Predicting Hotel Cancellations

ERIC ZHAO · ALEXANDER RUDRA · SIENNA ZHU · MELODY LAM · BRETT LIN

Agenda

1. Problem Statement
2. Dataset Overview
3. Logistic Regression
4. Support Vector Machine
5. Decision Trees
6. Random Forests
7. Multilayer Perceptron
8. Key Takeaways

1

Problem Statement

The business issue being examined and our goal

BAC

Our Motivation and Goal

Basis of Our Project

- The recent average cancellation rate was **40%**
- Research paper published in 2019 that details **aggregated hotel booking data** over three years

Problem Identification

- Hotel cancellations are a **risk** that hotels deal with, making **revenue management & forecasting difficult**

Goal

- To **accurately predict** if a given hotel booking will be cancelled, on an **aggregate industry level** as well as on a **specific hotel-type level**

2

Dataset Overview

Diving into the data used for this project

A Deeper Look at the Data

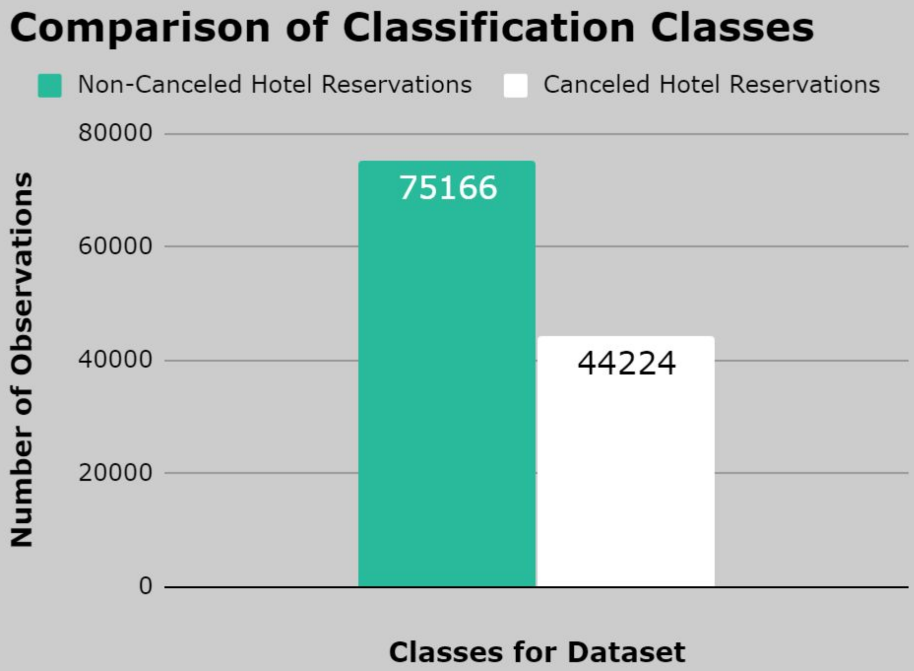
Out of nearly **120,000 rows** of data and **32 features** in our data set, below are the **six features** that we hypothesized would be **especially relevant** for predicting cancellations

Average Daily Rate	Hotel Type	Lead Time
Previous Cancellations	Deposit Type	Customer Type

Dependent Variable

Cancellation Outcome

SMOTE Data Resampling



Problem

IBM mentors suggested poor results could be a result of imbalanced classes — a deeper look revealed a strong disparity in counts

Solution

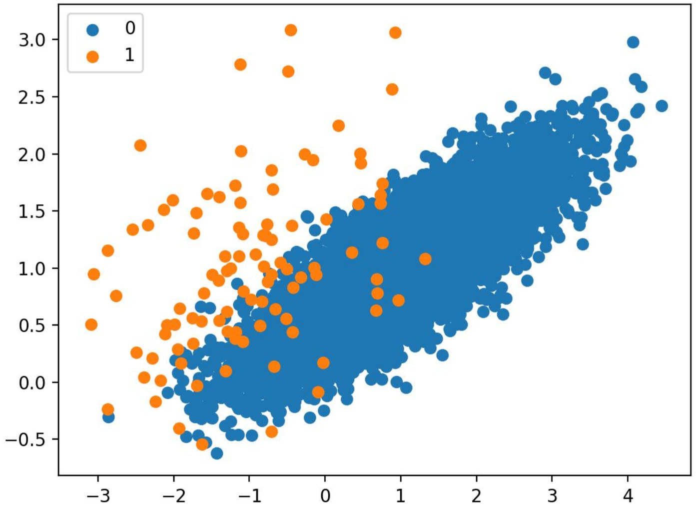
Utilize the Synthetic Minority Over-Sampling Technique (SMOTE) to even out the classes for improved model performance

SMOTE(ENN) Method

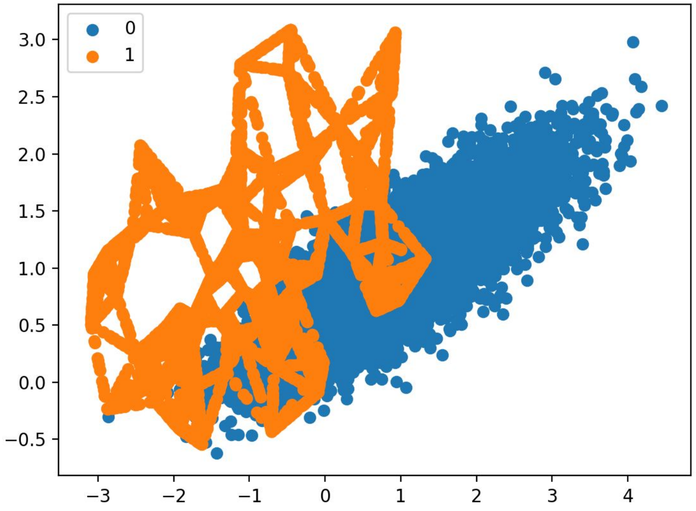
For the minority class (canceled hotel reservations), new observations were synthetically created from a nearby neighbor

A Graphical View of SMOTE

Prior to Resampling: Imbalanced

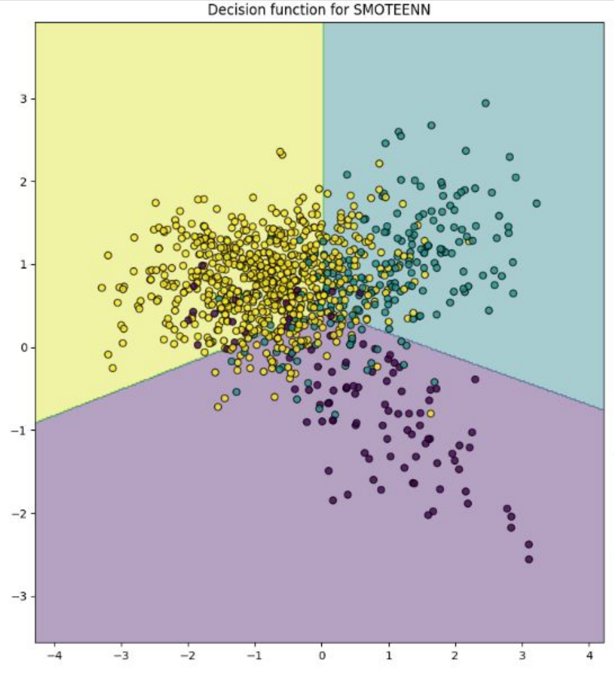


After Resampling: Balanced

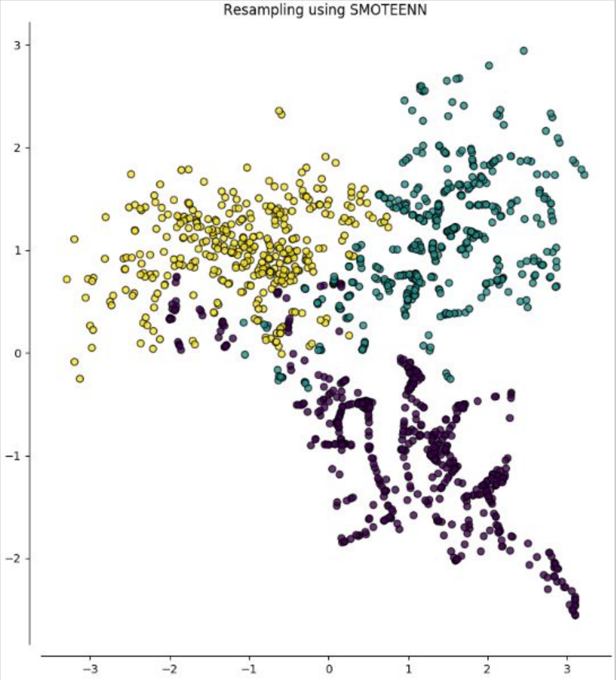


A Graphical View of SMOTEENN

Prior to Resampling: Imbalanced



After Resampling: Balanced

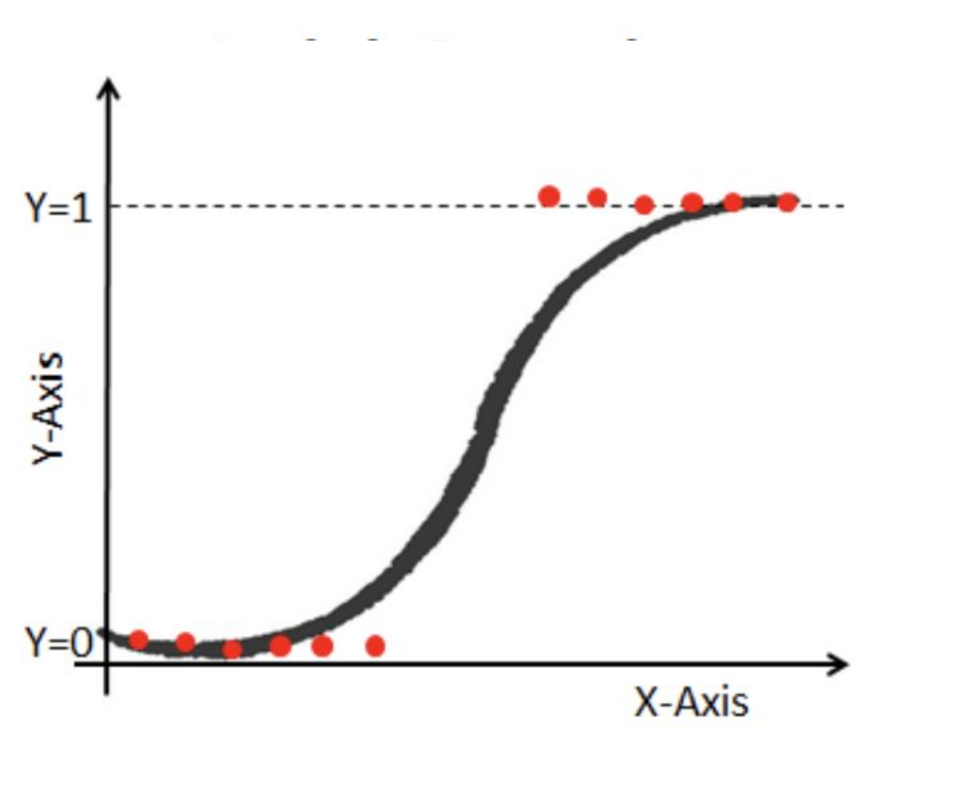


3

Logistic Regression

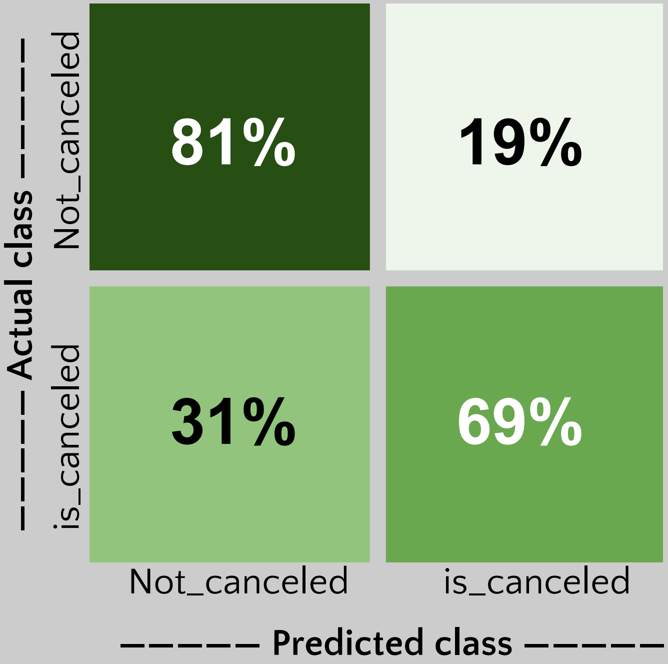
Examining our logistic regression model application

Logistic Regression



Logistic Regression on Aggregate Hotels

Normalized Confusion Matrix



Model Classification Report

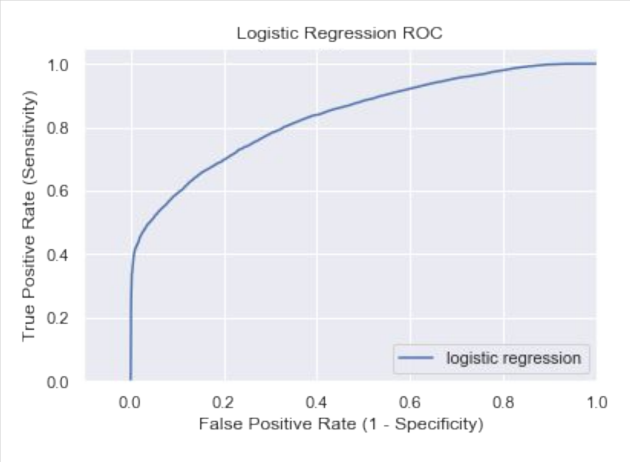
	Precision	Recall	f1-score
Not_canceled	0.72	0.81	0.76
Is_canceled	0.78	0.69	0.73

Accuracy

0.749

ROC-AUC

0.836



Logistic Regression on City Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	84%	16%
is_canceled	33%	67%

Model Classification Report

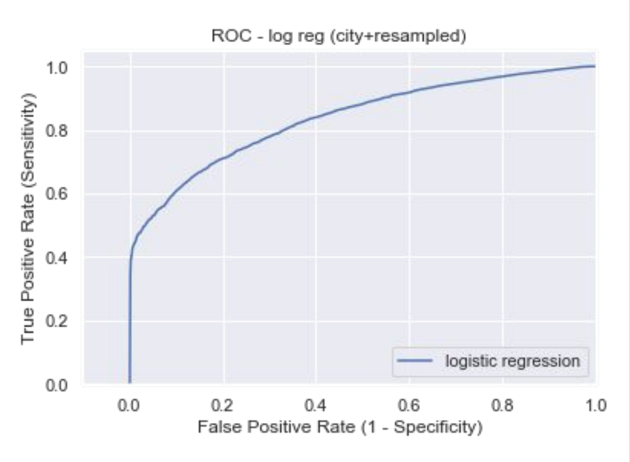
	Precision	Recall	f1-score
Not_canceled	0.72	0.84	0.78
Is_canceled	0.81	0.67	0.72

Accuracy

0.758

ROC-AUC

0.836



Logistic Regression on Resort Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	74%	26%
is_canceled	27%	73%

Model Classification Report

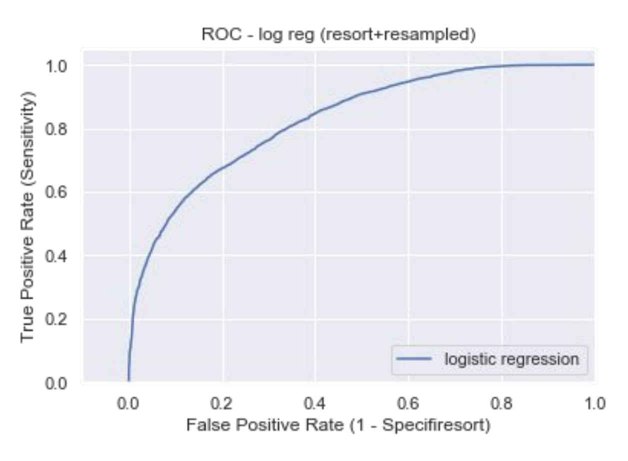
	Precision	Recall	f1-score
Not_canceled	0.73	0.74	0.74
Is_canceled	0.73	0.73	0.73

Accuracy

0.733

ROC-AUC

0.829



Logistic Regression - Insights

Feature	Data Type	Coefficient	Δ Odds	Effect on cancellation
No Deposit	Categorical	-3.34	0.0354	-
Required Parking	Categorical	-3.09	0.0454	-
Previous Cancellations	Continuous	1.98	7.29	+
Summer	Categorical	-1.41	0.380	-
Repeated Guest	Categorical	-0.87	0.416	-

BAC

Summary of Logistic Regression

Advantages

Easy to visualize features

Able to show +ve/ -ve features with our target

+

Disadvantages

Low accuracy

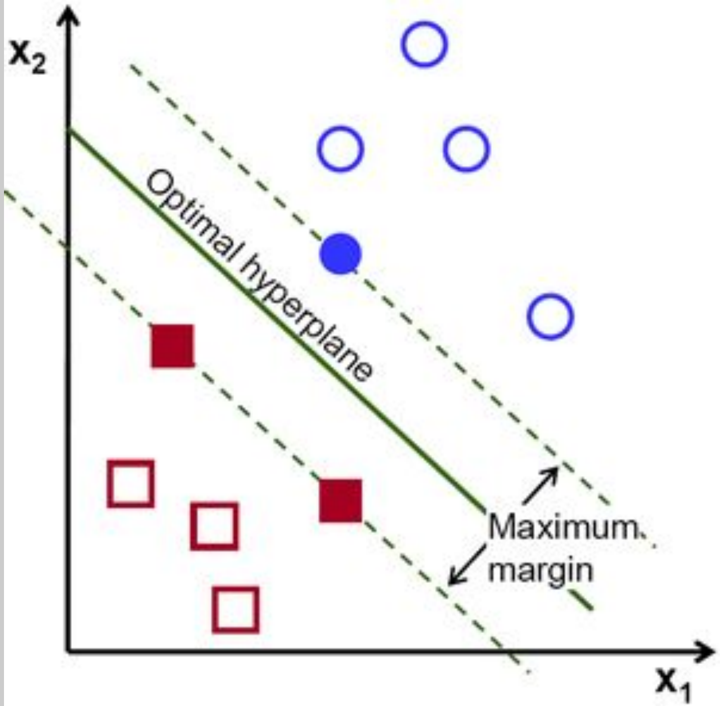
Low predictive power

4

Support Vector Machines

Looking at further SVM application to logistic regression

SVM Overview



Hyperplane

The decision boundary that classify the data points.

Support Vector

The points that help the model identify the hyperplane. These lie on the margins.

Margins

The distance between the hyperplane and its support vectors.

SVM on Aggregate Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	68%	32%
is_canceled	43%	57%

Model Classification Report

	Precision	Recall	f1-score
Not_canceled	0.61	0.68	0.64
Is_canceled	0.64	0.57	0.60

Accuracy

0.623

Cross Validated Accuracy

0.468

SVM on City Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	71%	29%
is_canceled	45%	55%

Model Classification Report

	Precision	Recall	f1-score
Not_canceled	0.61	0.71	0.66
Is_canceled	0.65	0.55	0.60

Accuracy

0.625

Cross Validated Accuracy

0.398

SVM on Resort Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	65%	35%
is_canceled	40%	60%

Model Classification Report

	Precision	Recall	f1-score
Not_canceled	0.62	0.65	0.64
Is_canceled	0.63	0.60	0.61

Accuracy

0.622

Cross Validated Accuracy

0.467

SVM Feature Selection

Feature	Data Type
Arrival Date (month)	Object
Meal	
Country	
Reserved Room Type	
Assigned Room Type	
Deposit Type	
Customer Type	

Categorical

Features that cannot be directly quantified

Feature Scaling

SVM requires that standardized data before analysis.

Selection

Due to the requirements of SVM, could only use 17 of 32 total features

BAC

Summary of SVM

Advantages

Alternative approach to logistic regression

Emphasizes the importance of categorical features

Disadvantages

Unable to use categorical features

Low Accuracy

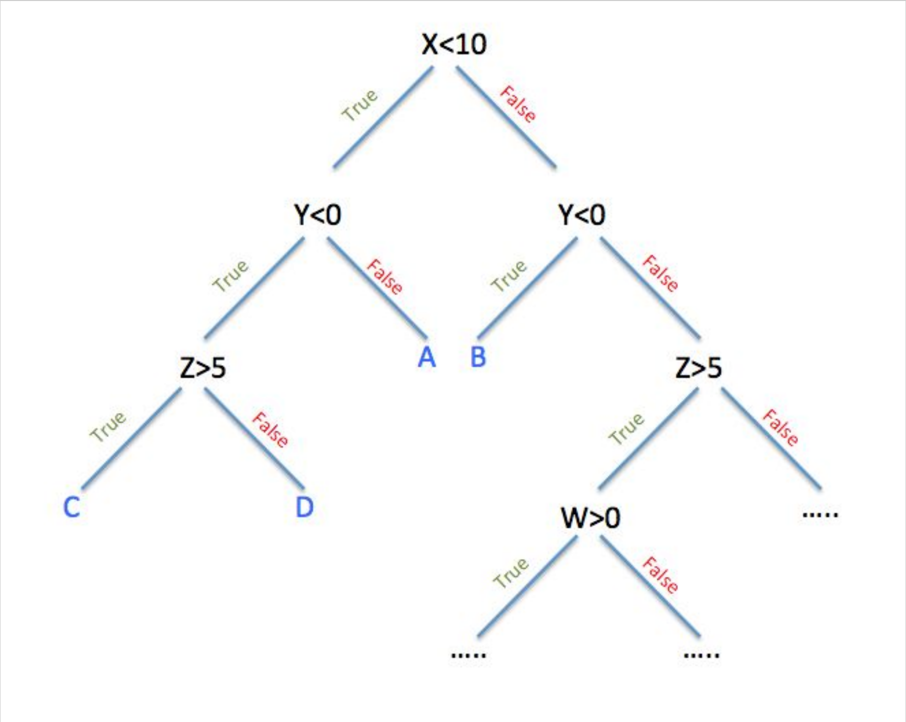
"Blackbox"

5

Decision Trees

Examining features through our decision tree model

Decision Trees



Decision Tree on Aggregate Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	83%	17%
is_canceled	24%	76%

Model Classification Report

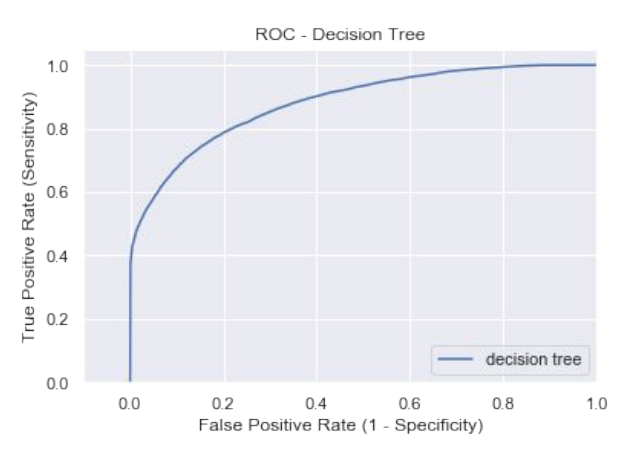
	Precision	Recall	f1-score
Not_canceled	0.78	0.82	0.80
Is_canceled	0.81	0.76	0.79

Accuracy

0.796

ROC-AUC

0.885



Decision Tree on City Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	73%	27%
is_canceled	23%	77%

Model Classification Report

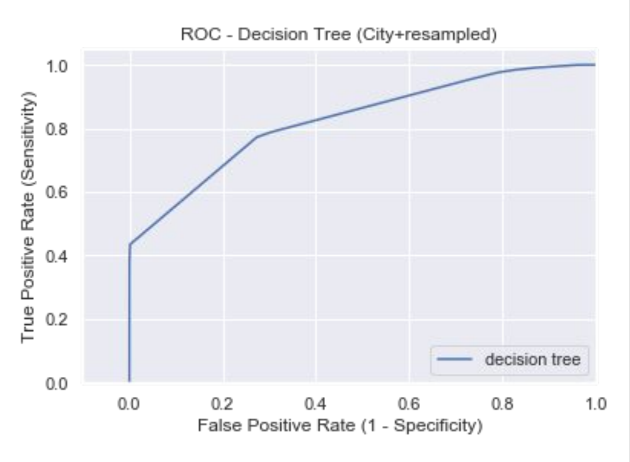
	Precision	Recall	f1-score
Not_canceled	0.76	0.73	0.74
Is_canceled	0.74	0.77	0.75

Accuracy

0.749

ROC-AUC

0.826



Decision Tree on Resort Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	77%	23%
is_canceled	21%	79%

Model Classification Report

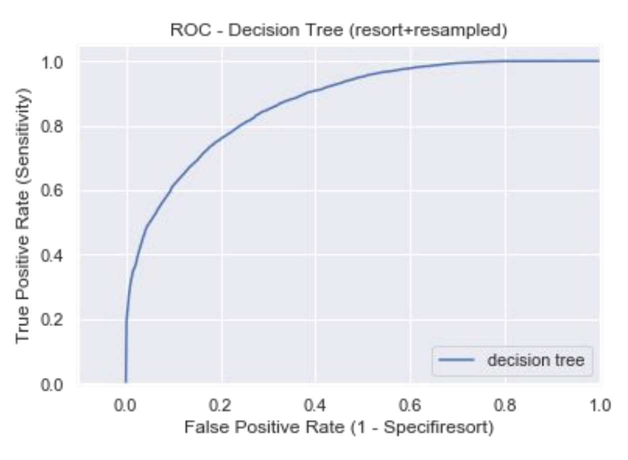
	Precision	Recall	f1-score
Not_canceled	0.79	0.77	0.78
Is_canceled	0.77	0.79	0.78

Accuracy

0.779

ROC-AUC

0.873



Decision Trees - Insights

Feature	Importance (%)
No Deposit	42.0%
Lead Time	16.1%
Average Daily Rate	9.52%
# of Special Requests	8.50%
# of Booking Changes	6.58%
Required Parking	4.97%
# of Previous Cancellations	4.20%

Summary for Decision Tree

Advantages

Able to visualize splits through the tree

Disadvantages

Possibility for overfitting

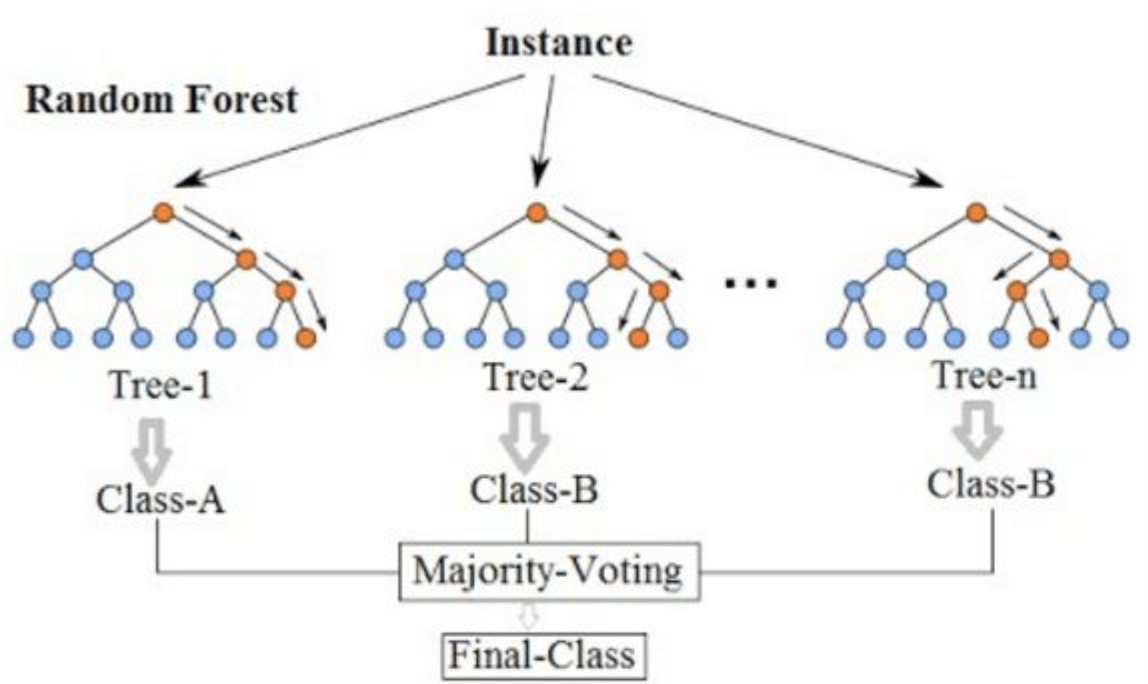
Trees can get complex to visualize with deeper trees

6

Random Forests

Exploring our Random Forest Model for cancellation prediction

Random Forest



Random Forest on Aggregate Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	86%	14%
is_canceled	16%	84%

Model Classification Report

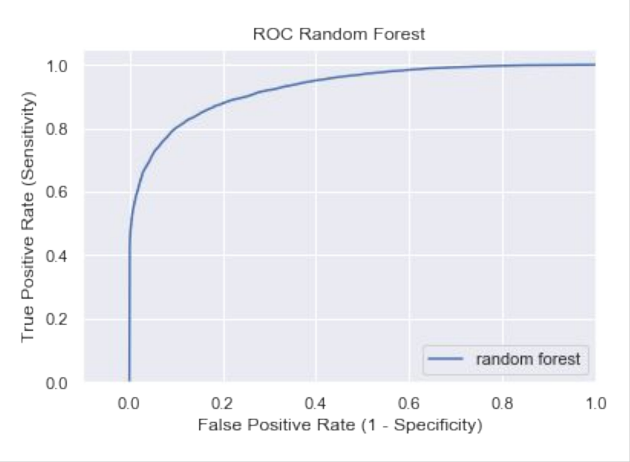
	Precision	Recall	f1-score
Not_canceled	0.84	0.86	0.85
Is_canceled	0.85	0.84	0.85

Accuracy

0.849

ROC-AUC

0.930



Random Forest on City Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	85%	15%
is_canceled	17%	83%

Model Classification Report

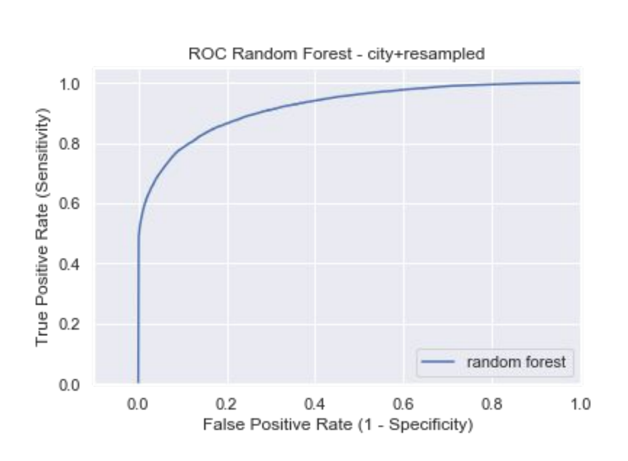
	Precision	Recall	f1-score
Not_canceled	0.83	0.85	0.84
Is_canceled	0.85	0.83	0.84

Accuracy

0.841

ROC-AUC

0.922



Random Forest on Resort Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	84%	16%
is_canceled	14%	86%

Model Classification Report

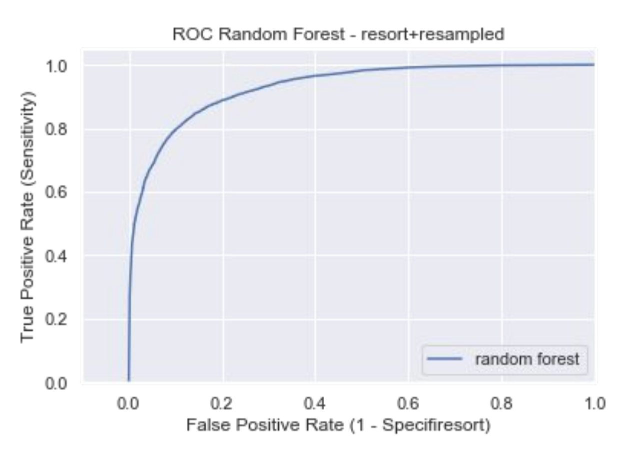
	Precision	Recall	f1-score
Not_canceled	0.86	0.84	0.85
Is_canceled	0.84	0.86	0.85

Accuracy

0.850

ROC-AUC

0.931



Random Forest - Insights (Aggregate)

Feature	Importance (%)
Lead Time	32.1%
Average Daily Rate	16.12%
No Deposit	16.8%
# of Special Requests	5.34%
# of Booking Changes	2.78%
Required Parking	2.42%

Summary for Random Forest

Advantages

High Accuracy

High Predictive Power

Disadvantages

More "black-boxed"

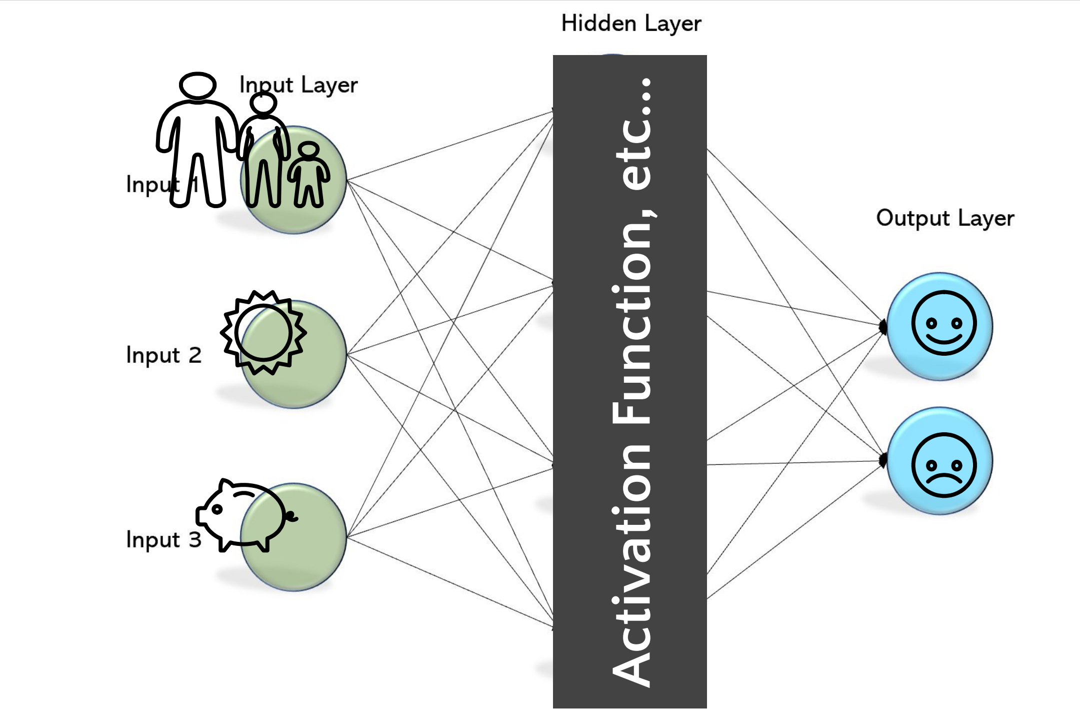
Unable to explore a feature's +-ve or -.ve effect on our outcome

7

Multilayer Perceptron

Looking into Neural Network, the first step in Deep Learning

MLP Overview



Model Parameters

Input Variables: **25**
Output Classes: **2**

Hidden Layer: **1**
Neurons on first layer: **20**

Activation Function: **Logistic**
Solver Function: **Adam**

Max Iteration: 1000

MLP on Aggregate Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	88%	12%
is_canceled	12%	88%

Model Classification Report

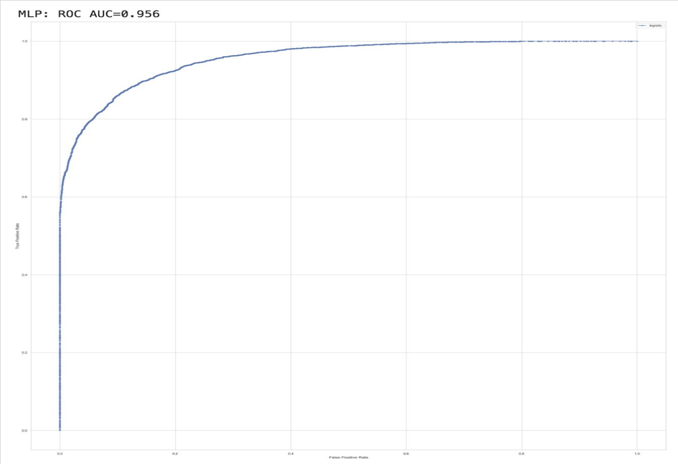
	Precision	Recall	f1-score
Not_canceled	0.84	0.88	0.86
Is_canceled	**0.91**	0.88	0.89

CrossVal Score

0.88

ROC AUC

0.956



MLP on City Hotels

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	92%	8%
is_canceled	12%	88%

Model Classification Report

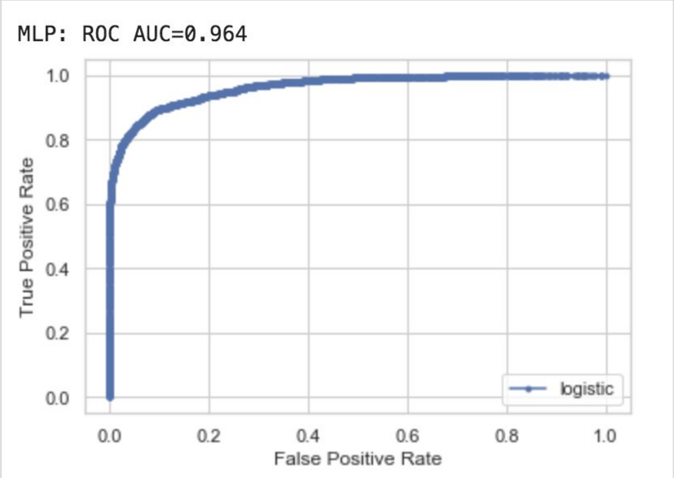
	Precision	Recall	f1-score
Not_canceled	0.86	0.92	0.89
Is_canceled	**0.93**	0.88	0.91

CrossVal Score

0.87

ROC AUC

0.964



MLP on Resort Hotel

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	81%	19%
is_canceled	9%	91%

Model Classification Report

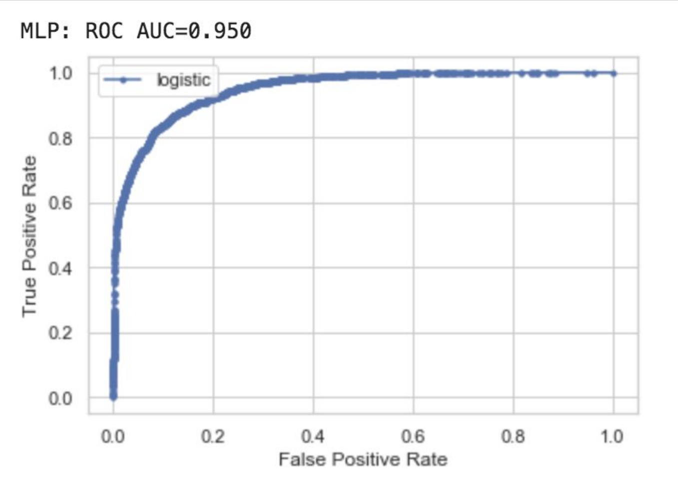
	Precision	Recall	f1-score
Not_canceled	0.86	0.81	0.84
Is_canceled	**0.88**	0.91	0.89

CrossVal Score

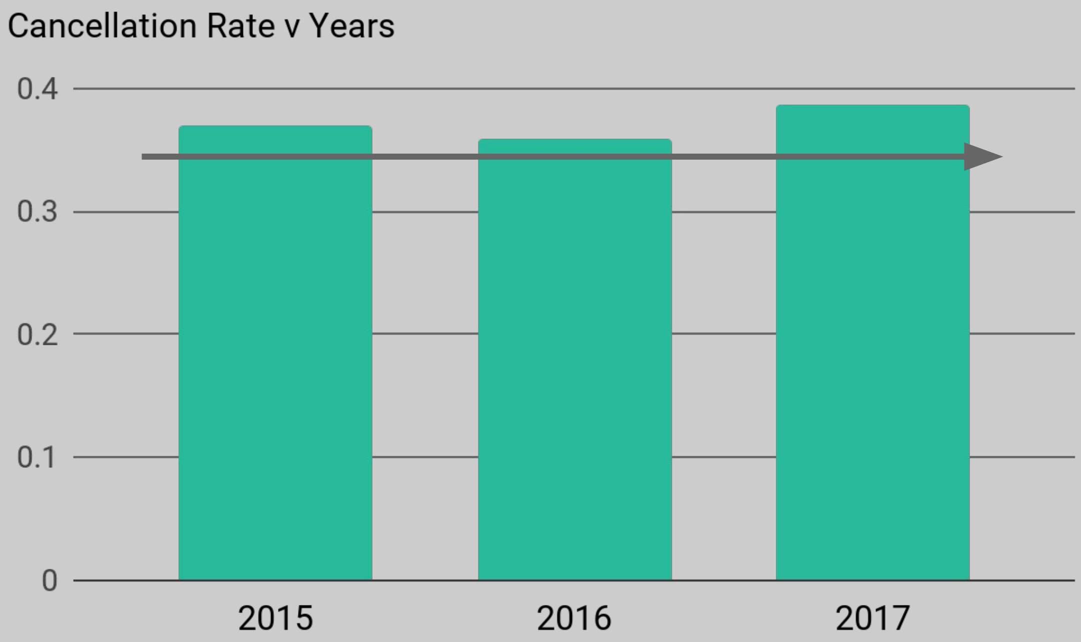
0.86

ROC AUC

0.950



Cancellation Rates over Years



No time series correlation between booking cancellation and years

Summary of MLP

Advantages

High accuracy

High ROC AUC

Good predicting power

+

Disadvantages

Limited business insights

No feature importance

“Blackbox”

8

Key Takeaways

Identifying our best models, the best important features, and next steps

Framework for the Ideal Model

We're trying to predict booking cancellations on an **industry level** AND on a **hotel-type specific level**

- **Industry-Wide Considerations**
  - More of a research/conceptual model
  - Can afford more complexity
- **Hotel-Type Specific Considerations**
  - Much more business-oriented
  - Must be mindful of the benefit/complexity tradeoff
  - Watch the False-Negative rate of each model

Best Overall Aggregate Model

Decision Tree

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	83%	17%
is_canceled	24%	76%

Accuracy

0.796

ROC-AUC

0.885

Random Forest

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	86%	14%
is_canceled	16%	84%

Accuracy

0.849

ROC-AUC

0.930

MLP

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	88%	12%
is_canceled	12%	88%

CrossVal Score

0.88

ROC AUC

0.956

Verdict: Best Model on the Industry Level is MLP

- Highest Accuracy and ROC scores
- Best Confusion Matrix

Best Overall City Model

Decision Tree

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	73%	27%
is_canceled	23%	77%

Accuracy
0.749

ROC-AUC
0.826

Random Forest

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	85%	15%
is_canceled	17%	83%

Accuracy
0.841

ROC-AUC
0.922

MLP

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	92%	8%
is_canceled	12%	88%

CrossVal Score
0.87

ROC AUC
0.964

Verdict: Best Model on the City Level is Random Forest

- Performance in comparable to the MLP, but slightly worse
- That trade-off is warranted by the significant reduction in model complexity

Best Overall Resort Model

Decision Tree

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	77%	23%
is_canceled	21%	79%

Accuracy
0.779

ROC-AUC
0.873

Random Forest

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	84%	16%
is_canceled	14%	86%

Accuracy
0.850

ROC-AUC
0.931

MLP

Normalized Confusion Matrix

Actual class \ Predicted class	Not_canceled	is_canceled
Not_canceled	81%	19%
is_canceled	9%	91%

CrossVal Score
0.86

ROC AUC
0.950

Verdict: Best Model on the City Level is Random Forest

- Performance was even more similar to the MLP
- The benefit from model MLP model complexity is even more marginal

Key Determinants of a Cancellation

Across all models, we saw recurring predictive features:

- No Deposit
- Required Parking
- Previous Cancellations
- Lead Time
- # of Special Requests
- Avg. Daily Rate

8

Q&A

APPENDIX: MLP Feature Processing

Adults

Children

Babies

Family
(binary variable)

arrival_date

arrival_months

arrival_year

arrival_week

Seasons
(categorical variable)

TOTAL

25 input variables, **2** output classes

Optimal Layer & Neuron Size

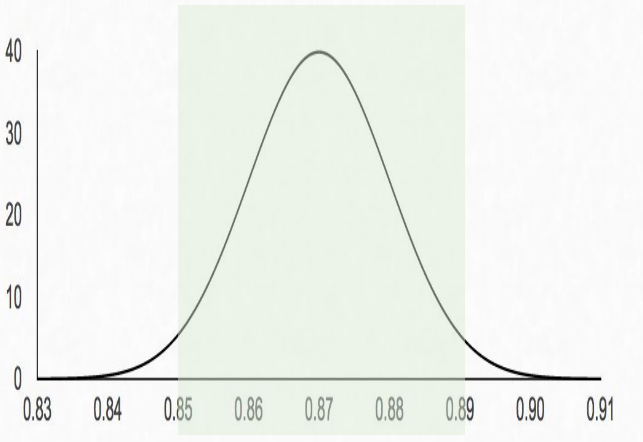
1 input layer, **1** hidden layer, **1** output layer, **20** neurons on hidden layer

APPENDIX: MLP Cross Validation

City Hotel

Mean: 0.87

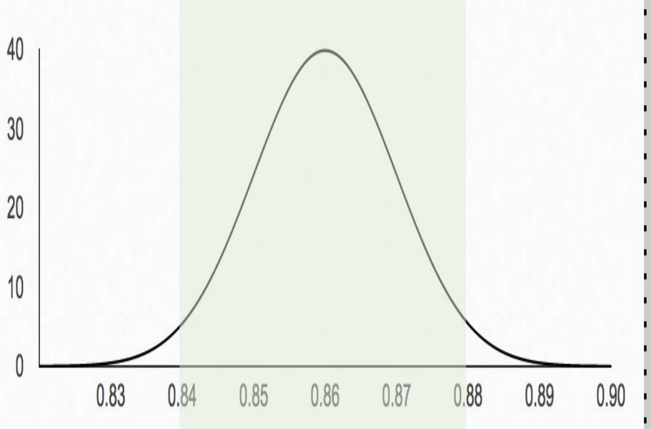
Sigma: 0.01



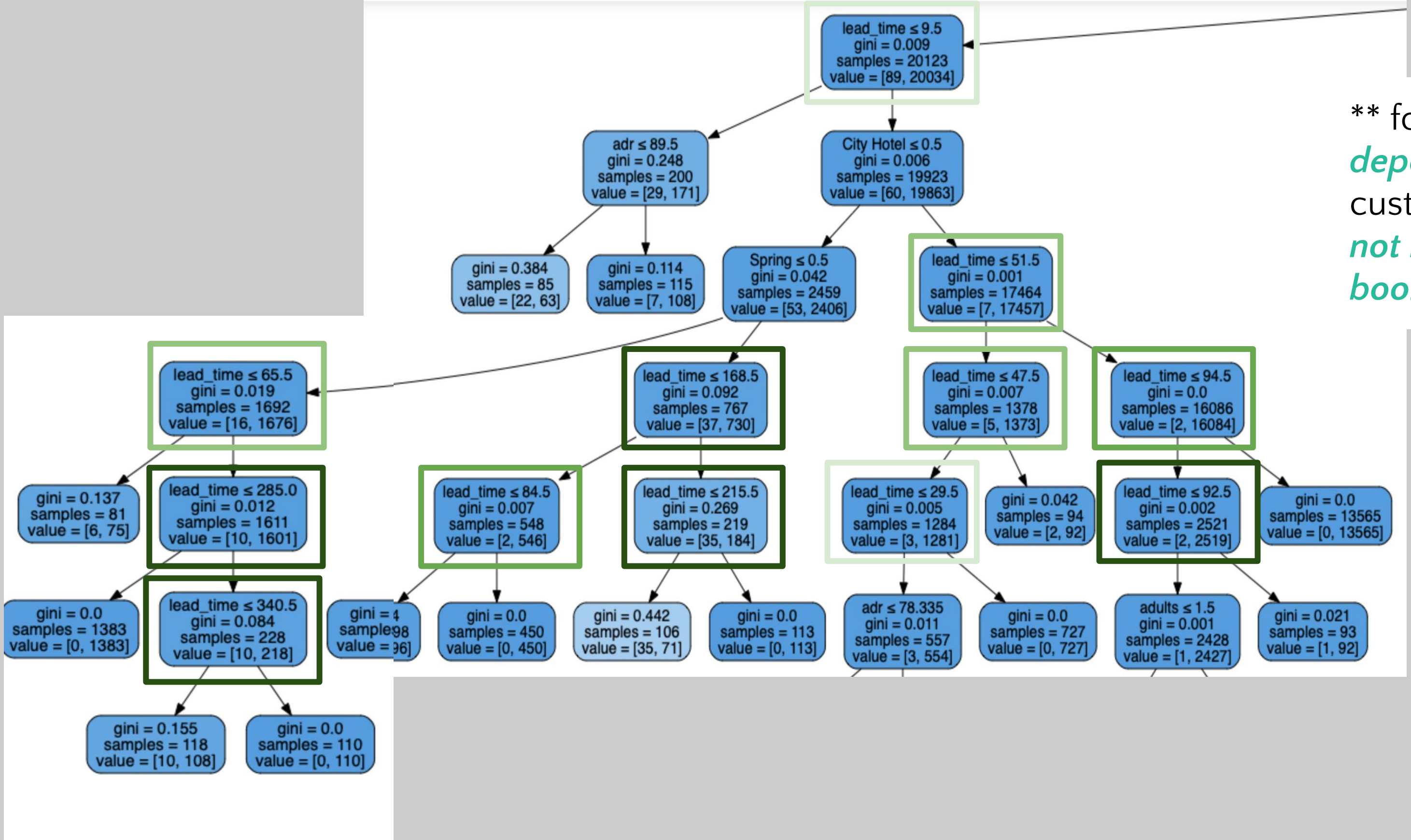
Resort Hotel

Mean: 0.86

Sigma: 0.01



APPENDIX: DECISION TREE SPLITS



** for hotels with *no deposit needed* and customers who *did not have any booking changes*

APPENDIX: ROC Curves