Hackathon 2021

June 4, 2021

Yonatan Shemesh, Noam Kesten, Roy Kreiner, Kfir Mizrahi

0.1 Choice Of Problem

We have chosen to take task 1 - predicting movie revenues and average votes, given general movies data.

What lead us to take this task is the fact that the data given seemed richer in terms of features.

0.2 Presenting our work

Most of our work boiled down to feature analysis. For that sake, we have utilized the tools we have learned

in the course: Analyzing the covariance and Pearson correlation between the different features. For

demonstration, see the HTML included in the zipped folder, where we exported a complete profile

of the features. In the meanwhile, we have established a pipeline used to train and evaluate different regression models;

some seen in class, and some we have not seen but read about (see model_eval.py module).

0.2.1 Feature Analysis:

Some Idea's on how we have managed to squeeze more out of the given features:

1. Adding a "superstars" columns: we have presented one hot encoding which determine weather

a famous actor is present in the movie. Analyzing the correlation, we have managed to see as expected that

known actors are correlated with high revenue and average vote.

2. Learning the date: we have embedded a "time passed" column to the data, allowing the model to learn

how to consider the time passed between release and prediction.

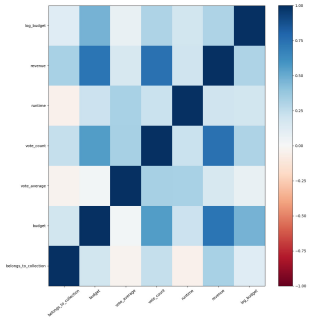
3. Considering inflation: we have added inflation information in the US, which lets the model consider more complex cases.

4. implementing ordinal data: We have divided some features into ordinal data, i.e. a director with higher revenue > a

director with lower revenue.

5. Combinations of features with interesting semantics: log(budget), $\frac{\text{budget}}{\text{runtime}}$ etc.

Some snapshots of feature analysis:



0.2.2 Model Choosing:

We followed the best by test attitude.

In model_eval.py we have built a pipe for evaluating a learner. After tuning hyper parameters, and running

training, we have pickled the model to disk only If it's performance increased on the previous combination of

hyper parameters. The models that did best where Linear Regression and Random Forest Regression, both we

have learned about in class.