Sentiment analysis: drug reviews

Jonathan Marks

Business Problem

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

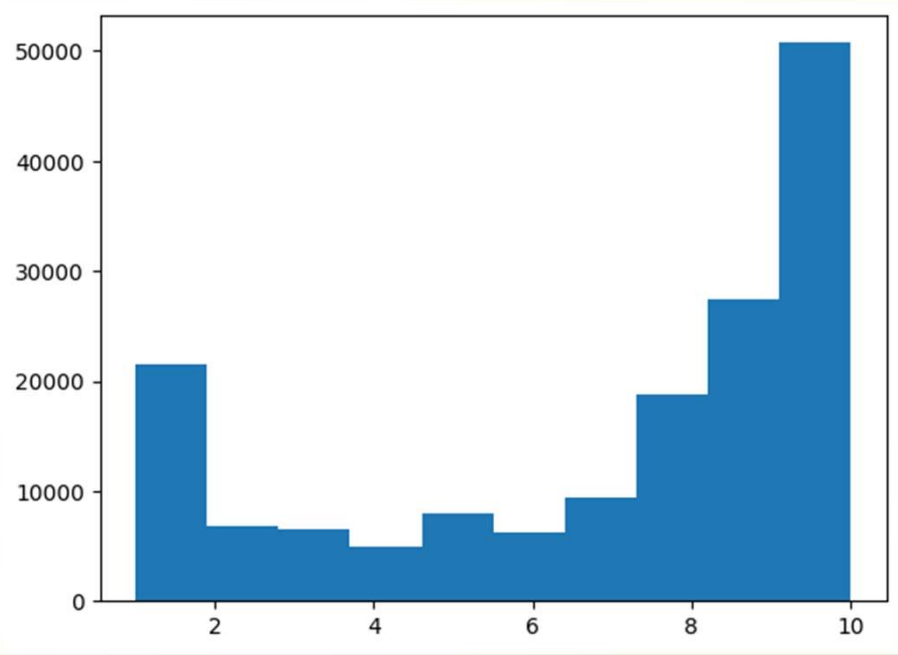
Data Understanding /Preprocessing

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

- 160,000 samples
- Short paragraphs
  - 800 unique conditions and
  - 3400 unique drugs

- Non-normal distribution of target

- Text and meta-deta



Preprocessing / nlp techniques

Tokenizing

TF-IDF

Word embeddings

(Queen is to king as women is to man.)

RMSE cross-model comparisons

- Linear regression model performs best.

	No word embeddings (TF-IDF)			Word embeddings	
	Baseline Decision Tree Regressor	**linear regression**	Random Forest Regressor	linear regression	Decision Tree Regressor
Train rmse	3.38	**2.84**	3.44	3.03	3.14
Val rmse	3.38 (3.09 w/deeper tree (5 leaves).	**2.88**	3.44	3.10	3.22

Recommendations/future work

- Deployment of linear regression model
- Gather insights on how patients rate drugs
  - "Doctor", "horrible", "worse", "love"
- Combine the tf-idf and word embedding models.
- Use the "meta-data" as features. (i.e. the drug evaluated)

Thank you, questions?