**Nielsen Ratings Predictor**

**Sandra Rajanovic SW17/2014**

1. Motivation

Nielsen is the leading measurement company when it comes to TV ratings, and all TV networks rely on their service. Predicting future TV ratings for their prime-time shows is something they all need in order to know how well will their future shows preform amongst the audience, but it’s still a question that hasn’t been fully addressed.

2. Research questions

My goal here was to try to predict ratings of the episodes based on various data. The dataset was manually made, and it consists of attributes that I’ve considered crucial when it comes to this topic. The attributes are the name of the show, name of the TV network that the show is on, genre, season number, episode number, year when it was aired (or more precisely – season, which can be fall or spring season), day of airing, time of airing, how long the episode is and the number of viewers. At the end, I’ve discarded the season number, as it turned out that parameter wasn’t helpful. Everything else contributes to the final model. The day of the airing matters, because e.g. shows aired on Friday rate less, the time is also important because shows aired later rate less, season matters too because episodes aired in fall season rate higher than the ones aired in spring, etc.

Since the data was collected manually, there are around 2500 samples split in train and test dataset. It took a lot of time, otherwise there would’ve been more.

3. Related work

There isn’t a lot of related work out there, because as I’ve said, it’s a question that hasn’t been fully addressed yet. The ones that I’ve found have used ensemble models (GBM) and linear regression.

4. Methodology

I’ve decided to use ensemble models, because I’ve read a paper where they’ve used ensembles and it seemed like the best solution. Since this is a problem that can be solved with regression, I’ve tried three models: Random Forest Regressor, Decision Tree Regressor and Extra Trees Regressor.

5. Discussion

First thing I did is convert categorical values to numerical, and I did that using LabelEncoder from sklearn library.

Next thing was to normalize my dataset, and then I tried three ensemble models.

I’ve tried Random Forest Regressor first with the following parameters: n\_estimators=100, max\_depth=None, max\_features=9, min\_samples\_split=2.

Next was Decision Tree Regressor with the following parameters: criterion='mae', splitter='best', max\_depth=None, max\_features=9.

At last I’ve tried Extra Trees Regressor with the following parameters: n\_estimators=700, criterion='mse', max\_depth=None, min\_samples\_split=2, max\_features=9.

The results were evaluated with RMSE. Random Forest Regressor has RMSE of 0.17, Decision Tree Regressor had RMSE of 0.32 and Extra Trees Regressor had RMSE of 0.14.

Extra Trees Regressor gave the best results. Since the ratings to be predicted go in range of [0.3, 5.2], I consider the error of only 0.14 to be pretty good.

6. References

PDFs included with the project.