United Flight Satisfaction Prediction

Team Members

Michael Crosson

Saaket Joshi

Luke Leon

Quinlan O’Connell

Austin Yeh

Table of Contents

Data Context

Data Context

Dataset Focus: *Predict if a customer had a satisfactory flight experience*

Passenger Age

Average Age: ~ 40
Highest Age: 85
Lowest Age: 7

Flight Distance

Average Distance: 1191 miles
Farthest: 4983 miles
Shortest: 31 miles

Passenger Class

Business Class: 11,462 people
Eco Plus: 1,752 people
Economy: 10,765 people

14 other flight service metrics were ranked from 0-5 by respondents

~ 43% of passengers were classified as "satisfied", while ~ 57% of passengers were classified as "neutral/dissatisfied."

Data Preparation

UNITED

Missing Values

Removed 80 NA's in Arrival.Delay.in.Minutes

Categorical Predictors

One-hot encoded nominal categorical variables like Gender, Travel Class, etc.

Data Dictionary

- Customer Satisfaction - Binary

Binary Variables

- Sex
- Loyal Customer
- Business Travel
- Business Class
- Economy Plus Class

Numerical Variables

- Age
- Flight Distance
- Departure Delay (mins)
- Arrival Delay (mins)

Range Variables (0 - 5 scale)

- Inflight Wifi Service
- Convenience of Arrival Time
- Ease of Online Booking
- Gate Location
- Food and Drink
- Online Boarding
- Seat Comfort
- Inflight Entertainment
- Onboard Service
- Leg Room Service
- Baggage Handling
- Check in Service
- Inflight Service
- Cleanliness

Background Info

Market Size

17.8%	17.4%	17.3%	16.0%
Delta	Southwest	American	United

Airline Domestic Market Share May 2023 - April 2024: https://www.transtats.bts.gov/

Problem

United Airlines continues to funnel a variety of passengers through their planes. Customers each have a different satisfaction level, depending on how they view the experience.

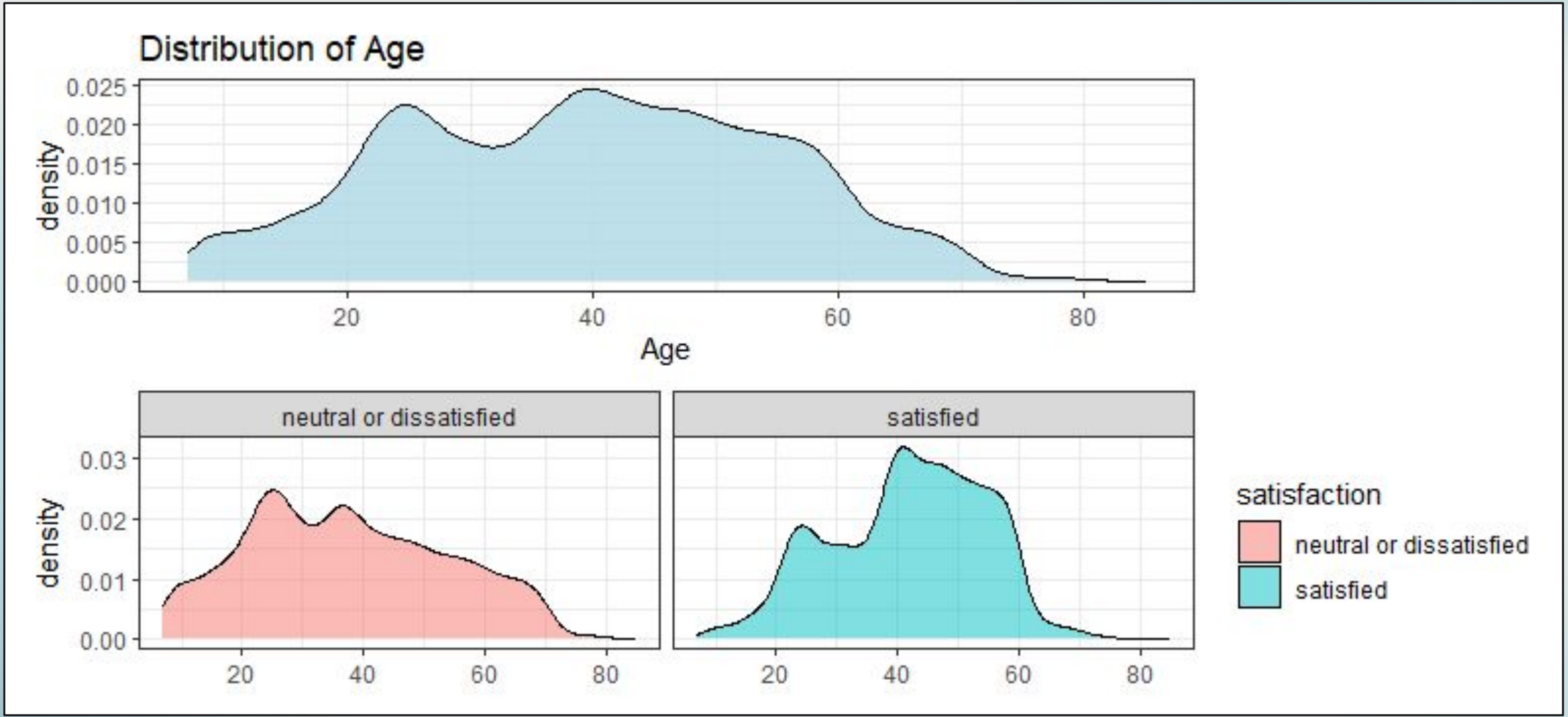
How can they make sure this experience is consistent across the company?

Question

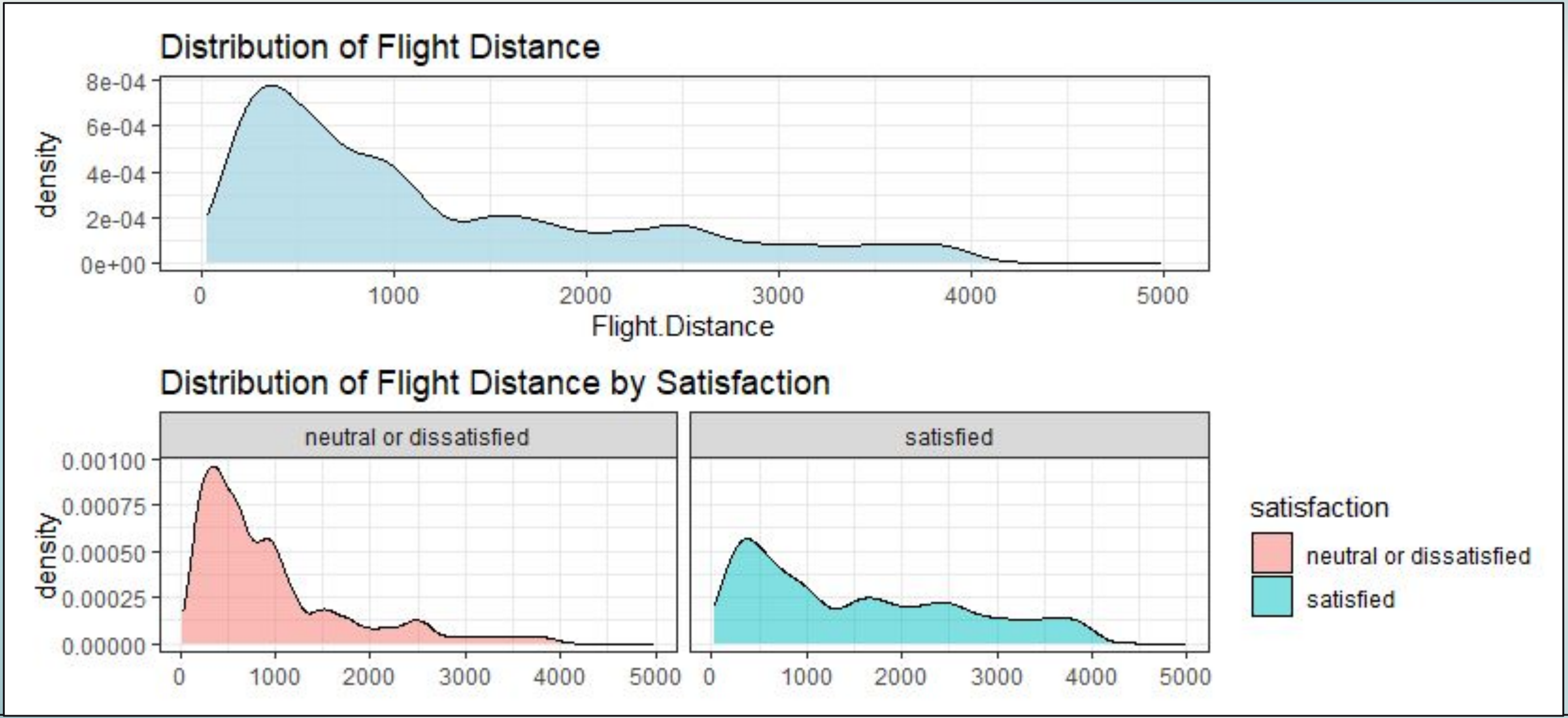
Can we predict customer satisfaction on United Airline Flights?

Exploratory Data Analysis

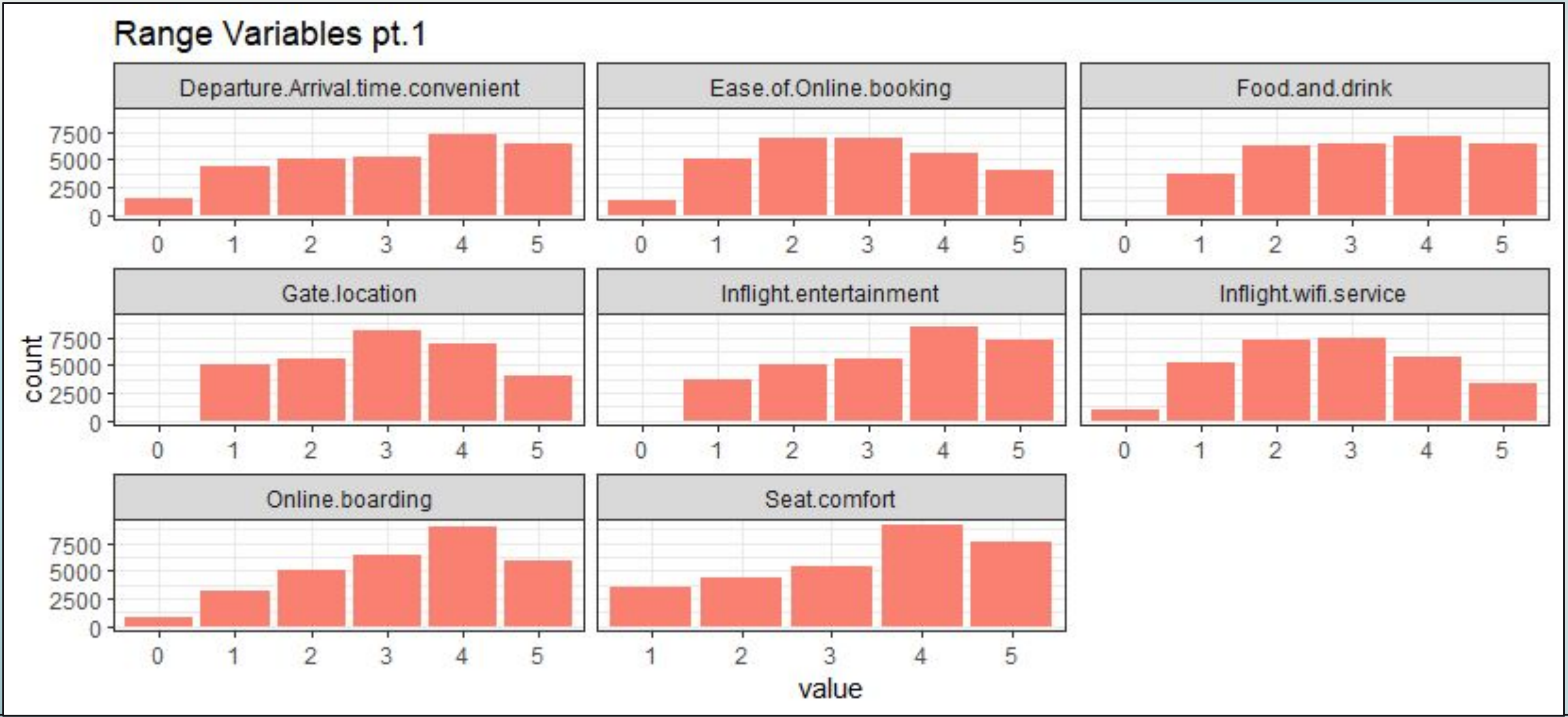
Continuous (Age)



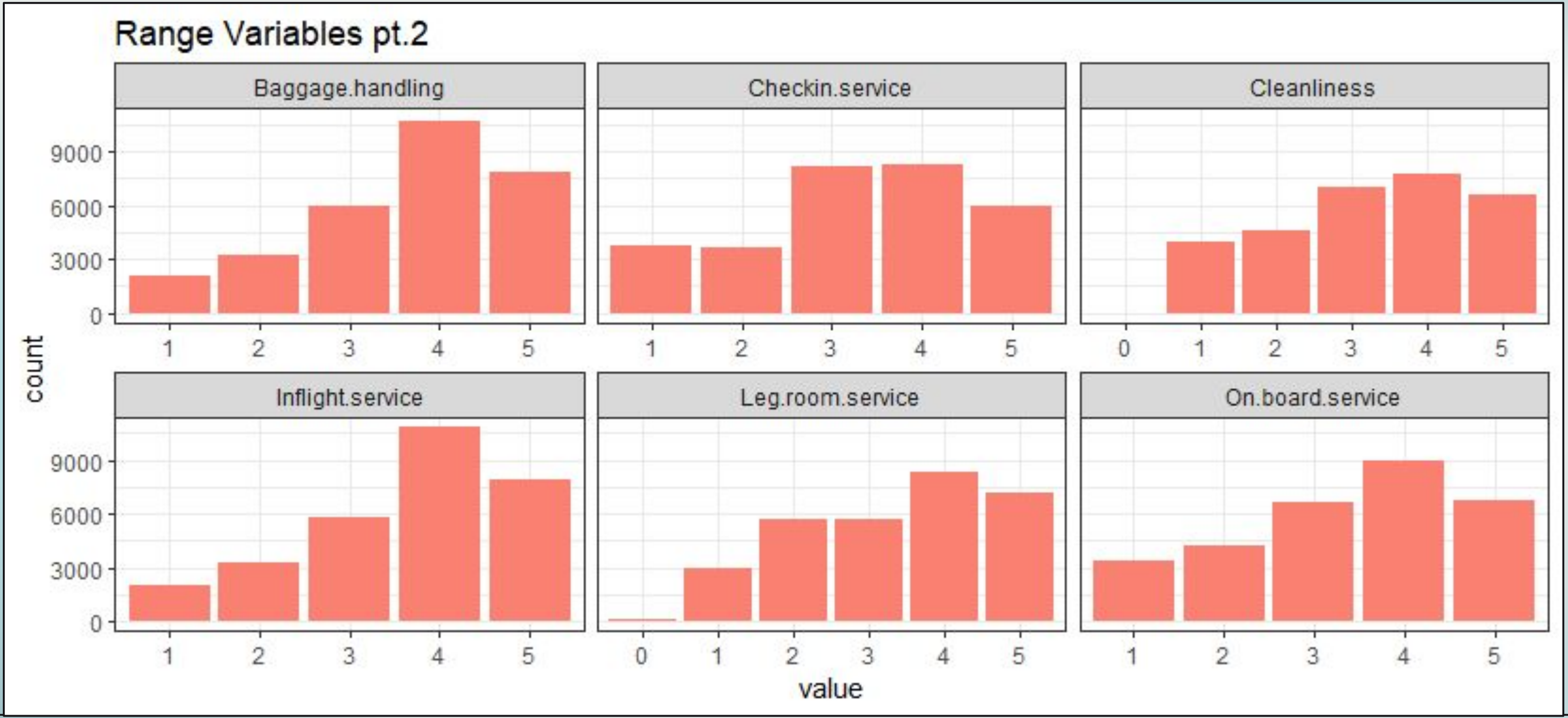
Continuous (Flight Distance)



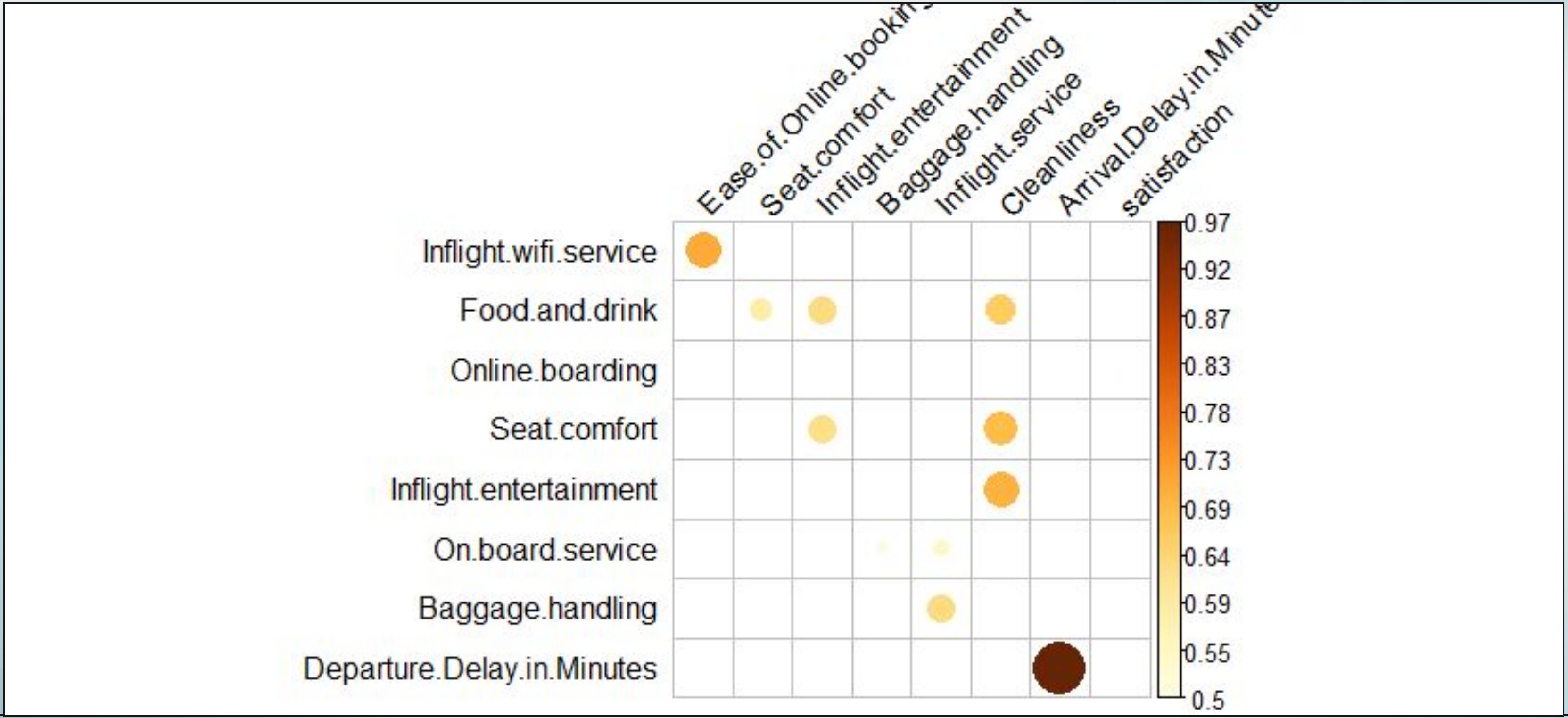
Range Variables PT. 1



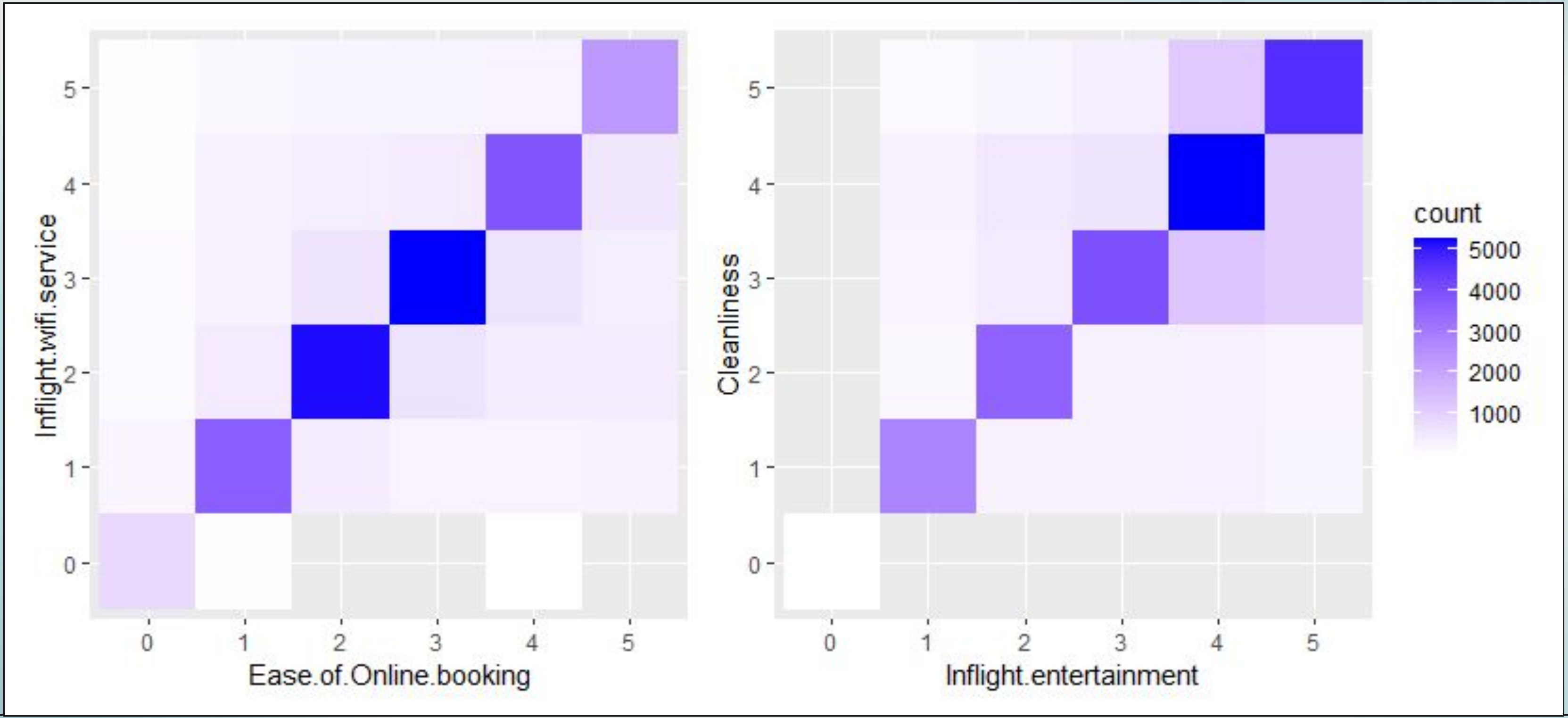
Range Variables PT. 2



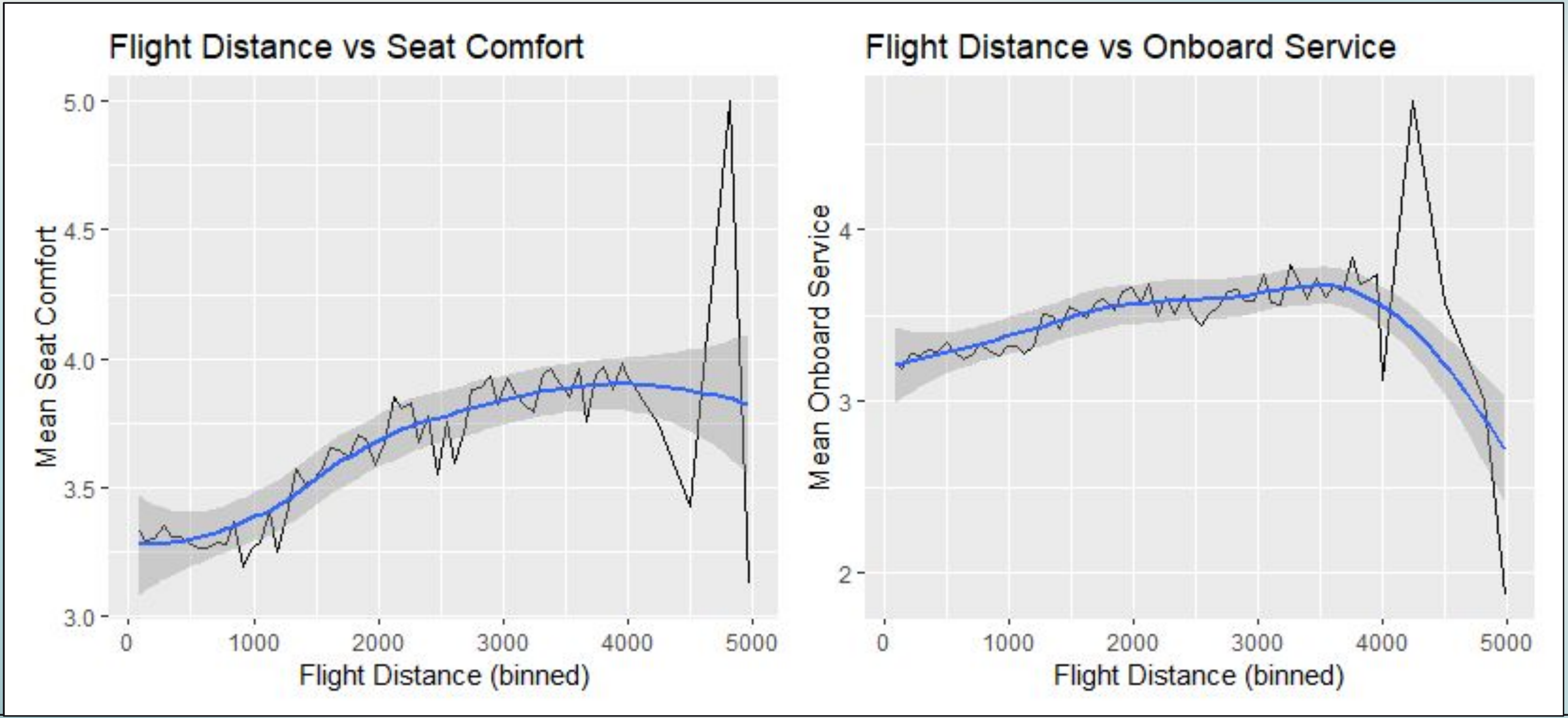
Corr. Matrix (Filtered)



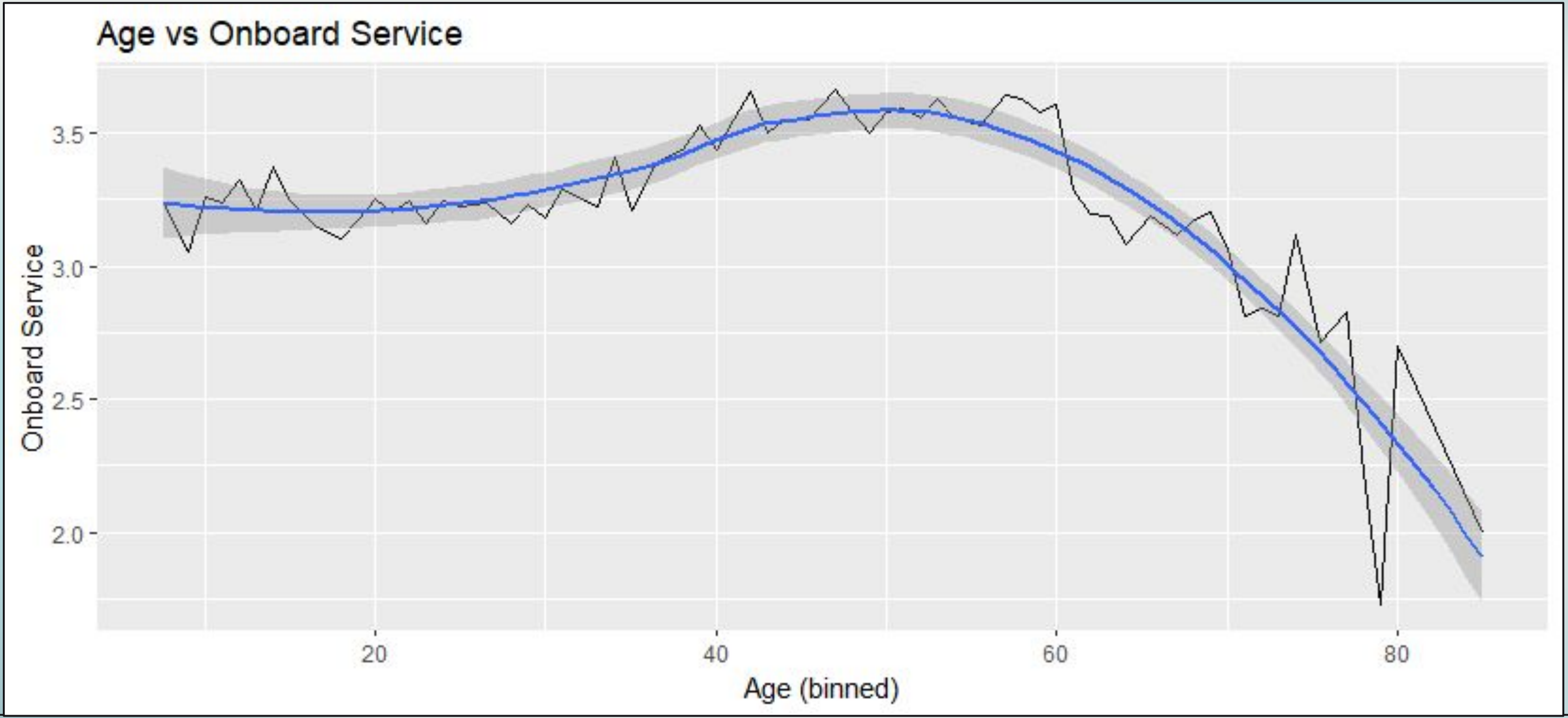
Highly Correlated Predictors



Highly Correlated Predictors



Highly Correlated Predictors



Statistical Learning Models

Models Used

KNN

Logistic Regression

Single Tree

Boosting

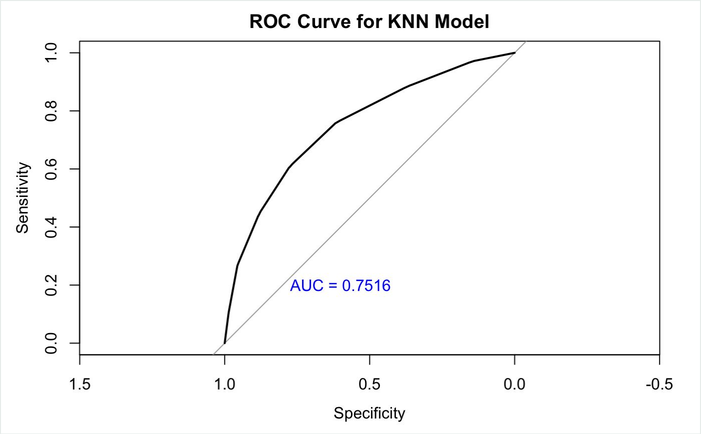
BART

Bagging

KNN Model

Key considerations

- 10-Fold Validation on K from 5 to 13
  - Optimal K was 7
- Lowest AUC & accuracy → established as baseline model



Actual

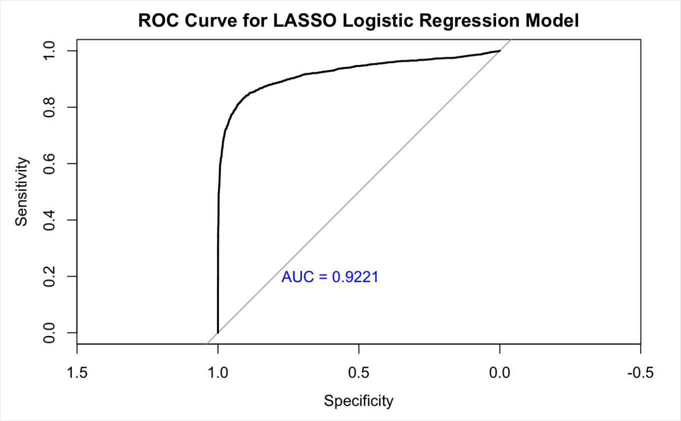
Predicted

	Not Satisfied	Satisfied	
Not Satisfied	2601	1037	
Satisfied	749	1593	Accuracy 70.13%

Logistic Regression Model

Key considerations

- Use of LASSO
  - Optimal alpha was 0.8
  - Optimal lamba 0.003268
- Trained on full model



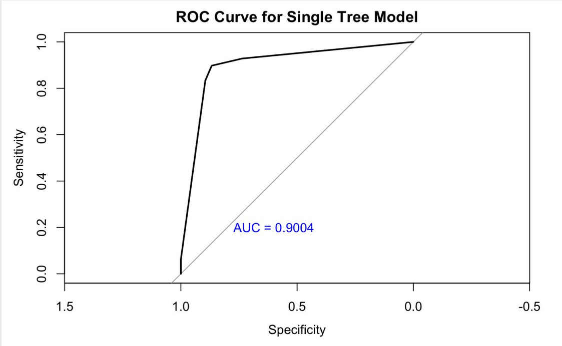
		Actual		
		Not Satisfied	Satisfied	
Predicted	Not Satisfied	3060	456	
	Satisfied	290	2174	Accuracy 87.53%

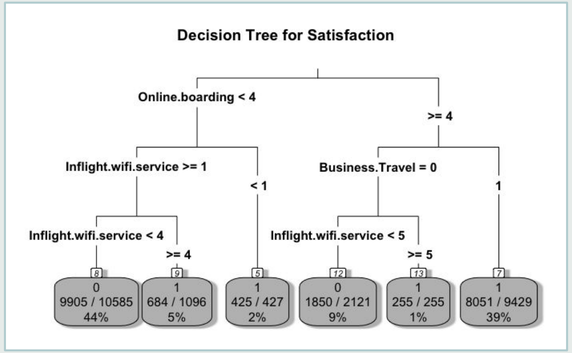
Single Tree

Key considerations

- Trained w/o:
  - 10-fold validation
  - Pruning
- Baseline for tree models
- Performed surprisingly well
- Very high number of false positives (costly errors)

		Actual		
		Not Satisfied	Satisfied	
Predicted	Not Satisfied	2906	270	
	Satisfied	444	2360	Accuracy 88.06%

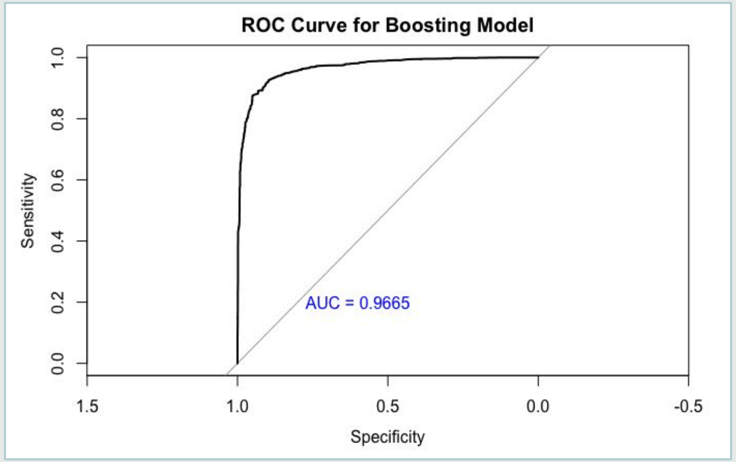




Boosting Model

Key considerations

- Use of gradient boosting classifier
- Bernoulli distribution for binary classification tasks



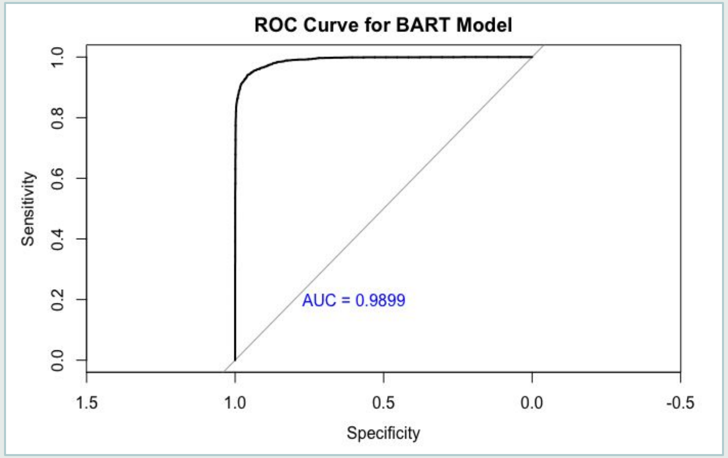
Actual

Predicted	Not Satisfied	Satisfied	
Not Satisfied	3110	285	
Satisfied	240	2345	Accuracy 91.22%

BART Model

Key considerations

- lbart function → used for classification tasks
- Extremely lengthy runtime
- Lowest value of false positives between all models (the more costly prediction)



Actual

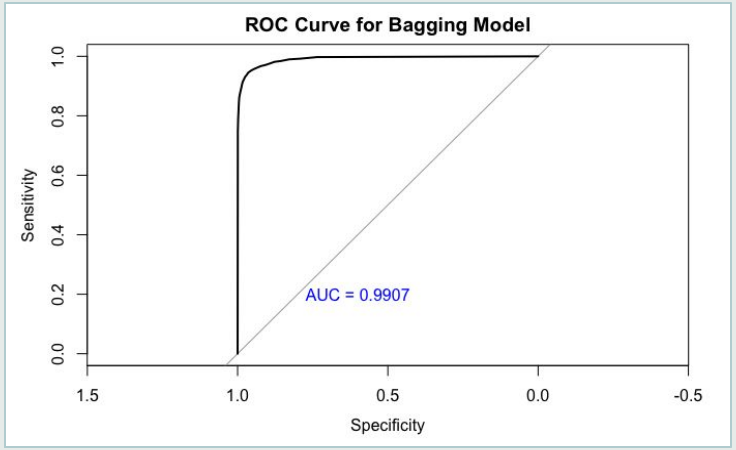
Predicted

	Not Satisfied	Satisfied	
Not Satisfied	3257	189	
Satisfied	93	2441	Accuracy 95.28%

Bagging Model

Key considerations

- No parameter tuning for "treebag" method
- Highest number of classified "True Positives"

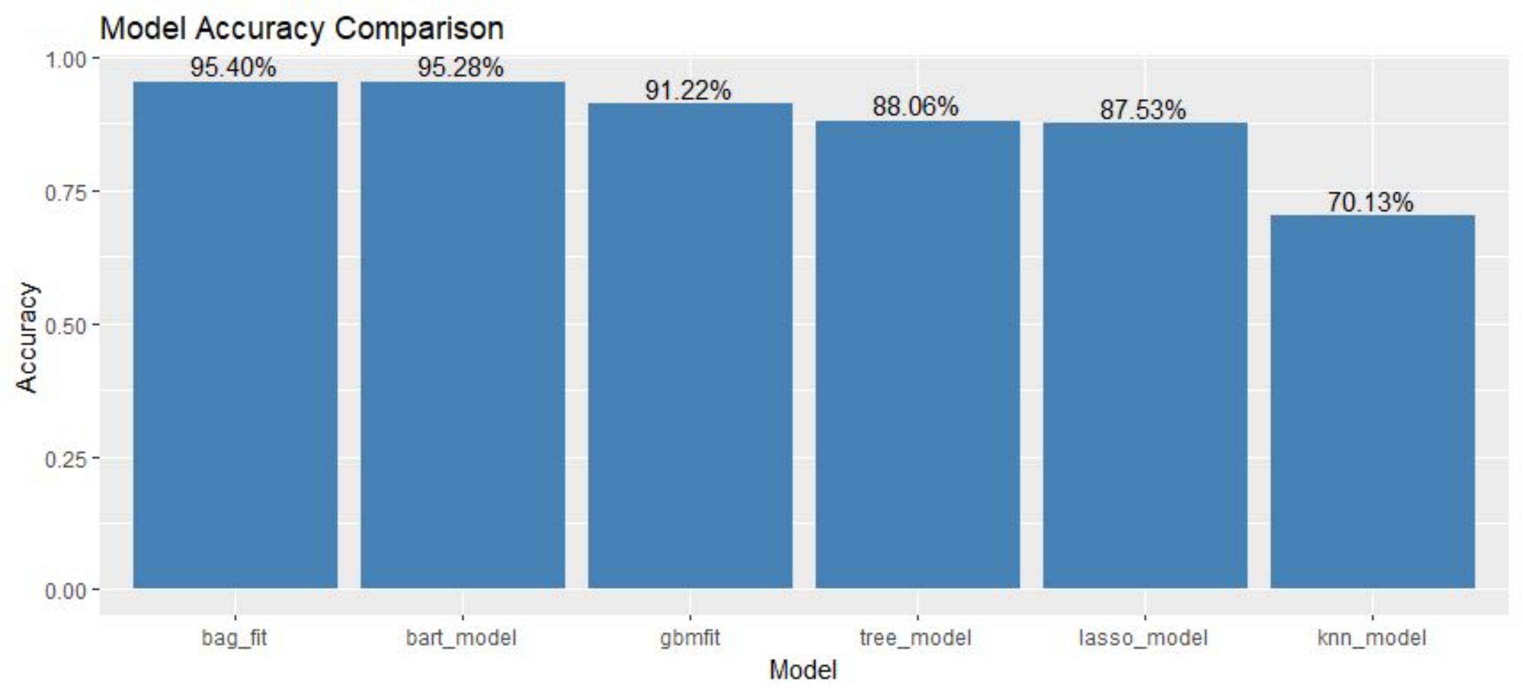


Actual

Predicted		Not Satisfied	Satisfied	
	Not Satisfied	3249	174	
	Satisfied	101	2456	Accuracy 95.40%

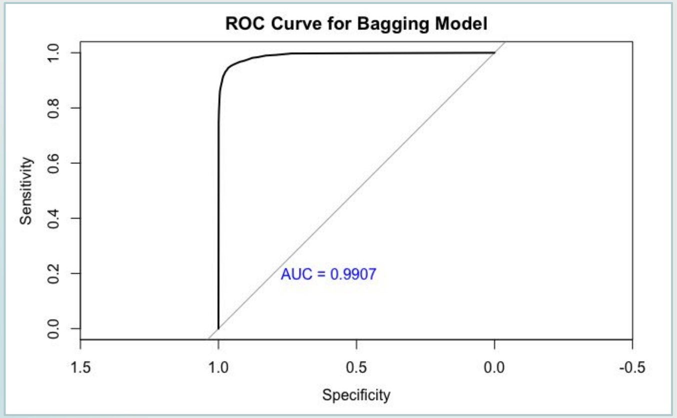
Final Model Selection

Model Comparison



Final Model Analysis

Final Model → **Bagging Model**



Actual

Predicted		Not Satisfied	Satisfied	
	Not Satisfied	3249	174	
	Satisfied	101	2456	Accuracy 95.40%

Top predictors:

- Traveling for Business
- Online boarding Satisfaction
- Wifi service and entertainment

	Overall
ClassBusiness1	100.000
Online.boarding	94.443
Business.Travel1	91.774
Inflight.wifi.service	90.537
Inflight.entertainment	70.886
Leg.room.service	36.417
Baggage.handling	31.252
Age	22.488
Flight.Distance	22.136
Ease.of.Online.booking	20.898
On.board.service	19.434
Inflight.service	19.380
Checkin.service	18.715
Seat.comfort	14.557
Cleanliness	13.404
Gate.location	11.375
Departure.Arrival.time.convenient	9.688

Conclusion

Conclusion

United Airlines could capitalize on this information through:

Wifi Priority	→	Potentially increasing wifi availability / decreasing wifi costs
Economy Focus	→	Investing in ways to improve the economy travel experience
Data Collection	→	Understanding how customer preferences change through new data

Thank You

Any Questions?

Bonus - SVM Model

Created an SVM model to broaden our scope and test a model we did not specifically cover in class. Upon research, SVM Models are supposed to perform well on binary classification problems. However, this model had extremely long runtimes and performed in the middle of the pack compared to the other models. The confusion matrix for this model can be seen below:

		Actual		
		Not Satisfied	Satisfied	
Predicted	Not Satisfied	3081	477	
	Satisfied	269	2153	Accuracy 87.53%

Would make it the 5th best performing model