Titanic: Machine Learning from Disaster

Ivon Saldivar, B16

Problem

The sinking of the Titanic is one of the most infamous shipwrecks in history. While there was some element of luck in surviving, there were some groups of people who were more likely to survive

The purpose of this competition is to build a predictive model that answers the question “what sorts of people were more likely to survive?” using passenger data (such as name, age, gender, socio-economic status, etc).

Approach

First, I cleaned up the data. This involved normalizing numeric data and discarding or reformatting categoric data. For example, cabin number was simplified down to just cabin deck (from C43 for C deck, for example) to make it easier to one-hot encode.

After cleaning the data, I fed it into two different models: A simple linear regression model, and a more complicated 5-layer model. Both models used Mean Squared Error for their loss functions. The simple model used Stochastic Gradient Descent as its optimizer, whereas the 5-layer model used Adams.

Links

- https://www.kaggle.com/c/titanic/overview
- https://github.com/IvonTSaldivar/Titanic

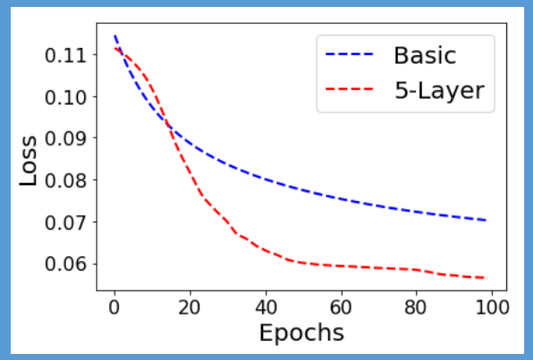
Analysis and Results

From analyzing the training data, I learned that women and first-class passengers were the most likely to survive the shipwreck.

	First Class	Second Class	Third Class
Percent of Total Passengers	23%	21%	56%
Percent of Survivors	39%	26%	35%

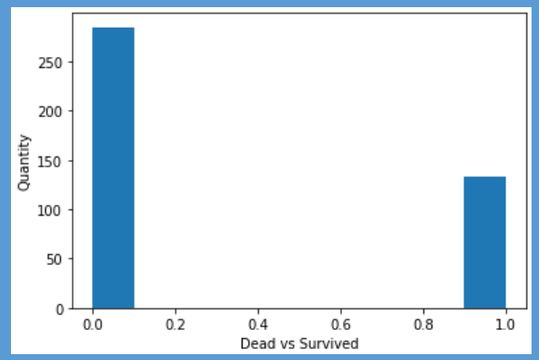
	Percent of Survivors	Percent of Total Passengers
Women	74%	35%
Men	19%	65%

There was no substantial difference between the Kaggle scores of the basic model and the 5-layer model. The 5-layer model scored 78.229% accuracy versus 76.076% for the basic model. In comparison, a submission that assumed there were no survivors achieved 60.22% accuracy.



To the left is the loss values as epochs increase.

To the right is the ratio of dead to survived predicted by the 5-layer model.



Conclusions

My big takeaway from this project was the reason this field is called “Data Science”. The similarity between the results for my two models shows that in this case, there is minimal benefits to be gained from a more complicated model. However, I believe I could increase the accuracy significantly by extracting additional fields from categories that I had discarded initially, such as “Name”, or by simplifying fields, such as “Siblings” and “Parents” into a single “Relatives” category.