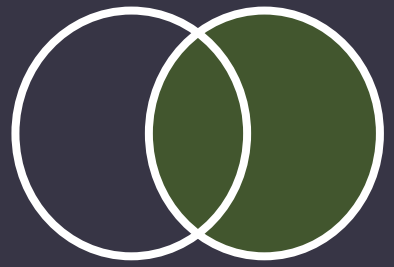


# DataQuest 2023

Team 7: Oren Joffe, Aaron Pogrin, Ryan Rakusin


**Michael Scott**

CEO of Brescia Norton

BRESCIA NORTON

# Message from the CEO

> **“The company is losing money from cancellations. Fix it.”**

# Problem, Idea, Implementation

## Problem

*The hotel's management team identified the challenge of predicting and managing booking cancellations*

Problem

## Idea

*The hotel is looking to leverage machine learning algorithms to predict booking cancellations.* .

Idea

## Implementation

*Determine how Brescia Norton can use machine learning to predict booking cancellations using the dataset provided*.

Implementation

# Data Exploration

**29,020**
**Bookings**

**19**
**Variables**

*15 numerical, 4 categorical*

**33%**

**Ratio of Cancellations**

# Data Cleaning

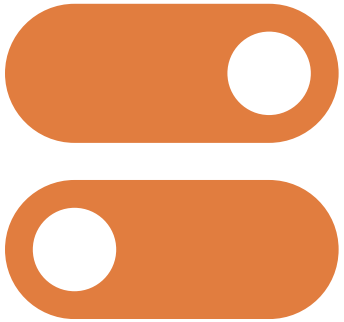
*Although the dataset was almost perfect there were some minor changes made.*

01

**Checking For Missing Values**

02

**Dropping Unnecessary Variables**



03

**Converting Categorical Variables to numerical**

# Feature Engineering

## Five new features were added to the dataset.

These features aim to get a more accurate model of the dataset by incorporating multiple variables together. The fifth variables added was converted all dates to days since the first booking in the dataset to get a better representation of time.

```
BookingStatus                1.000000
LeadTime                     0.452386
ArrivalYear                  0.189561
DaysSinceFirst               0.186090
AvgRoomPrice                 0.165365
Nights                       0.100442
NumGuests                    0.094195
NumWeekNights                0.088575
MealPlan                     0.088085
NumAdults                    0.087969
NumWeekendNights             0.061888
NumChildren                  0.039264
RoomType                     0.035358
ArrivalDate                  0.012961
BookingID                   -0.001129
ArrivalMonth                -0.013979
NumPrevCancellations        -0.033401
CancelRatio                 -0.042983
NumPreviousNonCancelled     -0.063662
Parking                     -0.082415
RepeatedGuest               -0.113404
MarketSegment               -0.123802
SpecialRequests             -0.236206
Name: BookingStatus, dtype: float64
```

### Cancel Ratio

Previous Cancellations / (Previous Cancellations + Previous Non-Cancellations)

### Total Number of Guests

Number of Children + Number of Adults.

### Total Nights Stayed

Number of Weeknights + Number of Weekend Nights

### Total Price

Average Price per Room x Number of Nights Stayed

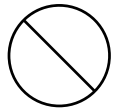
# Model Selection Criteria

## Accurate

Has high percentage of predictions that are correct.

## Specific

High percentage of "no" cases that were predicted as "no".



## Precise

Has high percentage of "yes" predictions that are correct.

## Optimal ROC Curve

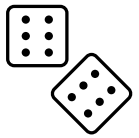
The Receiver Operating Characteristic curve must be as close to (1,1) as possible.

## Sensitive

Has high percentage of "yes" predictions that were predicted as "yes".

## High Area Under Curve Score

"Probability that the model ranks a random positive example more highly than a random negative example." – GoogleDevelopers



# Logistic Regression

## Model Idea 1

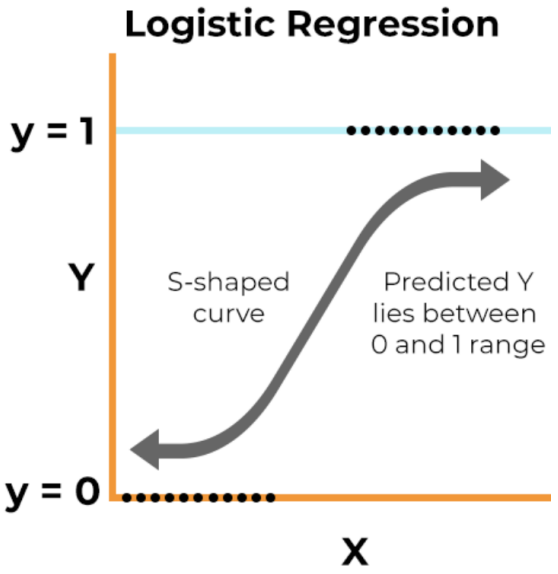
*Logistic regression is a statistical method used for binary classification problems, where the output variable takes on only two possible values. It uses a logistic function to model the probability of the outcome variable as a function of the input variables.*

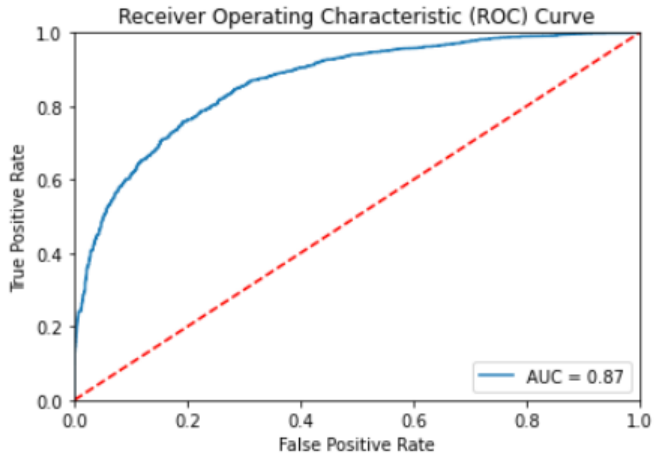
**Accuracy**
*0.806*

**Recall**
*0.891*

**Precision**
*0.832*

**F1 Score**
*0.861*



# Decision Tree Classifier

## Model Idea 2

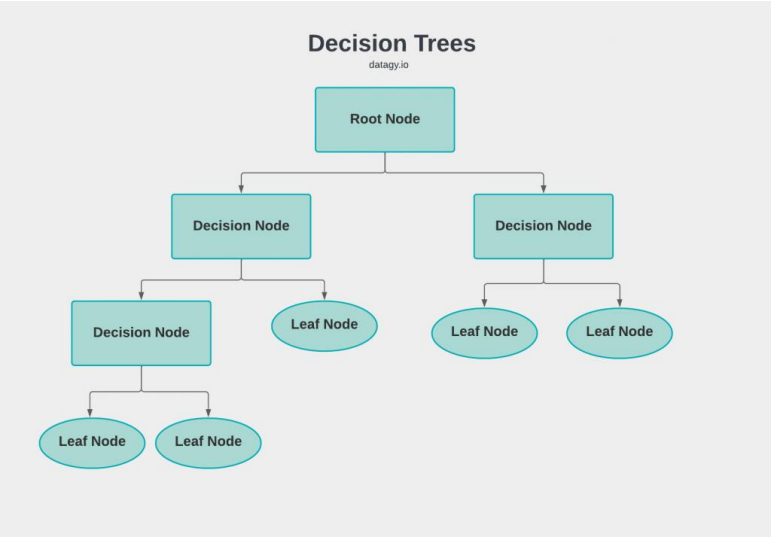
*A decision tree model is a type of supervised machine learning algorithm that predicts a target variable by recursively partitioning the input data into smaller and smaller subsets based on the most informative features. It can be used for both classification and regression tasks.*

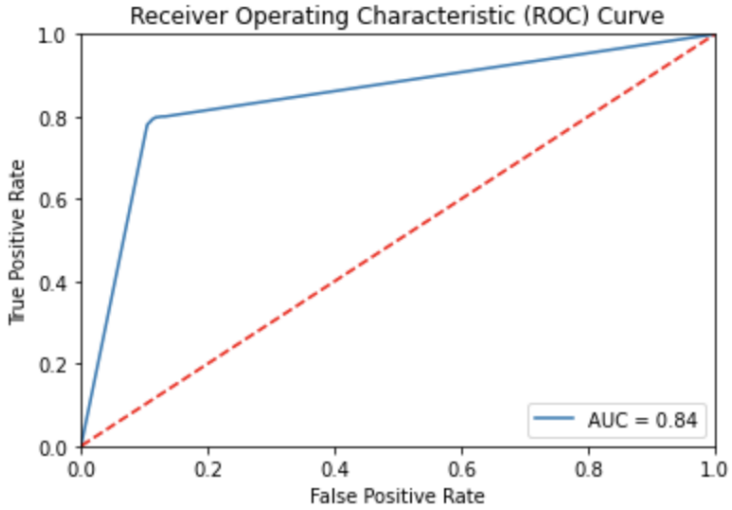
## Accuracy
*0.863*

## Recall
*0.782*

## Precision
*0.799*

## F1 Score
*0.791*



# Multilayer Perceptron Classifier

## Model Idea 3

*A Multi-Layer Perceptron (MLP) model is a type of artificial neural network that consists of multiple layers of interconnected nodes, with each node performing a linear or non-linear operation on its inputs. It is commonly used for classification and regression tasks and can learn complex patterns in data by adjusting the weights and biases of the nodes during training.*

## Accuracy
*0.858*

## Recall
*0.911*

## Precision
*0.882*

## F1 Score
*0.896*

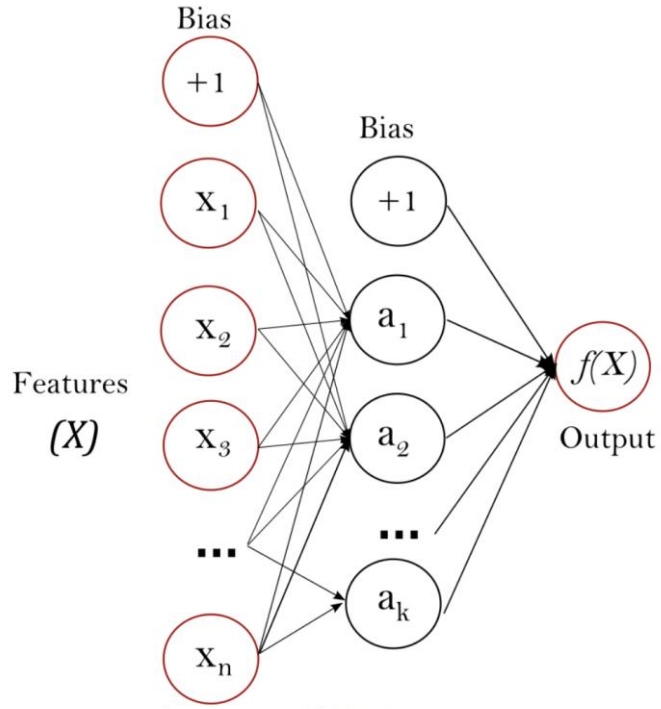
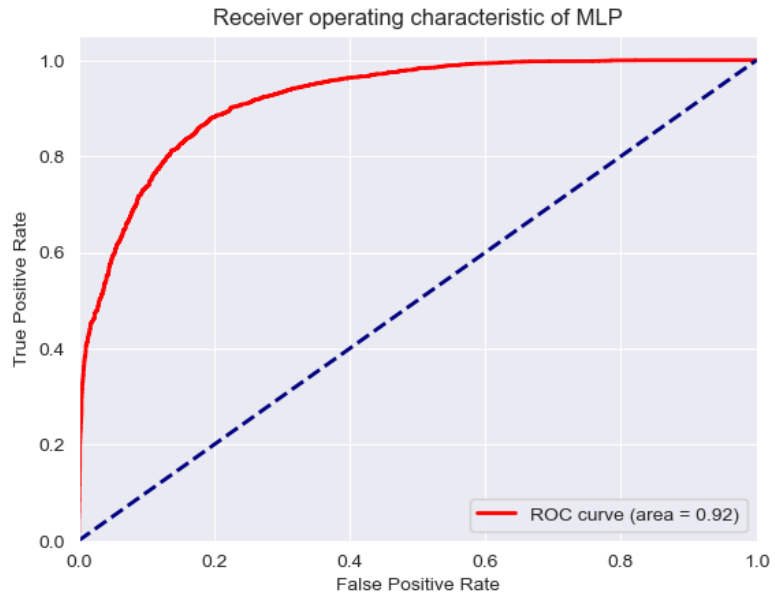



**Figure 1 : One hidden layer MLP.**

Source: Scikit Learn



# Random Forest Classifier

## Model Idea 4

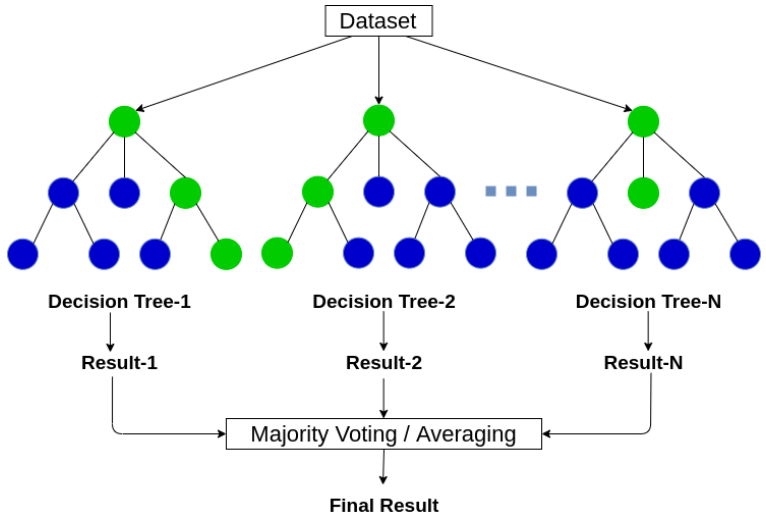
*Random Forest is a type of ensemble machine learning algorithm that builds multiple decision trees and combines their predictions to obtain a more accurate and robust prediction. Each decision tree in the Random Forest is built on a random subset of the input features, and the final prediction is made by averaging or voting over the predictions of all the individual trees. It can be used for both classification and regression tasks.*

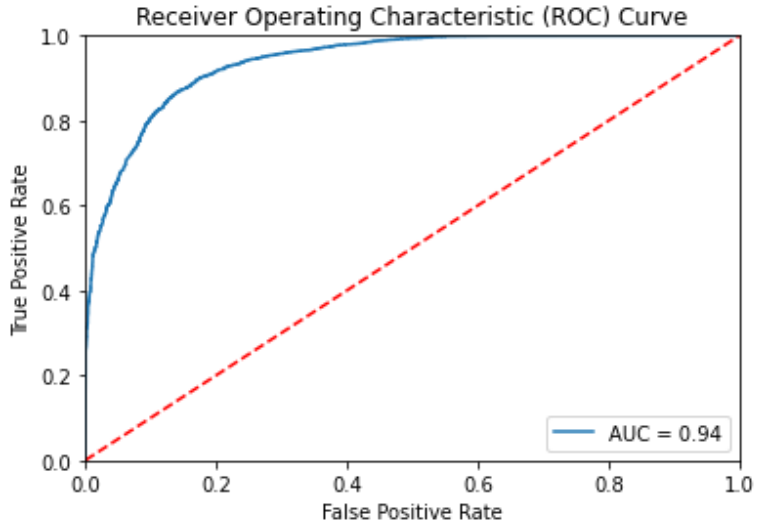
## Accuracy
*0.883*

## Recall
*0.930*

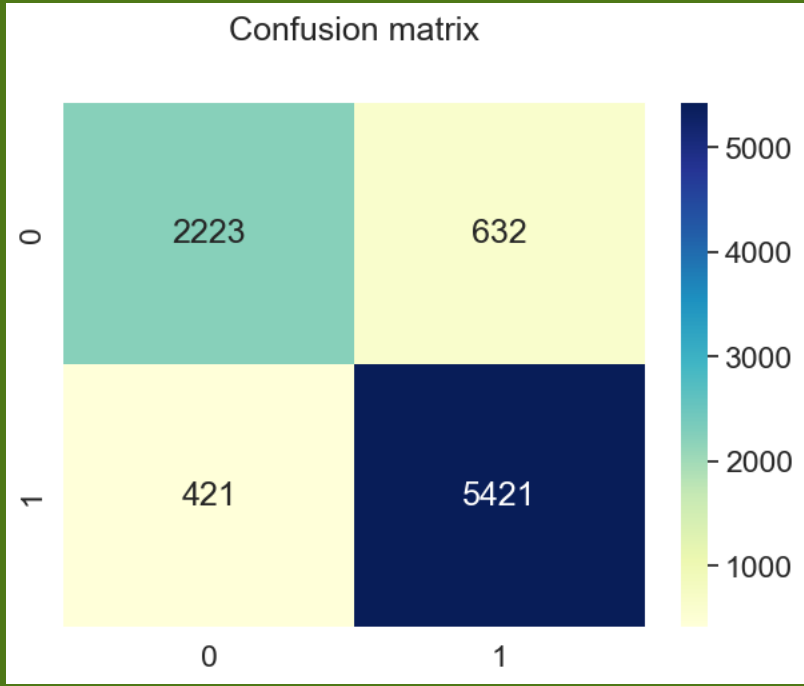
## Precision
*0.900*

## F1 Score
*0.915*



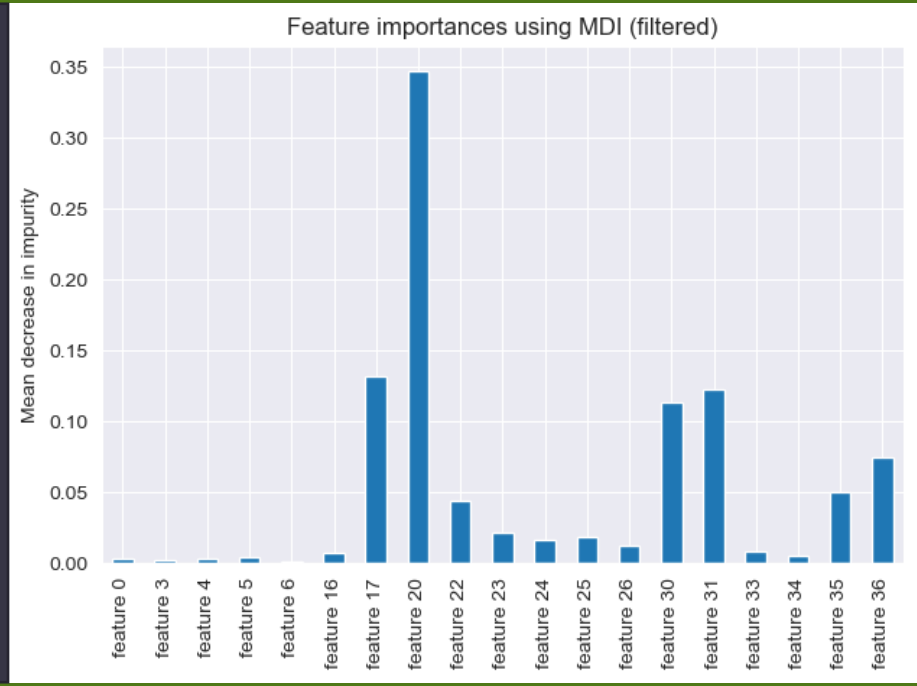



# Random Forest Metrics




**Train vs. Test Accuracy**

Train Accuracy - : 0.895
Test Accuracy - : 0.883
Difference - : 0.012



# Business Plan

01

## Develop

*Train a Random Forest model on Brescia Norton cancellation data.*

02

## Validate

*Test the model's performance on a separate dataset and validate its performance.*

03

## Deploy

*Deploy the model into production, which can involve integrating it with other Brescia Norton systems and ensuring it is scalable, robust, and secure.*

04

## Retrain

*Retrain the model on new data to improve its accuracy and performance over time.*

BRESCIA NORTON

# Bibliography


Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Brownlee, Jason. “How to Use Learning Curves to Diagnose Machine Learning Model Performance.” *MachineLearningMastery.Com* (blog), February 26, 2019. https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/.

Google Developers. “Classification: ROC Curve and AUC | Machine Learning.” Accessed March 12, 2023. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

Garg, Mohit. “Understanding Random Forest Better through Visualizations.” *Medium* (blog), February 2, 2019. https://garg-mohit851.medium.com/random-forest-visualization-3f76cdf6456f.

Google Developers. “Interpreting Loss Curves | Machine Learning.” Accessed March 12, 2023. https://developers.google.com/machine-learning/testing-debugging/metrics/interpretic.

“LandwehrHallFrankCameraReady.Pdf.” Accessed March 12, 2023.

https://www.cs.waikato.ac.nz/~eibe/pubs/LandwehrHallFrankCameraReady.pdf.
DeepAI. “Multilayer Perceptron,” May 17, 2019. https://deepai.org/machine-learning-glossary-and-terms/multilayer-perceptron.

Yiu, Tony. “Understanding Random Forest.” Medium, September 29, 2021. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

# Questions?