Project Proposal

Team: Cem Beyenal, Mason Menser

https://www.kaggle.com/competitions/titanic/overview

Intentions

We intend to predict whether a passenger on the Titanic survived based on a set of variables. By analyzing passenger data, we can quantify how demographic and socio-economic factors affected survival.

Methods

The first thing we would do is to pre-process the data. Some examples of cleaning the data would be to handling missing values and converting categorical variables into numeric values (gender, port of embarkation, ticket class). After processing the data, we would engage in exploratory data analysis to visualize trends regarding the data to give us a general direction of exploration. The next method would be to build different kinds of models and evaluate them to see which model performs best.

Data

Variable	Definition
survival	0=Died, 1=Survived
pclass	Ticket Class – 1=1st, 2=2nd, 3=3rd
sex	Sex of passenger.
age	Age of passenger (years).
sibsp	# of siblings on board
parch	# of parents on board
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of embarkation

The data will be obtained via the Kaggle competition here. It includes a training set, where we would train the model on the survival of passengers based on the variables given in the data set. The test set would not be used in training, and it would be used to test our model's predictive capabilities.

There are several interesting ways we could manipulate/add to the data. We could engineer new features like family size which can be extrapolated from # of siblings/parents. We can add a marriage variable by looking at title, age, name, and embark. We can also extract the cabin level from the cabin number to come up with localization data.

Background

Relevant background work would include researching and gathering information about the Titanic, specifically about passengers. Where they came from, what nationality, what motivations, etc. Another intriguing aspect could be the location of impact and how it coincides with cabin location.

Next would be understanding the technology we'll use to analyze the data. Random forests, regression, logistic regression, K-nearest neighbors, decision trees, etc.

Watch the Titanic movie by James Cameron.

Tentative Plan

Date	Goal
10/15	Data understanding/cleaning
10/25	Progress Report Due
10/30	Model creation
11/15	Iterating on model
12/1	Prepare findings
12/9	Final Report