Fork me on GitHub

Easy Review Sentiment Analysis with `pandas` and `scikit-learn`

Arnout Devos
indy.epfl.ch

Problem

Predict whether a new review has a rather positive (+) or negative (-) sentiment, given the experience from known labeled (+/-) reviews.

Dataset

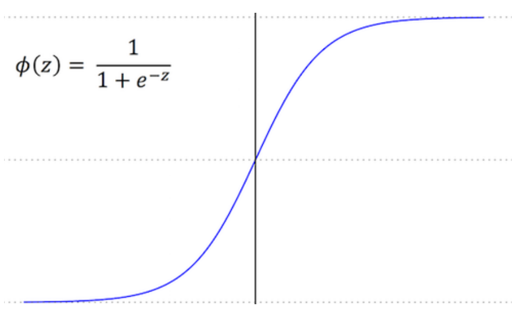
Around 400,000 Amazon product reviews
Sentiment data: 50% positive, 50% negative

Preprocessing

Python: `pandas` and `scikit-learn`
80% Train, 20% Test (every 5th sample ∈ test)
Convert *textual reviews* into *numerical vectors* with `CountVectorizer()`

Algorithms

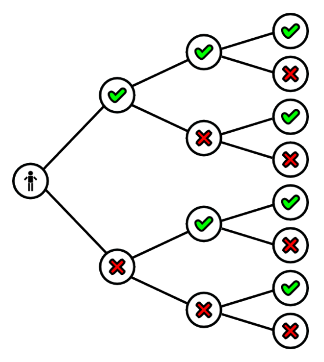
1. Logistic Regression



- Natural choice for binary classification problems with vectors
- Simple interpretable model (coefficients)

`LogisticRegression()`

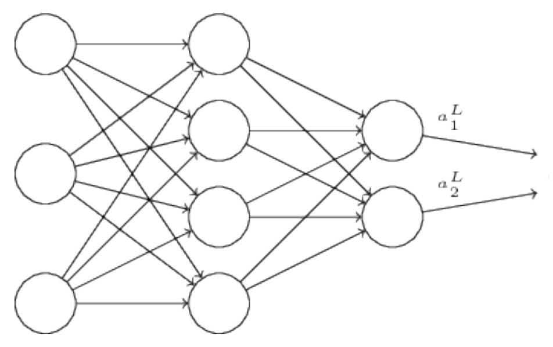
2. Decision Tree



- Natural choice for decision problems in general (low-dimensional, human-size)
- Simple interpretable model (tree)

`DecisionTreeClassifier()`

3. Neural Net

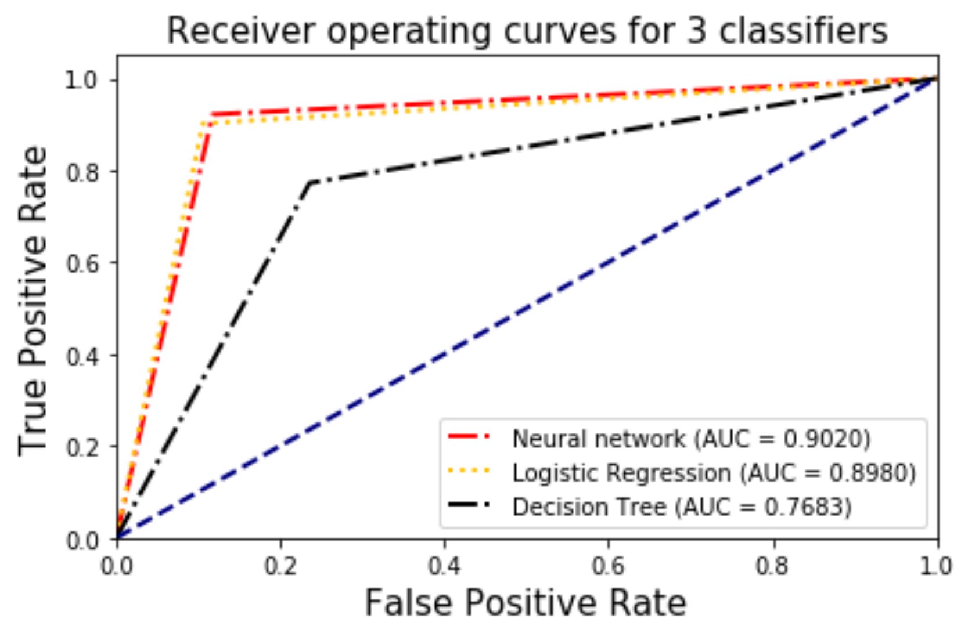


- Popular choice with Word2Vec features
- Can fit highly nonlinear functions/ distributions

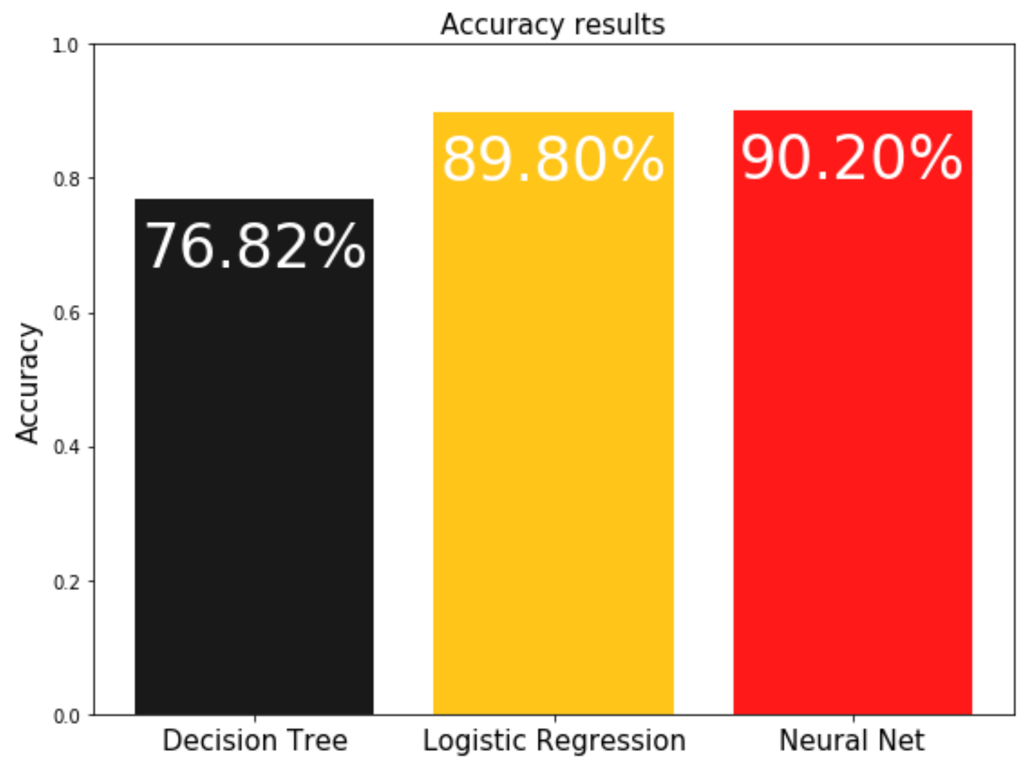
`MLPClassifier()`

Results

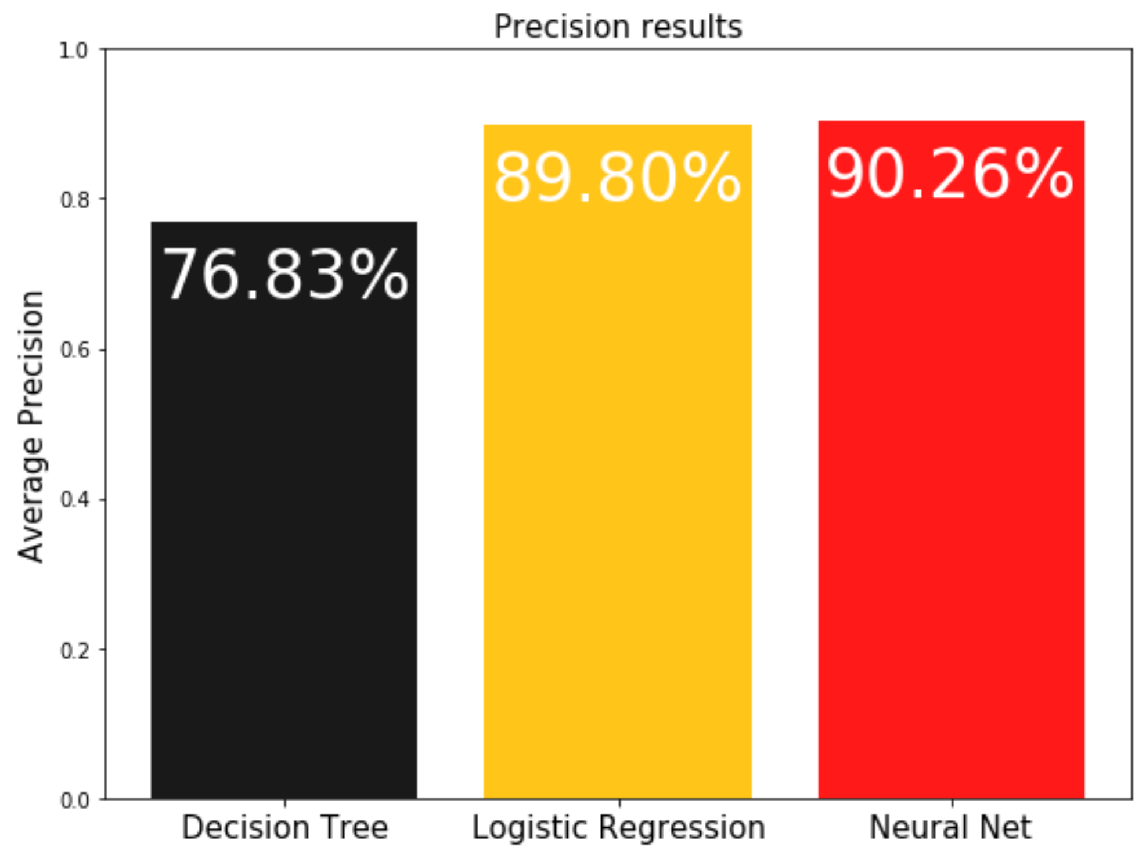
ROC



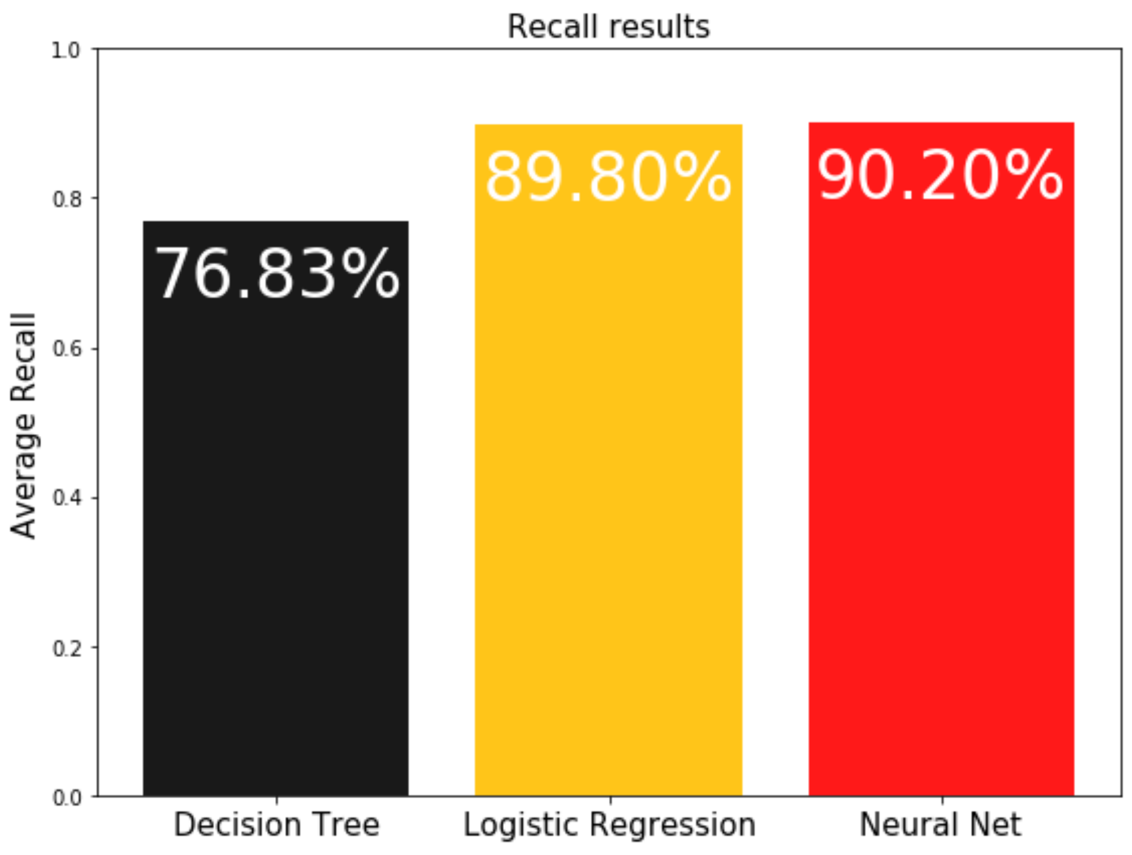
Accuracy



Precision



Recall



- ***Neural Net*** shows the best overall performance
- ***Decision Tree*** has trouble with high dimensionality and performs the worst
- ***Logistic Regression*** benefits from vectorization, but performs slightly worse than the Neural Net