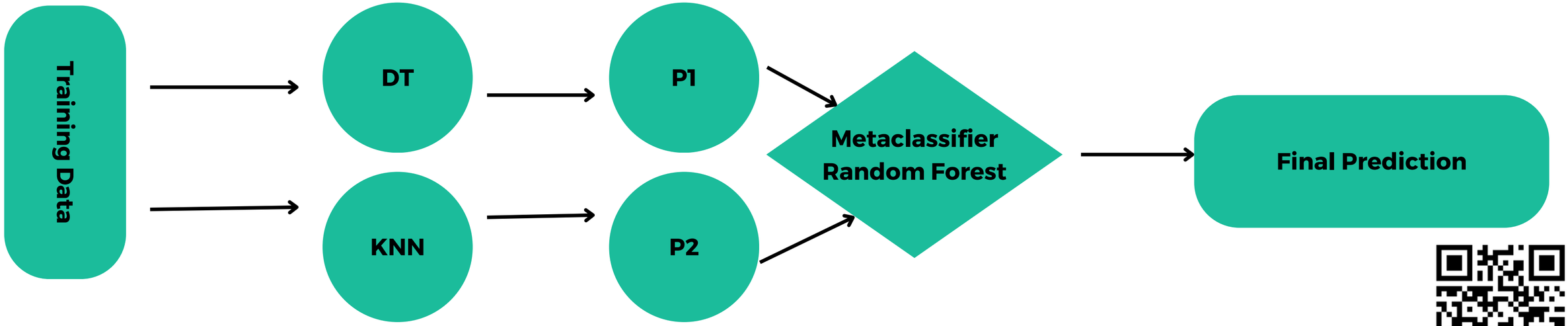
Photo Credits:
https://www.researchgate.net/figure/Activation-function-ReLu-ReLu-Rectified-Linear-Activation_fig2_370465617
https://www.smartsheet.com/neural-network-applications

Stacking Classifier



Training Accuracy:
0.96

Testing Accuracy:
0.94



More about stacking classifiers if interested:
https://tinyurl.com/3963fb74

Naive Rule:

0.56

Model	Training Accuracy	Test Accuracy
K-NN (K = 9)	0.77	0.70
Decision Tree (Full)	1.0	0.93
DT Tuned /w CCP Alpha	0.97	0.94
DT Tuned /w CCP Alpha & Bagging/10 Fold CV	0.98	0.95
DT Tuned /w Random Hyper Parameters, 10 Fold CV	0.88	0.88
DT Ideal /w Tuned Grid search CV/10 Fold CV	0.94	0.93

Naive Rule:

0.56

Model	Training Accuracy	Test Accuracy
Logistic Regression	0.81	0.81
LR Ridge Gridsearch CV	0.87	0.87
LR Lasso Gridsearch CV	0.87	0.87
Random Forest	1.0	0.96
RF /w Tuned CV	0.90	0.90
Neural Networks	0.87	0.87
Stacked Classifier - bases were DT and KNN - meta was RF	0.96	0.94

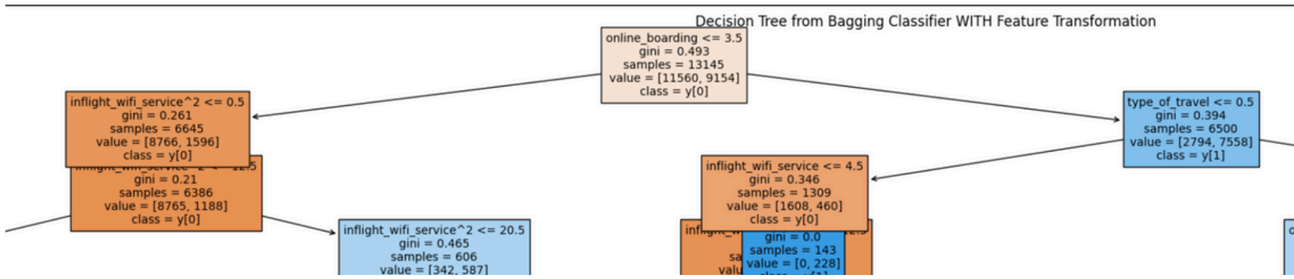
Feature Transformation

Inflight Wifi Service high value of satisfaction is overall the 2nd most important feature in DT and RF

What would be the effect of creating a column called “inflight_wifi_service^2” which is inflight_wifi squared?

Would it then become the most important feature?

Run the DT with Ideal CCP and Bagging



Having a satisfactory Online Boarding experience is still the most explanatory variable to guest satisfaction!

```
df['inflight_wifi_service^2'] = df['inflight_wifi_service'] ** 2
df
```

Final Models Chosen

For Accuracy:

Decision Tree Bagging Classifier
Tuned with BEST CCP Alpha
10 Fold CV

Training: 0.98
Testing: 0.95

For Interpretability:

Decision Tree
Arbitrary Hyperparameters
10 Fold CV

Training: 0.88
Testing: 0.88

For Generalizability:

Decision Tree
Ideal Hyperparameters
GridSearchCV
10 Fold CV

Training: 0.94
Testing: 0.93

CONCLUSIONS

Next steps

Recommendations

From our Feature Importance Model it recommend that for new airline companies focus on prioritizing these features as they contribute to a good review in a passengers airline experience:

Online Boarding: Passengers prefer having the ease of checking in and boarding via their phone

Inflight Wifi: Who doesn't want wifi when on a plane

Ease of online booking: Having a easy to use booking site can be greatly beneficial for customers

In Flight Entertainment: This can appeal to all ages and give more options in entertainment having both wifi and in flight entertainment

Challenges and Considerations

Challenges

- RF bagging + tuned + cv no matter what parameters were changed still had testing accuracy > training
- Neural network was more simple than it should be
- Most intensive part was the predictive modeling part because data was pretty clean when we got it
- Computationally intensive computers would help modeling

THANK YOU!

Any Questions?