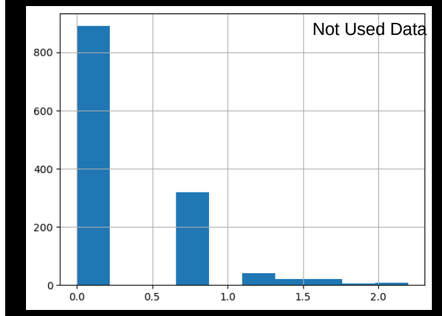
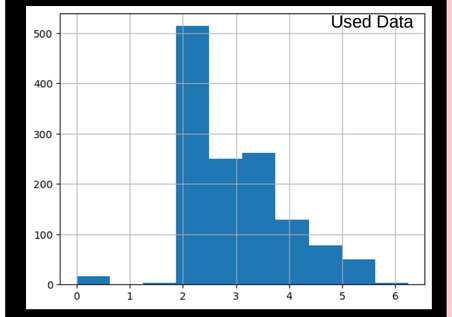
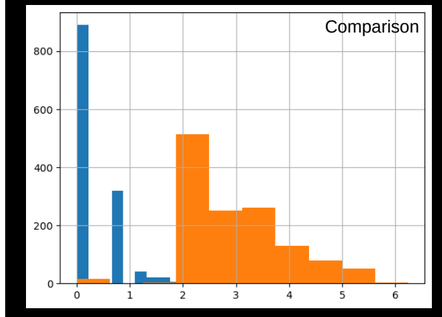
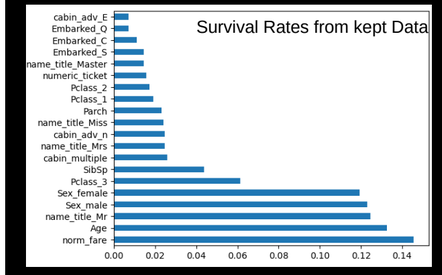








Titanic - Machine Learning from Disaster

Rylei Mindrum
A02352206
CS 4320

Goal

To predict whether or not a person survived the Titanic shipwreck in relation to demographic information regarding to the passengers

Introduction

Background: The sinking of the Titanic is one of the most infamous shipwrecks in history.
Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.
While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

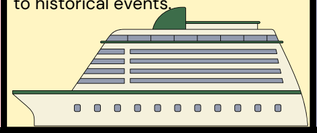
The Challenge: Build a predictive model that answers the question: "what sorts of people were more likely to survive?" using passenger data.

Dataset Description

The Kaggle Titanic - Machine Learning from Disaster Competition is a **Binary Classification** Problem. The starter code provides a large set of data much of which needs splitting or dropping. With an end goal of being able to predict passenger survival based on features.

The missing data within the dataset provides the opportunity for the competitor to display and experiment with handling capabilities.

Beyond the technical skill required in the competition it encourages discussions about the ethic aspects of data science in relation to historical events.



Dropped

- **PassengerId** - No predictive power over survival
- **Name** - Doesn't contribute to survival prediction.
  - However titles from name are kept to reflect social status, age, and marital status.

Kept

Survived: The target survival variable we are trying to predict **(0 = No, 1 = Yes).**
Pclass: Ticket class which is a proxy for socio-economic status **(1 = 1st, 2 = 2nd, 3 = 3rd)**.
Sex: The gender of the passenger **(male or female).**
Age: Age in years. It's fractional if less than 1 and is estimated if in the form of xx.5.
SibSp: Number of siblings/spouses on board
Parch: Number of parents/children on board
Fare - Info about socio-economic stats
Cabin: Cabin number.
- Split into: **cabin_adv** and **cabin_multiple**

Embarked - Provides information about socio-economic status
Ticket - Kept as **numeric_ticket**
Title - From **Name** as **name_title**

Preprocessing

- Dropped **PassengerId** and **Name**
- Did categorical transformations on all of the data.
- Scaled all data with a **0-1 scale** for standardization.

Model Analysis

I have chosen **Random Forest** as the best model for this project. RF combines outputs from **multiple decision trees to reach a single result or score**. A single decision tree does not have the highest accuracy (as seen to the right) but a combination of multiple can provide high accuracy in predictions.

On my tests it scored highest of all the models. With additional tuning I think that a perfect score could be achieved with this model.

Results

Model	Score
Gaussian	.722173
Decision Tree	.776207
Dummy Classifier	0.80143
XGB	.817774
Logistical Regression	.821183
KNN	.830172
SVC	.832431
Random Forest	**.835802**
Competition Best	**1.0**

Conclusion

I was able to use multiple different models to predict survival rates of Titanic passangers based on passenger data. The model that has the highest score was the **Random Forest** Model with a score of .**835802**. This model is the best choice do to its performance during my project as well as its high scores across many different Titanic Kaggle competition entries.

Future Goals

My primary future goal is to **get a perfect score on this project**. This is completely possible with proper tuning of the Random Forest Model.

Plans:
- possible New Model Exploration
- Different Optimization Techniques (Baysain)
- Further redundancy checking (embarked is very similar to Pclass)
- Explore the ethical sides of Titanic related data science