Supervised Learning

Muhammad Haris Masood  
[mmasood30@gatech.edu](mailto:mmasood30@gatech.edu)

# Introduction & Data

## Overview

My two classification problems were the hotel reservation problem. These were both found using Kaggle. The first dataset, the hotel reservation dataset, provides comprehensive information about online bookings, and identifies if they were cancelled. The second dataset is medical data, the features are several health markers for a patient, and the label identifies if the patient is at a high risk for a heart attack. The goal of the first dataset is to identify if a particular reservation will be cancelled. The goal of the second dataset is to identify if a patient is at a high risk of a heart attack.

## A blue and orange pie chart Description automatically generatedDatasets Explored & Why they are interesting.

1. A close-up of a graph

   Description automatically generatedHotel Reservation Dataset
2. Heart Disease Dataset

The exact characteristics of the figures can be observed from Figure 1 & 2. In some sense, the datasets are very good opposites of each other. The hotel reservation dataset is quite unbalanced, is a relatively large dataset, and has more numerical features than categorical. The heart disease dataset is quite the opposite, the labels are much more balanced, the dataset itself is significantly smaller, and there are more categorical features in the data.

Utilizing PDP (Partial Dependence Plots) & Pearson R correlations also led to me identifying that there was strong interaction between many of the features in both datasets, and no one feature by itself, would lead to a successful model, relationships between features would also need to be considered. Resultantly, I concluded that both datasets provided nontrivial problems, allowing for a good resource to test different algorithms on. My exploration also did lead me to identify the attack detection dataset had more features with high correlation to the label, and, generally more interaction in the PDP plots.

Finally, the datasets in themselves are quite interesting, modelling unique, and practical problems. However, the risk for heart attack detection problem is quite high stakes, which is quite different than the hotel reservation cancellation problem. For this reason, we employed different metrics to test the success of our algorithms, further diversifying our experimentation, and allowing us to draw more generalizable conclusions.

# Hypothesis

Based on my data exploration I had a few hypotheses.

1. I believe the heart attack risk dataset is the more complex of the two hence I believe Neural Nets should do well on this dataset.
2. The heart attack risk dataset has most of its features as categorical variables; hence it may do well with Decision Trees & Gradient Boosted Classifiers.
3. The heart attack risk dataset has a small amount of data compared to the hotel reservations dataset, so, overfitting will be a problem.
4. The hotel reservation dataset has two features that have a strong relationship with the label (high correlation), namely, the “lead time” and “no\_of\_special\_requests” and I believe KNN & SVM will do well on this dataset (Whereas for the heart dataset, most features have a strong correlation with the label)
5. Due to the unbalanced nature of the hotel reservation dataset, I believe, ensemble methods may work well on this dataset.

# Testing methodology

## Preprocessing Data

For the heart attack risk dataset, I removed the Booking ID column, as that was just an index value. For the hotel reservation dataset, I combined the date and time columns to produce two features. The first identified if the booking was in the H1 or H2 of the month, the second, identified if the booking was in the summer or winter. All categorical features were one-hot encoded to avoided the learner identifying an incorrect ordering pattern.

## Metrics

For the heart attack dataset, I used a custom metric, which I refer to throughout the report (and in the graphs) as F Two, or F2. This is like the F1 score, but the weights have been revised to give recall a higher importance. Recall is twice as important as precision in my testing, as a False Negative is significantly more harmful than a False Positive for this data. For the hotel dataset I used accuracy as it is a staple metric, and works quite well with the dataset (even though the dataset is roughly 33% class 1)

## Procedure

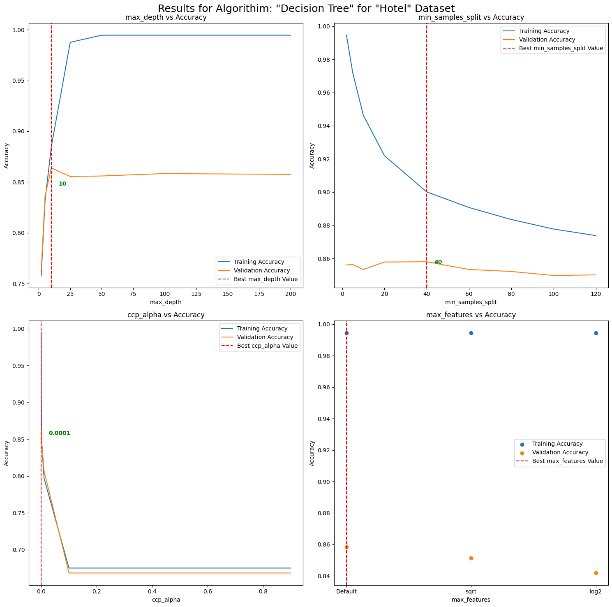
Each algorithm went through hyperparameter tuning. After the hyperparameter various experiments were then conducted on the learner to identify strengths and weaknesses of the learner. Finally, a narrower range of hyperparameters were used alongside GridSearchCV to create an optimizer learner. The optimized learner was used to evaluate the learner. For hyperparameter tuning and experimentation, train and validation data was used. For the final evaluation, training and validation were combined for a mega training set, and evaluation was performed on a separate test set (kept always separate from the data).

# Decision trees

## Introduction

For Decision Trees I did not normalize the data. I chose to do this because decisions trees do not require normalization, and this is particularly noted strength of the tree. When we normalize the data, it adds an additional layer of bias (we can only normalize the training data and assume that is a good representation of the test data). To mimic this, I trained a scikit learn standard scaler on the training data, I used that to transform the validation or test data.

## Hyperparameter Tuning



1. Hyperparameter Tuning for Decision Tree

Here we note as max depth increase.

Link for high dimensional and KNN : chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://bib.dbvis.de/uploadedFiles/155.pdf

Link for sigmoid <https://medium.com/swlh/the-support-vector-machine-basic-concept-a5106bd3cc5f>

NNET:

Sigmoid issues:  
Understandingthedifficultyoftrainingdeepfeedforwardneuralnetworks

DeepSparseRectifierNeuralNetworks

Learning Rate;  
The Need for Small Learning Rates on Large Problems

For learning rate reference for NNet

Sparse Data:

"Sparse deep learning: A review and challenges

" by Lu et al. (2019) - This paper provides an overview of techniques for training deep learning models on sparse

data and discusses challenges and strategies for handling sparsity.

"Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?" by Olshausen and Field (1997) - This classic paper discusses sparse coding in the context of neural representations, which can be relevant for understanding the challenges of sparse data in machine learning.

High-Dimensional Data:

"High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality" by Bellman (1987) - This seminal paper discusses the challenges of high-dimensional data and strategies for addressing them.

"On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes" by Ng and Jordan (2002) - This paper compares the performance of discriminative and generative classifiers on high

-dimensional data.