Phase 5 - Capstone Project

Stack Overflow - Question Modeling for Quality-of-Life

Will Dougherty - Flatiron School - Online Data Science

Overview

Stack Overflow Question Modelling

- Quality Prediction

- Tag Suggestion

Goals

- In-the-moment text analysis
  - **Alert** users if post might be low-quality
  - **Suggest** tags based on text of the question

- **Save time and resources** for Stack Overflow's moderation team
- Improve **quality** of questions
- Improve **search** and **related-post** features

Question Quality

Dataset and Methodology

“60k Stack Overflow Questions with Quality Rating”

- Classifies posts as ‘high’ or ‘low’ quality

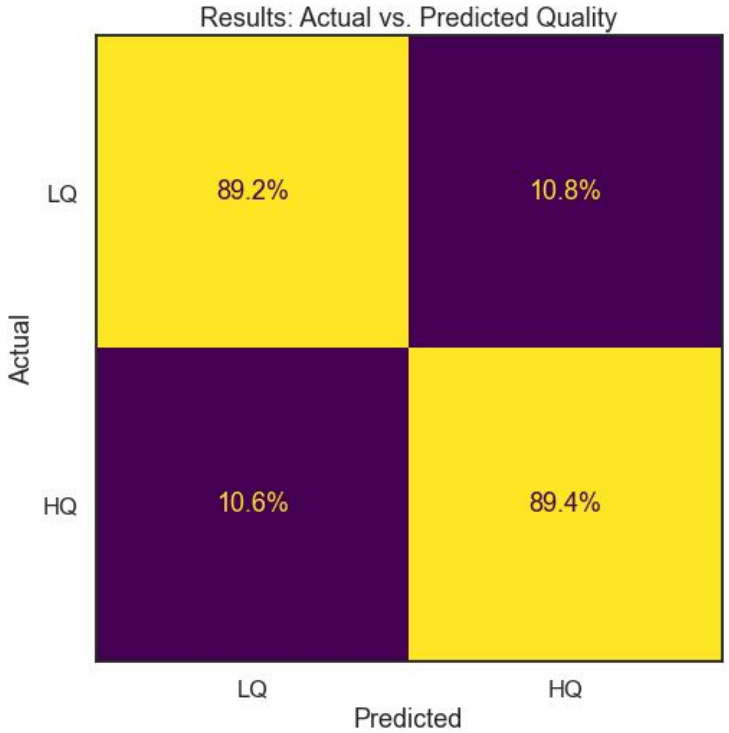
Model Details

- Using text to predict target
- Text cleaning, vectorization, target class balancing
- Logistic Regression

Results

The model does well in distinguishing between the two classes

89% of both the ‘high’ and ‘low’ quality posts are correctly identified



Tag Suggestion

Dataset and Methodology

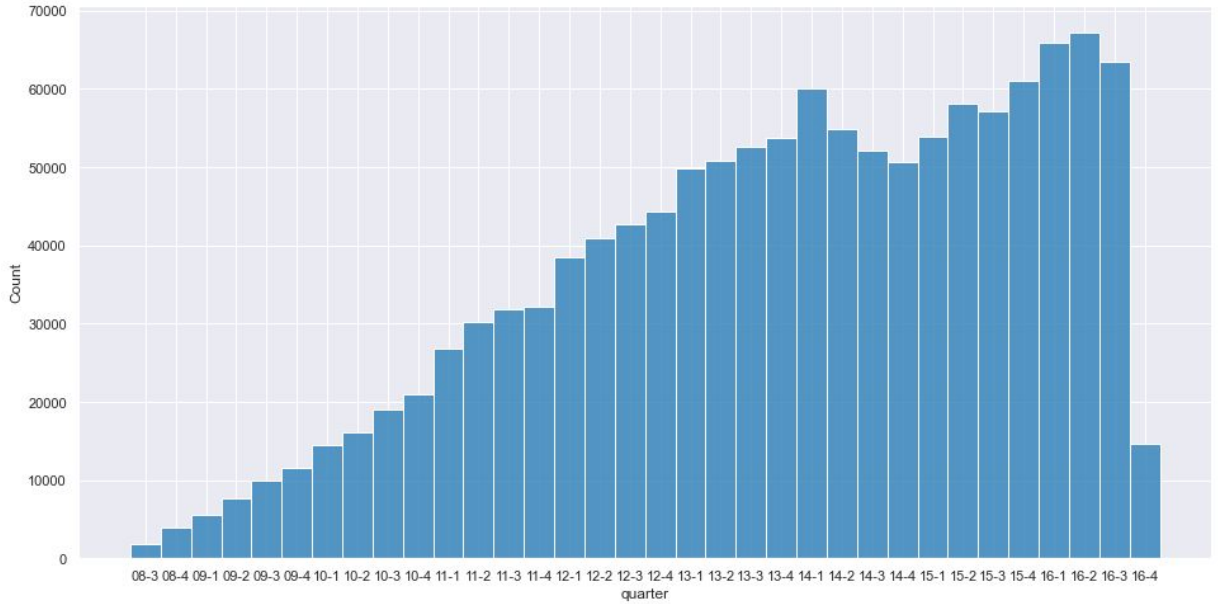
“StackSample:
10% of Stack Overflow Q&A”

- Full set: **1.2 million** posts
- Modelling set: **200,000** posts

Modelling

- One-vs-Rest Classifier
  - one model per tag
- Predict probability for each tag

Posts per Quarter - 2008-2016

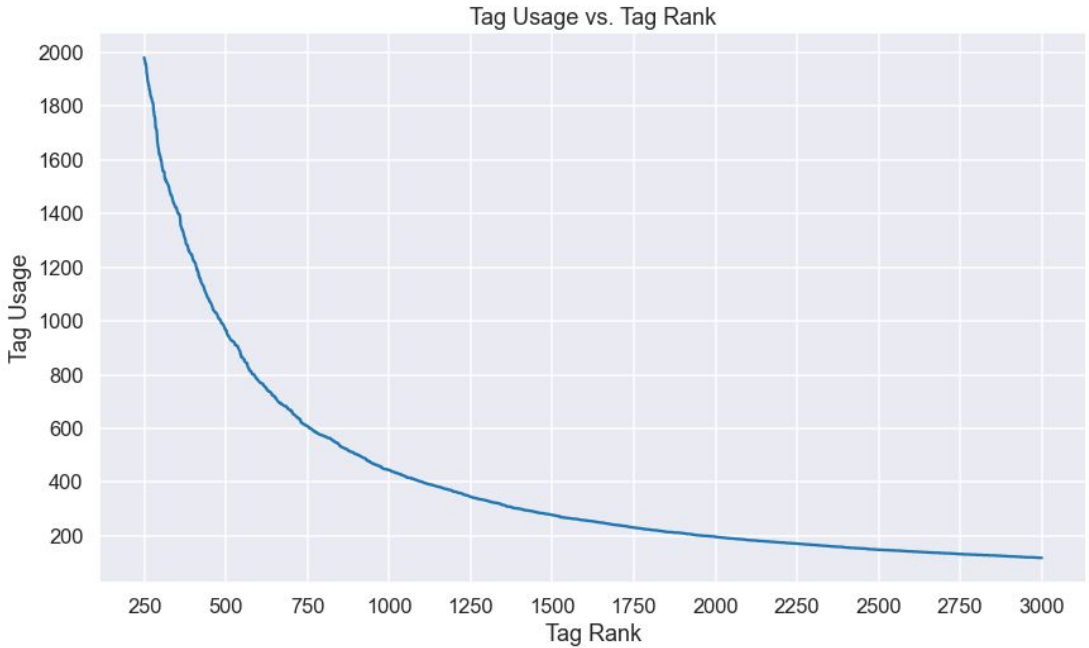


Top Tags

37,000 unique tags - very few account for most tag uses

Cutoff point of **2,000** tags gives model access to:

- **5%** of unique tags
- **83%** of tag uses
- **98%** of posts



Results

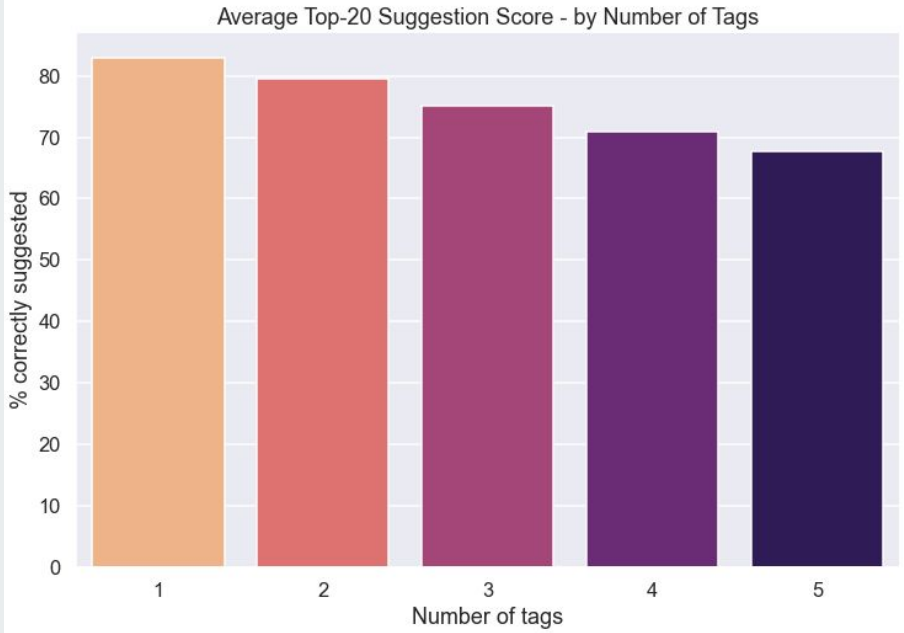
Model gives probabilities that tags apply to a given post

Tags are ranked, and a top-20 list is generated

- **89%** of modelled tags are correctly suggested
- **75%** of all tags are correctly suggested

Posts with lower # of tags are modelled more accurately

Ranging from **83-68%** of tags correctly suggested



Conclusions

Features seem feasible to implement

Quality Prediction Alerts

- Even with small dataset, high accuracy is achieved
- Better post quality, gives chance for users to edit question before posting

Tag Suggestion

- Remind users, save time
- More thorough tagging

Limitations, Future Work

- More samples/tags need more resources
    - Though most tags have very few uses, so they are hard to model
- Further modelling opportunities
    - Prescriptive quality alerts
    - Tag trend prediction

Thank you!