J. Jedediah Smith

BIFX-551

Final

5/12/2022

**TASK:** Use what we have learnt to create a model that predicts which passengers survived the Titanic shipwreck. The data will contain the details of a subset of the passengers on board (891 to be exact) and importantly, will reveal whether they survived or not, also known as the “ground truth”. Your model will be based on “features” like passengers’ gender and class.

You are expected to include:

* Feature engineering to create new features.
* Split the data into training and testing to perform the prediction
* Include all metrics (F-1 Score, Precision, Recall, Accuracy)
* Include the code for the model reading directly from the file titanic.csv that should be located in the same folder as the code (no need to submit the file).

**Variable Notes:**

Pclass: A proxy for socio-economic status (SES)

* 1st = Upper
* 2nd = Middle
* 3rd = Lower

Age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

SibSp: The dataset defines family relations in this way...

* Sibling = brother, sister, stepbrother, stepsister
* Spouse = husband, wife (mistresses and fiancés were ignored)

Parch: The dataset defines family relations in this way...

* Parent = mother, father
* Child = daughter, son, stepdaughter, stepson
* Some children travelled only with a nanny, therefore parch=0 for them.

Embarked: Port of embarkation

* C = Cherbourg
* Q = Queenstown
* S = Southampton

**Feature Engineering**

After importing the data, I do several things to improve the information and make it more usable and accurate. First, I binned the continuous variables Age and Fare. Then, I modified the variable Cabin into correspond with a particular deck level on the Titanic. While doing this, I addressed unknowns and N/A values in the data. Rather than being blank, I set them equal to U or 0 depending on the variable.

**Decision Tree**

Once the data was in a usable state, I split it into train and test groups. Approximately 10% of the entries were saved for the test group. My model was created using the train group and I excluded the variables Passenger ID, Survived, Name, and Ticket. These are all either unique values or the result we are trying to predict, so it seemed appropriate to exclude them from the tree. I boosted the model 25 times but this did not appear to have a significant impact on the results.

Non-Boosted Decision Tree

Text

Description automatically generated

Boosted Decision Tree (6 of 25)

A picture containing text

Description automatically generated

Non-Boosted Cross Table

Table

Description automatically generated

Boosted Cross Table

Table

Description automatically generated

**Metric Analysis**

With the models complete, I set about calculating the different metrics. For the original model, sensitivity and recall was 0.9091, specificity was 0.7069, precision was 0.6383, and F-1 Score was 0.7500. For the boosted model, sensitivity and recall was 0.8857, specificity was 0.7143, precision was 0.6596, and the F-1 Score was 0.7561. Over all, the models are very similar, but the final F-1 Score for the boosted model was slightly, but probably not significantly, higher.

Non-Boosted Confusion Matrix w/ Metrics

A screenshot of a computer

Description automatically generated with low confidence

Boosted Confusion Matrix w/ Metrics

A screenshot of a computer

Description automatically generated with low confidence