Project Milestone 1

Alexa L. Wittlieff

College of Science and Technology, Bellevue University

DSC680-T301: Applied Data Science (2231-1)

Dr. Catherine Williams

September 4, 2022

Proposal and Data Selection

Topic

One of the greatest beginners Kaggle competitions is “Titanic – Machine Learning from Disaster” in which participants use machine learning to build models predicting which passengers survived this historically significant shipwreck. I will refer to this data science project by the competition name.

Business Problem

In 1912, the “unsinkable” Titanic ship set off on her maiden voyage. Unfortunately, the ship collided with an iceberg and sank. The lack of lifeboats led to the deaths of 1502 of the 2224 passengers and crew. The data science problem I plan to solve is “what sorts of passengers were more likely to survive the sinking of the Titanic?”

Datasets

Kaggle’s Titanic dataset comprising of passenger information will be used for the “Titanic – Machine Learning from Disaster” project. There is one test and one train csv file. The dataset includes the following variables:

- Survival
- Pclass
- Sex
- Age
- Sibsp
- Parch
- Ticket
- Fare
- Cabin
- Embarked

Survival is the target variable for prediction for this project. Zero represents “No” the passenger did not survive, and one represents “Yes” the passenger did survive.

Methods

My analysis methods for the completion of this project may evolve as the project progresses. If the classes are not balanced, a method like SMOTE may be used. Feature engineering may also benefit the model.

Initially, I plan to leverage a k-nearest neighbors (KNN) model to understand if the features of Titanic survivors naturally fall into groups with similar characteristics. I will also likely construct a logistic regression model and a random forest model for evaluation and comparison.

Ethical Considerations

As the fates of the passengers on the Titanic were already determined, the ethical considerations of this project are limited. I will accurately present my data and follow any Kaggle competition rules and regulations.

If this model would be applied for predicting survival odds or directing rescue efforts, then steps must be taken to avoid any bias against groups of passengers based on key demographic details contained in the dataset. For example, a specific gender or ticket class should not be given preferential treatment relative to survival.

Challenges/Issues

As a best practice when embarking on a data science project, the potential risks and challenges must be investigated. The dataset may feature incomplete data. If this occurs, I must decide how to best handle the null values. Additionally, the feature engineering may pose challenges in determining the best features. Despite these potential hurdles, I look forward to embarking on this fascinating data science project, “Titanic – Machine Learning from Disaster.”

References

Titanic - Machine Learning from Disaster | Kaggle. (n.d.). Kaggle. https://www.kaggle.com/competitions/titanic/data