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. At the time of writing, there are 13,731 teams that have participated. It is a well established problem.

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

My plan 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.

Datasets and inputs

Training and test datasets have been made available. Training includes a target column: Survived. The features of the data sets, with some descriptions, are:

- PassengerId - Unique ID
- Pclass - Passenger class (1, 2 or 3)
- Name
- Sex - Male or Female
- Age
- SibSp - Total number of passenger’s siblings and spouse
- Parch - Total number of passenger’s parents and children
- Ticket - Ticket number
- Fare - Ticket price
- Cabin - Cabin number
- Embarked - port of embarkation (Cherbourg, Queenstown or Southampton)

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
- 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
- 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