Capstone Project Proposal

Sam Stelzner - January 2023

Domain Background

This project is based on the Titanic Kaggle competition. The competition is presented as the best introduction to Kaggle competitions. It was introduced in 2012 by Jessica Li, Will Cukierski. Competition entries from the last two months are presented on a rolling leaderboard. At the time of writing, there were 13,731 teams enrolled. It is a well established problem.

The competition forms part of the field of binary classification. As stated on learndatasci.com, binary classification, in machine learning, is a supervised learning algorithm that categorises new observations into one of two classes. Authors Kumari and Srivastava perform a comprehensive review of work done with binary classification with the aim of detecting sockpuppets. They provide a substantial list of types of classifier algorithms used in this area.

Problem Statement

Disaster struck in 1912 when the RMS Titanic sank after hitting an iceberg. 1502 out of 2224 passengers and crew died. It seems that some groups of people were more likely to survive than others. The problem to be solved here is to predict “what sorts of people were more likely to survive?” based on passenger data. Prediction takes the form of a binary 1 for survived, 0 for deceased.

Solution Statement

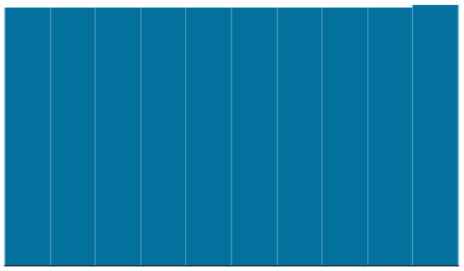
The proposal is to begin by finding an appropriate model using AutoGluon. Then, to perform hyperparameter optimization before training the final model and making inferences. There is also potential to try adding some additional features to the data set and preprocess the data to, for example, impute for any missing values. The project is to be done on the AWS, Sagemaker platform, using the account provided by Udacity for the capstone project of the ML with AWS nanodegree.

Datasets and inputs

Train and test datasets have been made available. Train has 891 unique rows, representing passengers. Test has 418. Train includes a target column: Survived. The features of the data sets, with some descriptions and detail of the distribution of the data within these features for the train dataset, taken from Kaggle, are:

- PassengerId - Unique ID

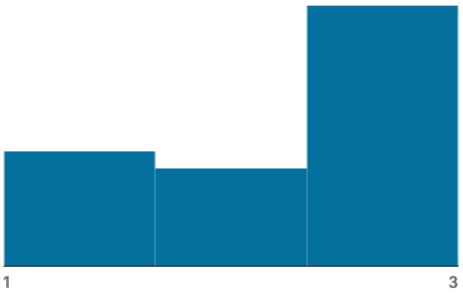
PassengerId



Valid	891	100%
Mismatched	0	0%
Missing	0	0%
Mean	446	
Std. Deviation	257	
Quantiles	1	Min
	223	25%
	446	50%
	669	75%
	891	Max

- Pclass - Passenger class (1, 2 or 3)

Pclass



Valid	891	100%
Mismatched	0	0%
Missing	0	0%
Mean	2.31	
Std. Deviation	0.84	
Quantiles	1	Min
	2	25%
	3	50%
	3	75%
	3	Max

- Name

Name

891
unique values

Valid	891	100%
Mismatched	0	0%
Missing	0	0%
Unique	891	
Most Common	Braund, Mr. ...	0%

- Sex - Male or Female

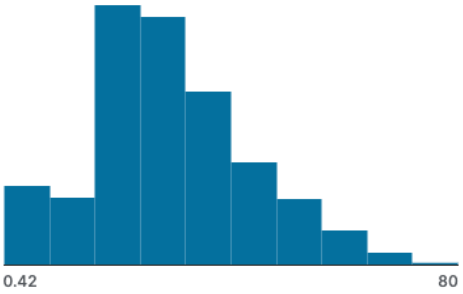
Sex

male	65%
female	35%

Valid	891	100%
Mismatched	0	0%
Missing	0	0%
Unique	2	
Most Common	male	65%

- Age

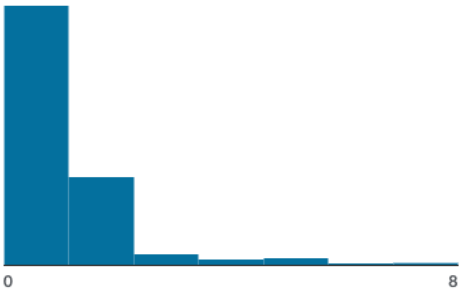
Age



Valid	714	80%
Mismatched	0	0%
Missing	177	20%
Mean	29.7	
Std. Deviation	14.5	
Quantiles	0.42	Min
	20	25%
	28	50%
	38	75%
	80	Max

- SibSp - Total number of passenger's siblings and spouse

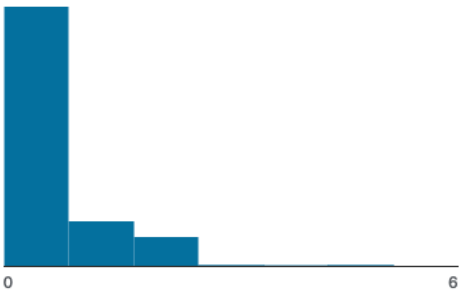
SibSp



Valid	891	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.52	
Std. Deviation	1.1	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	8	Max

- Parch - Total number of passenger's parents and children

Parch



Valid	891	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.38	
Std. Deviation	0.81	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	6	Max

- Ticket - Ticket number

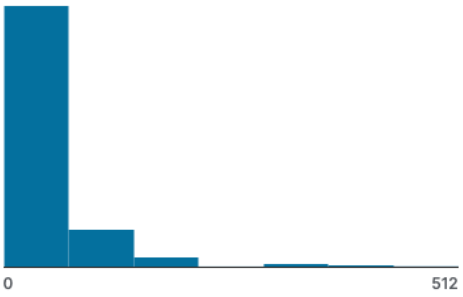
A Ticket

681
unique values

Valid	891	100%
Mismatched	0	0%
Missing	0	0%
Unique	681	
Most Common	347082	1%

- Fare - Ticket price

 # Fare



Valid	891	100%
Mismatched	0	0%
Missing	0	0%
Mean	32.2	
Std. Deviation	49.7	
Quantiles	0	Min
	7.9	25%
	14.5	50%
	31	75%
	512	Max

- Cabin - Cabin number

A Cabin

[null]	77%
G6	0%
Other (200)	22%

Valid	204	23%
Mismatched	0	0%
Missing	687	77%
Unique	147	
Most Common	G6	0%

- Embarked - port of embarkation (Cherbourg, Queenstown or Southampton)

A Embarked

S	72%
C	19%
Other (79)	9%

Valid	889	100%
Mismatched	0	0%
Missing	2	0%
Unique	3	
Most Common	S	72%

- Survived - 0 for deceased and 1 for survived

 # Survived

Valid	891	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.38	
Std. Deviation	0.49	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	1	Max

Benchmark Model

This popular model from Kaggle's public notebooks, upvoted 3755 times, can serve as the benchmark. It makes use of the randomForest classification algorithm and has an accuracy of 0.80382.

Evaluation metrics

The evaluation metric is accuracy of correct predictions of whether a passenger survived or not, between 0 and 100%.

Project Design

A project roadmap can be summarised as follows:

1. Exploratory data analysis - checking, for example, for completeness of data
2. Possible feature engineering and value imputation
3. Using AutoGluon to find an appropriate type of model
4. Hyperparameter optimisation
5. Training
6. Inference

References

- Source of problem: https://www.kaggle.com/competitions/titanic
- Academic paper in the same field of binary classification: https://www.researchgate.net/profile/Saurabh-Srivastava-8/publication/313779520_Machine_Learning_A_Review_on_Binary_Classification/links/5a140771aca27240e30848cf/Machine-Learning-A-Review-on-Binary-Classification.pdf
- Course notes on AutoGluon: https://learn.udacity.com/nanodegrees/nd189/parts/cd0385/lessons/9df6a213-9889-44f7-b85c-782738dea7d8/concepts/b121c5b0-9b82-45fb-909f-c70eb054bbfb
- AutoGluon documentation: https://auto.gluon.ai/stable/index.html
- Detail on train dataset from Kaggle: https://www.kaggle.com/competitions/titanic/data?select=train.csv
- Benchmark solution: https://www.kaggle.com/code/mrisdal/exploring-survival-on-the-titanic
- Model documentation from benchmark solution: https://cran.r-project.org/web/packages/randomForest/index.html