INFX 575 - Data Science III Project

Stack Overflow Data

Presented By:
Ayoush Mukherjee ~ Shipra Gupta ~ Shreya Agarwal

UNIVERSITY of WASHINGTON

Agenda

- Why Stack Overflow?
- Prior Research
- Research Questions
- Preliminary Findings
- Techniques
- Potential Challenges

Why Stack Overflow??

- Stack Overflow is the largest online community for programmers to learn, share their knowledge, and advance their careers.
- It aims to bring solutions to day to day problems faced by them.
- Stack Overflow offers a great interface for accessing all of its data and running any possible query in the questions/answers database.

Prior Research: What drives the SO Community

- Stack Exchange has a clear purpose
- The use of a voting system
- The use of reputation and badges
- The great content available
- The awesome moderation system
- The possibility to edit someone else's content
- A dedicated site just for problems, suggestions and bug reports
- Really good search options like combine tags, exclude tags, specific user, specific range of votes, specific range of number of answers, is closed, etc.

Research Questions

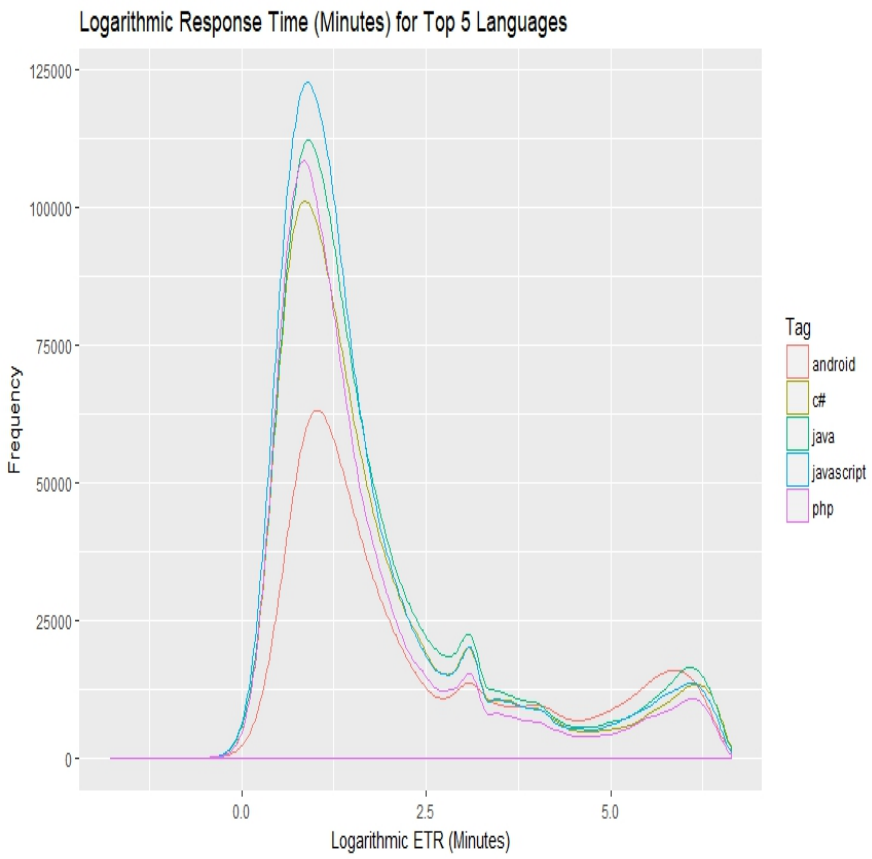
- Predicting how soon/late will a question posted on Stack Overflow get a response from the community

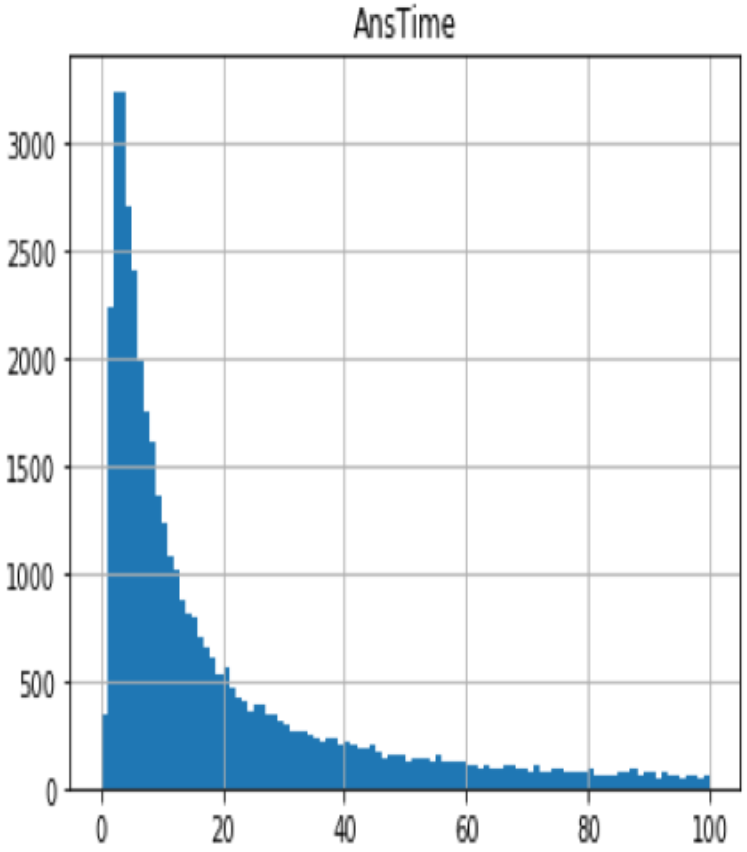
A) Based on the subject-matter of the question.

B) Based on other features of the question.

- Predicting which user is likely to answer a question posted on Stack Overflow from the top 15 users

Preliminary Analysis





Top Users Preliminary Analysis

For approaching the question of predicting who will be answering a question, we wanted to check for these top 5 languages, who are the top 5 users for each language and their response count. Below table shows top 5 users and their response count:

Language	UserID-Count	UserID-Count	UserID-Count	UserID-Count	UserID-Count
Javascript	19068 - 831	157247- 819	114251- 652	1048572-635	816620- 612
Java	22656-1774	23354-1026	17034-840	29407-702	34397-587
C#	115145-1811	1202025-324	501696-301	653856-273	1631193-228
PHP	118068-752	285587-511	476-511	1491895-430	367456-404
JQuery	114251-908	965051-535	519413-488	157247-429	13249-423

Techniques

> Text Preprocessing

- ❑ Noise Removal, Lexicon Normalization (Lemmatization, Stemming), Removing stop-words, TFIDF, bag of words(trigram), word2vec

> Machine Learning Models

- ❑ Naïve Bayes Classifier, Stochastic Gradient Descent, Support Vector Machine, Random Forest

> Evaluation Methods

- ❑ Confusion Matrix, Recall and Precision, Accuracy

Predicting Response Time for a Question

> Text Classification – Question Body

Model Used: Multinomial Naïve Bayes Classifier

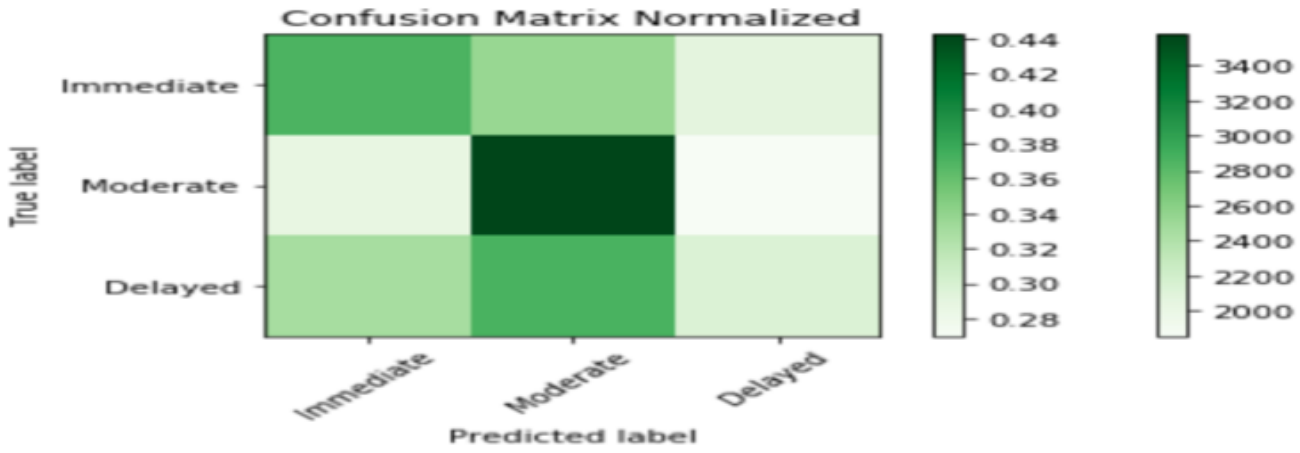
44% Accuracy

	precision	recall	f1-score	support
Delayed	0.40	0.71	0.51	6631
Immediate	0.52	0.49	0.51	7938
Moderate	0.39	0.13	0.19	7052
avg / total	0.44	0.44	0.41	21621

Predicting Response Time for a Question

- Model Used: Naïve Bayes and Random Forest
- Features: Tag Count, Question Score, Has Code, Has Link, Question Body Length, Question Title Length

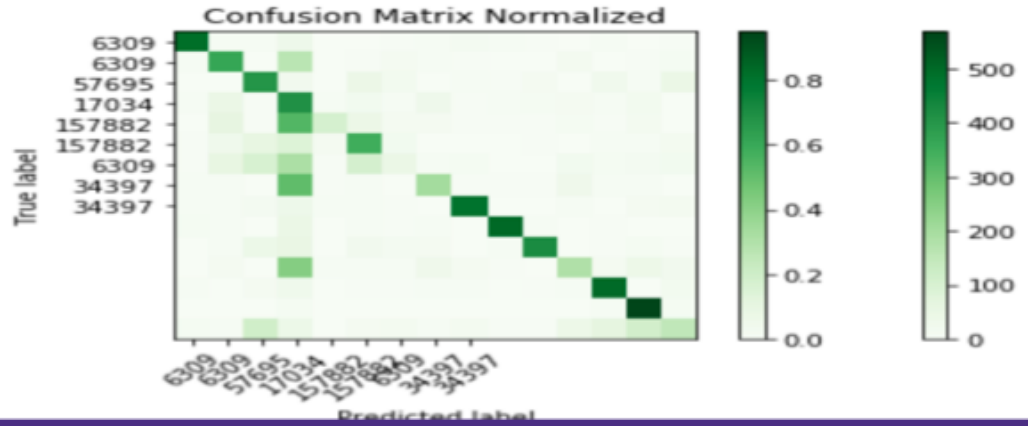
```
accuracy 0.373386984876
confusion matrix
 [[2359 2145 1854]
 [2327 3577 2181]
 [2370 2671 2137]]
(row=expected, col=predicted)
```



Predicting the Answerer for a Question

> Model(s) Used: Linear Support Vector Machine

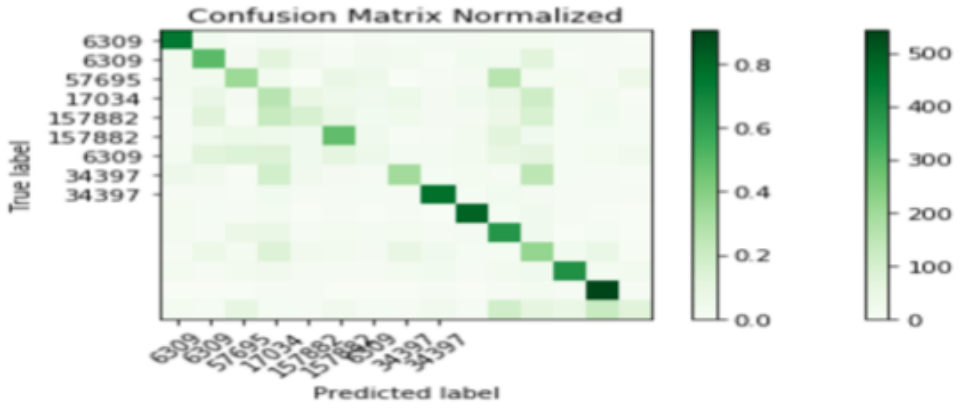
```
accuracy 0.626619110193
confusion matrix
 [[240   3   4  18   1   1   2   4   7   3   2   0   3   1   2]
 [  3 195   0  86   1   2   7   4   2   1   0   9   0   2   5]
 [  1   0 204  12   1  22  10   0   4   2   6   0  13   4  23]
 [  5  50   3 438   5  23   9  37   3   6  10  13   7  20   0]
 [  0  26   1 136  44  18   5   5   1   1   1   4   1   7   0]
 [  2  21  41  61   9 245  15   1   1   0   2   1   5   5  11]
 [  2  29  50  94   2  53  24   3   5   1   1  10   5   7  11]
 [  3   4   0 143   2   3   1  94   2   2   2  15   3   3   1]
 [  5   3   9  23   3   4   2   0 298   1   0   3   0   8  10]
 [  6   4   0  28   0   4   2   5   0 303   4   0   2   0   1]
 [  1   3  22  28   0  12   8   6   0   3 230   0   2   5   2]
 [  1   6   1 120   3   4   3  15   6   2   0  87   6  18  11]
 [  3   0   7  18   0   2   2   5   3   1   2   3 304   1  11]
 [  0   1   1  10   0   0   0   0   2   0   1   1   5 569   9]
 [  4   2  49  17   0   5   7   2   7   1   1  16  27  48  63]]
(row=expected, col=predicted)
```



Predicting the Answerer for a Question

- Stochastic Gradient Descent

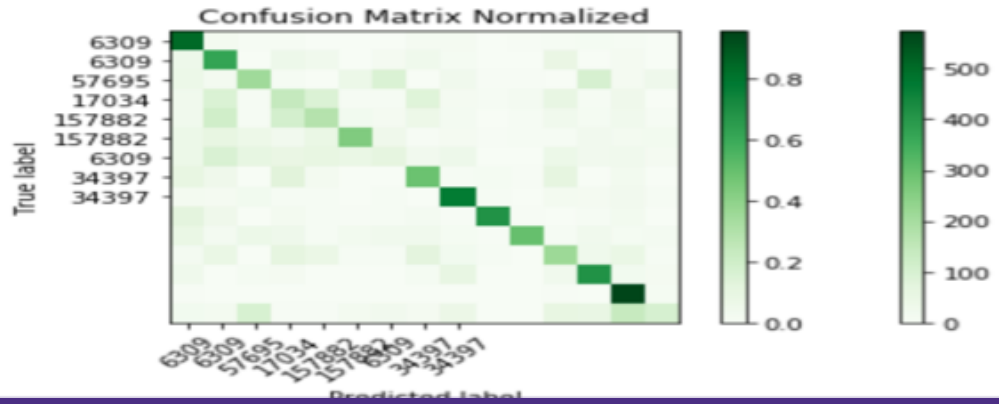
```
accuracy 0.508541392904
confusion matrix
 [[219  10   2   7   4   1   6   5   7   6   8   8   5   3   0]
 [ 10 159   6  38  12   4  12   9   2  11  10  37   2   3   2]
 [  9   9 104   9   1  23  18   0   4   6  81   8   8   3  19]
 [ 14  48  13 166  49  35  24  41  13  25  45 124   9  20   3]
 [  2  32   1  55  41  18   8  12   5   5  17  40   4   8   2]
 [  5  20  28  23  18 204  17   2   3  12  49  19   7   7   6]
 [  8  35  41  42  13  33  17   7   7   9  27  34   5   8  11]
 [ 16  10   1  50  10   5   3  92   4   7   2  70   4   3   1]
 [  5   4   2  12   5   5   6   3 288   6  13   7   3   5   5]
 [ 10   6   3  10   0   5   0   3   1 294   6  15   4   2   0]
 [  9   4  22  23   4   8   7   8   4   7 207  13   0   5   1]
 [  4  18   8  41  11   9   6  25  13   5   5 104  10  22   2]
 [ 10   3  11  13   4   5   3   9  12   4  11  16 240  10  11]
 [  1   1   2   6   5   0   2   2   6   0   7  15   5 542   5]
 [  7   4  24   4   3   9   4   3   9   3  47  27  19  54  32]]
(row=expected, col=predicted)
```



Predicting the Answerer for a Question

- Multinomial Naïve Bayes

```
accuracy 0.525811901633
confusion matrix
 [[246   8   3   4   1   1   1   6   6   1   2   3   5   4   0]
 [ 19 198   0  20  14   0   7  16   5   0   1  28   1   7   1]
 [ 21   8 107   3   0  21  48   0  13   2   2   1  52   6  18]
 [ 30 100   6 152  97   9   9  88  13   8  15  60   6  35   1]
 [ 12  50   3  46  73   9   5  16   6   0   2  11   4  11   2]
 [ 29  41  28  17  36 185  22   2   8   3   4   5  16  11  13]
 [ 20  48  30  26  29  25  34  12  18   1   1  21  12  14   6]
 [ 28  15   1  37   8   0   0 137   3   5   2  32   0   9   1]
 [ 11   7  13   4   3   0   3   1 288   3   0  11   7  15   3]
 [ 43  20   0  10   4   2   3   9   2 249   2   0   3  12   0]
 [ 27   8  22  19   4  10  15  15   6   2 162   4  14   6   8]
 [  7  24   0  31  22   3   4  34  10   3   0 103  14  24   4]
 [ 18   1   6   8   1   0   1   6  35   0   0  15 251  13   7]
 [  4   1   0   2   0   0   1   1   1   0   0   3   2 573  11]
 [  9   7  42   1   1   7  10   5  18   0   0  25  23  58  43]]
(row=expected, col=predicted)
```



Visualizations

Challenges and Limitations

- Various iteration to get 10%sample data from 50gb dataset
- No strong correlation between the features for performing regression analysis
- Improving the accuracy of the classification model
- Skewed and messy data

Thank You