KAGGLE COMPETITION PREDICTIONS REPORT

PREDICTION 1: HOUSE PRICES

PREDICTION 2: TELSTRA NETWORK DISRUPTIONS

DONE BY:

ESTHER NGOMBE NTUMBA

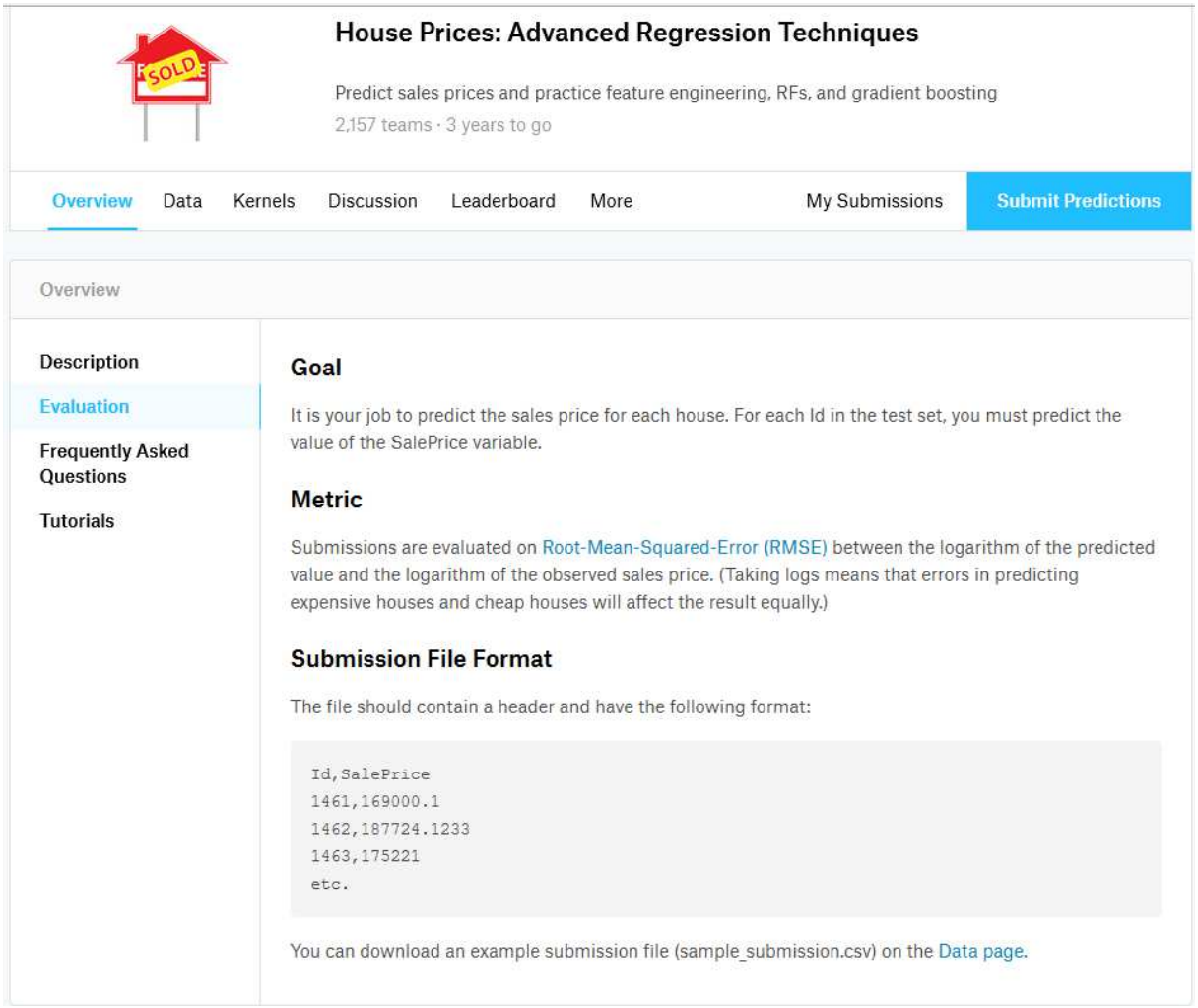
MOTIS8

MAY 2017

PROJECTS' GOALS:

1. HOUSE PRICES

My job was to predict the sale price for each house. For each Id in the test set, I had to predict the value of the Sale Price variable in the train set.

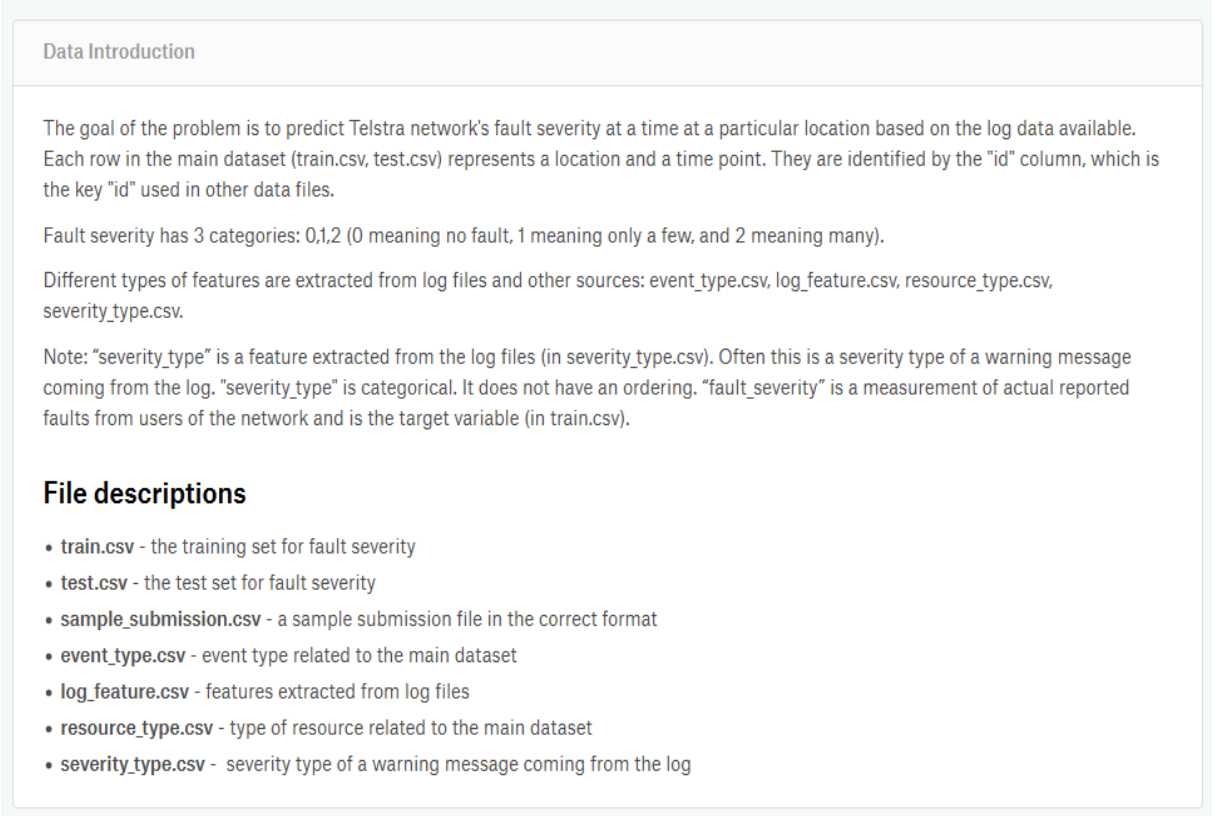


I have chosen this project because as a beginner in data science, this topic seemed to be easy and interesting for me to work on.

At first, after making my choice, I was wondering whether I would be able to work on this project…after reading the tutorials and making more research, I believe I did the right job that was assigned to me.

2. TELSTRA NETWORK DISRUPTIONS

The goal of the problem was to predict Telstra network's fault severity at a time, at a particular location based on the log data available.



Data Introduction

The goal of the problem is to predict Telstra network's fault severity at a time at a particular location based on the log data available. Each row in the main dataset (train.csv, test.csv) represents a location and a time point. They are identified by the "id" column, which is the key "id" used in other data files.

Fault severity has 3 categories: 0,1,2 (0 meaning no fault, 1 meaning only a few, and 2 meaning many).

Different types of features are extracted from log files and other sources: event_type.csv, log_feature.csv, resource_type.csv, severity_type.csv.

Note: "severity_type" is a feature extracted from the log files (in severity_type.csv). Often this is a severity type of a warning message coming from the log. "severity_type" is categorical. It does not have an ordering. "fault_severity" is a measurement of actual reported faults from users of the network and is the target variable (in train.csv).

File descriptions

- **train.csv** - the training set for fault severity
- **test.csv** - the test set for fault severity
- **sample_submission.csv** - a sample submission file in the correct format
- **event_type.csv** - event type related to the main dataset
- **log_feature.csv** - features extracted from log files
- **resource_type.csv** - type of resource related to the main dataset
- **severity_type.csv** - severity type of a warning message coming from the log

This work has been given by the professor during the course examination, because he found us capable of working it out.

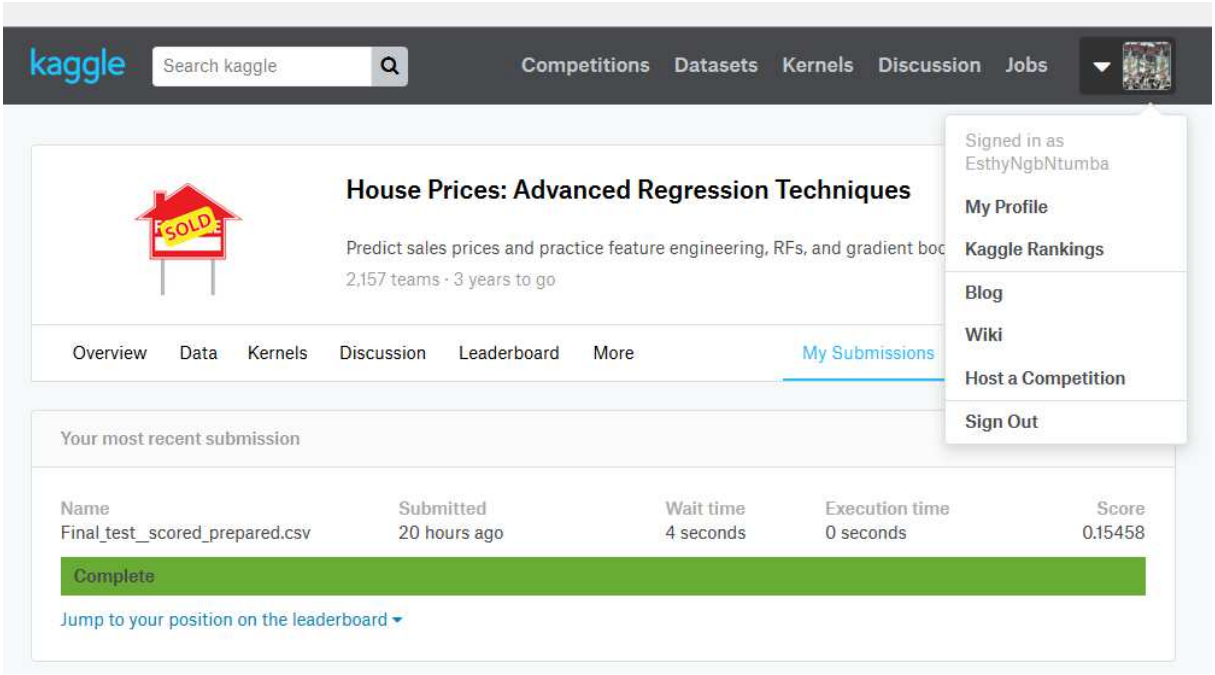
For both the projects, I have used the DSS platform with a little help from excel to solve each case.

A very good experience I had, especially the fact that at the beginning I had no clue at all to how to use it as my laptop could not run the program properly…

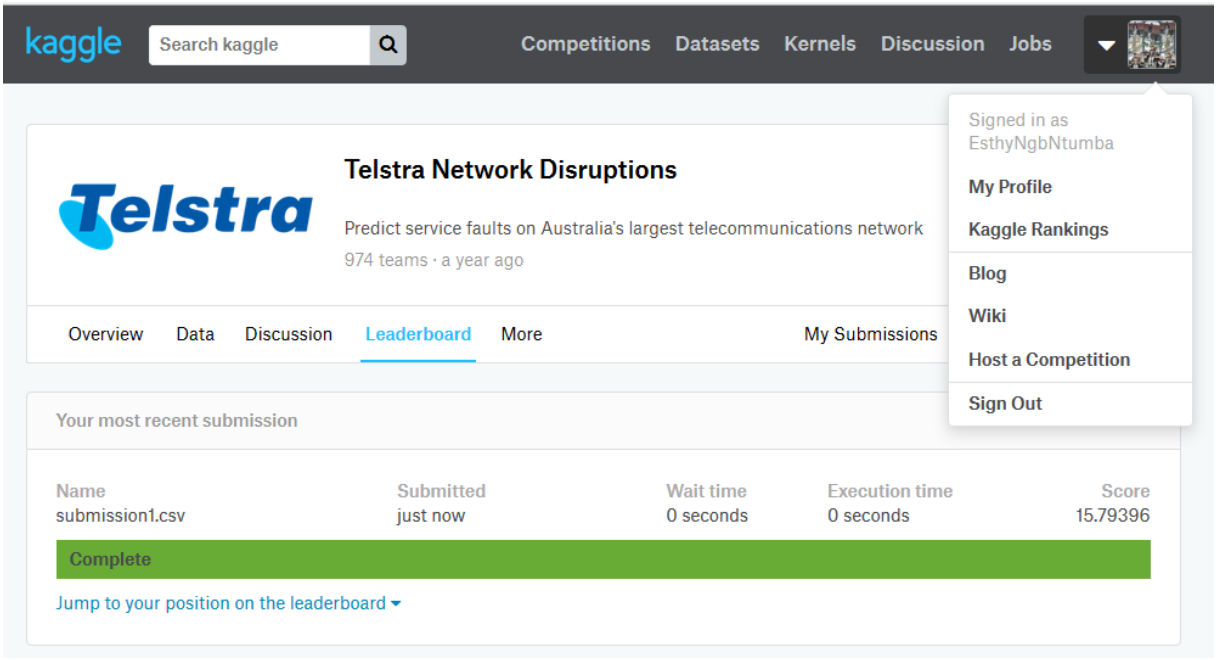
PREDICTION RESULTS

We were asked to submit all our KAGGLE competions' predictions results on the website with the aim of being among the top 500 leaderboard members to be awarded. Here are the result of the work I have submitted, hoping to be one of the best.

1. House Prices



2. Telstra Network Disruptions



REFERENCIES

https://www.kaggle.com/c/house-prices-advanced-regression-techniques#description

https://www.kaggle.com/c/telstra-recruiting-network

http://6xqkiwbr.i.cloud.dataiku.com/projects/HOUSEPRICESV/datasets/testA__scored_prepared/explore/

http://6xqkiwbr.i.cloud.dataiku.com/projects/TELSTRANETWORKDISRUPTIONS/datasets/submission1/explore/

http://6xqkiwbr.i.cloud.dataiku.com/projects/TELSTRANETWORKDISRUPTIONS/datasets/test_scored/explore/

THANKS