Titanic ML Competition

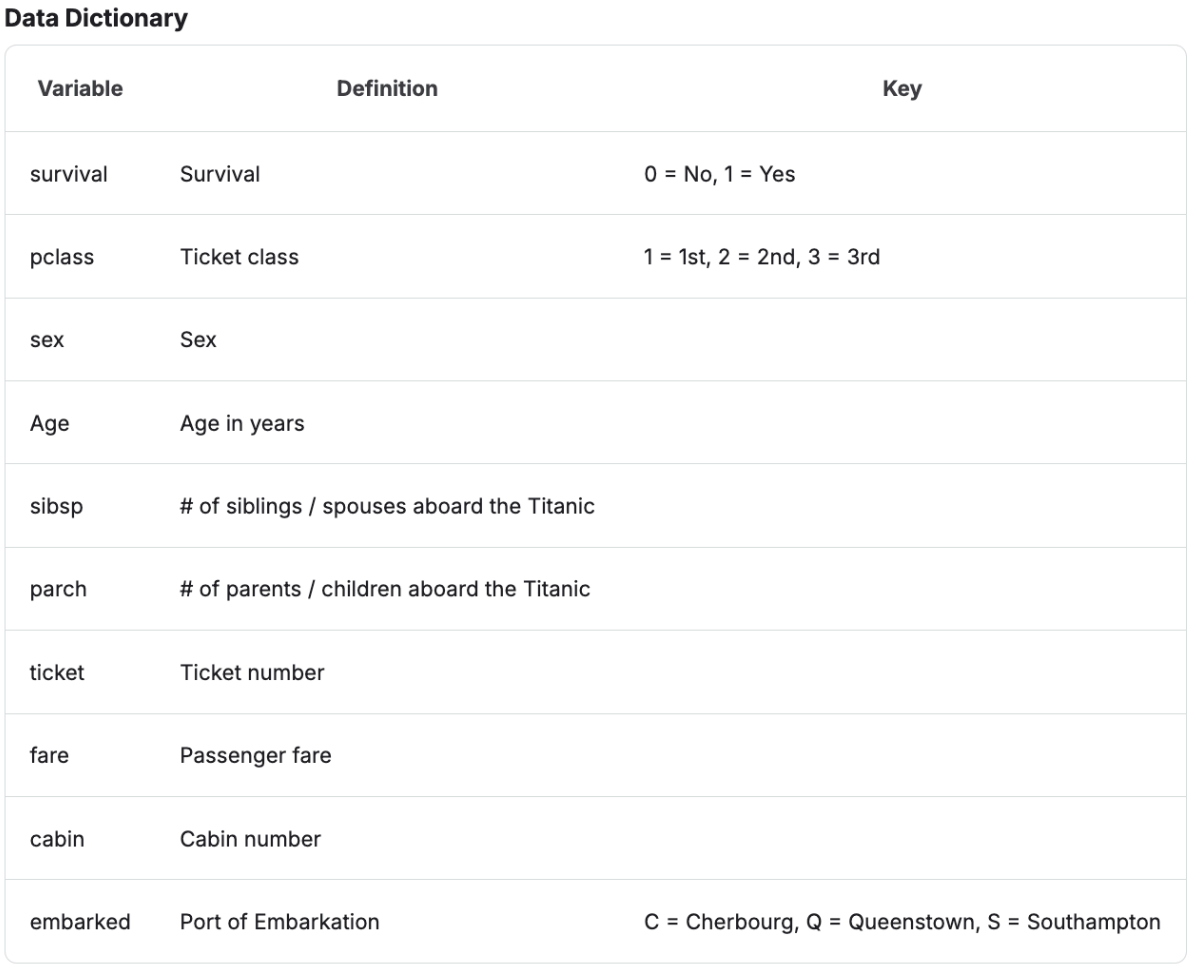
Harry Barrs

Intro

This project is an introductory exploration of using ML models. Kaggle has many ML competitions, their most introductory being the Titanic ML competition. The competition entails creating a model in order to predict whether Titanic passengers lived or died based on recorded information about their person, eg. Sex, Age, Ticket Fare, etc. My goal with this project was to introduce myself to using ML models, with a rather simple dataset and problem.

Data

The data can be found on Kaggle in the data tab of the Titanic ML competition. There are two data sets. One, for training your model (train.csv) and one for testing your model (test.csv). The difference being that the test data does not include whether or not each passenger survived or not. Below is a graphic taken directly from Kaggle's competition page, explaining the columns and information in the datasets.



Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Methodology

My first step was to research the commonly used ML models in order to determine which would best apply to the Titanic competition. Since the model's purpose would be to predict survival, which is a binary outcome (0 = No, 1 = Yes), my instinct and final decision was to proceed with a Support Vector Machine. With the prediction output being strictly binary, the challenge is to properly split input data into two groups, one for each outcome. This aligns perfectly with a Support Vector Machine. Furthermore, after examining the data visually using Tableau, I found that Sex was the strongest predictor of survival. I also decided to proceed only with *pclass, sex, age, sibsp, parch,* and *fare* as my influencing variables. Since sex was by far the strongest predictor, I decided to split my data into two groups (male and female) and train individual SVM models for each group. I hoped that this would better predict outcomes and help with outliers.

Results

I submitted three submission files as tests for my model.

1. The first file was derived using a SVM model set to linear. This received a score of **0.76555** on accuracy.
2. The second file was derived using two linear SVM models, one for female passengers and one for male passengers. This received a score of **0.77751** on accuracy.
3. The third file was derived again using two SVM models, one for male and one for female passengers. However, I optimized accuracy using GridSearchCV on both models to find the best performing settings. This received a score of **0.77751** on accuracy.

Conclusion

All in all, the SVM underperformed per my expectations. I was hoping to get above 80% accuracy. However, this still provided an interesting learning opportunity. I was pleased to see that splitting the data into male and female groups helped train the model better (if only marginally). I was very surprised that my optimized parameters did not yield better results than the basic settings. Were I to continue onward with attempting to get a better score, I would likely try integrating an additional model. However, for the purposes of this case, I am content with using only the SVM model and an accuracy score of 0.77751.

Data Source

Titanic ML competition