Background Information

A hospital or insurance provider is interested in efficiently extracting numeric ratings from patients' written review. To this end we build a model using labelled, numerically, patient reviews.

Data exploration

The data comes from Drugs.com and is accessed through UCI's website.

![Ratings distribution]('./Images/Ratings distribution.png')

Ratings are not normally distributed. Counts are highest at the worst and best ratings.

- 160,000 samples
- 800 unique conditions
- 3400 unique drugs

Data Understanding/Preprocessing

- Tokenizing and creating the tf-idf matrix.

Data modelling

- Baseline model (TF-IDF) with linear regression

- Decision Tree regression models

- Word embedding models.

Results/conclusions

![rmses]('./Images/rmses.png')

Conclusions

- Deployment for ‘rating extraction’ from written review.

- Gather insights on how patients rate drugs.
    - "doctor, love, worse" etc.