 **How helpful is your review?**

Junze Bao, Wensi Yin, Sandro Barnabishvi

# Motivation & Goal

Hard to identify the quality of online products simply via images, or videos. Other customers’ reviews are more convincing than product descriptions. Identify prominent features of helpful reviews. We define **helpfulness** as the ratio of number of likes over total number of votes

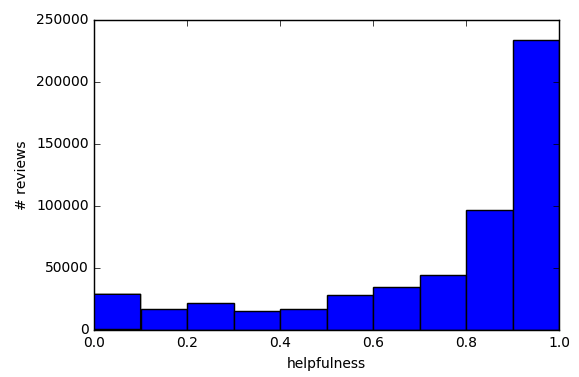
# Dataset

100GB+ json file containing users’ reviews on Amazon products

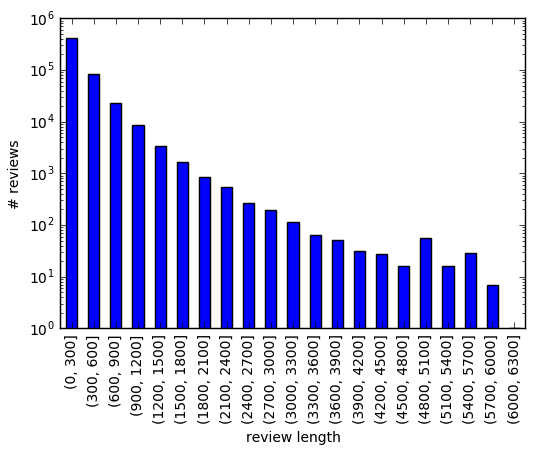
|  |  |
| --- | --- |
| Category | Size (Bytes) |
| Beauty | 731701483 (~730MB) |
| Books | 6952783152 **(~7GB)** |
| Electronics | 2110130573 **(~2GB)** |
| Kindle Store | 12501007 (~12MB) |
| Musical Instruments | 101143085 (~100MB) |
| Office Products | 258605466 (~250MB) |
| Pet Supplies | 324027342 (~320MB) |

# Statistics

1. *Helpfulness* distribution (more than 10 votes)



1. *Review length* distribution

**

# Features

|  |  |
| --- | --- |
| Category | Features |
| Intrinsic | review score  review length  summary length |
| Social | item rating (mean)  user rating (mean) |
| Linguistic | # various punctuations  # pronouns |
| Psychological | sentiment (anger, anxiety, etc.)  insight, analytic, authenticity |

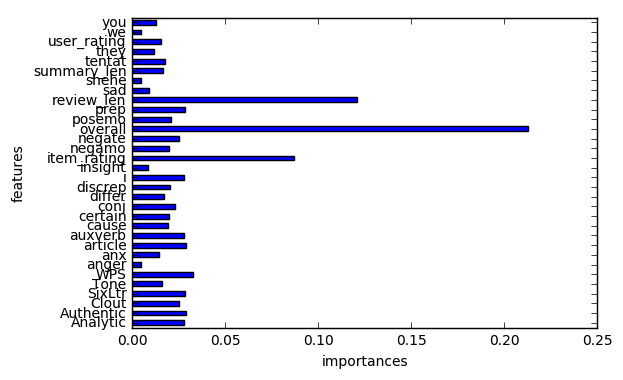
# Regression

MSE of different models of Electronic category

|  |  |
| --- | --- |
| Model | MSE |
| Multilayer Perceptron | 0.065 |
| Random Forest | 0.036 |
| Support Vector Regression | 0.257 |
| Ridge Regression | 0.056 |

# Feature Importance

Use random forest to see feature importance



Relationship between review length and helpfulness

# 

# Conclusion

* Most people tend to write short reviews and most reviews with at least 10 votes are useful
* Random Forest achieves the least error
* Most common important features cross all categories are ***review score, review length, item average rating, WPS (word per sentence) and i***.
* Different categories have specifically important features, e.g. *negate* in Beauty category