Stackoverflow – Time until Answer

Team:

Gaurang Mhatre (011432200)

Parth Upadhyay (011451219)

Amruta Dhondage (011416210)

Section 1: Introduction

StackOverflow.com is an online question-and-answer site for programmers. Started in fall 2008, its rich feature set brought rapid popularity: users can ask and answer questions, collaboratively tag and edit questions, vote on the quality of answers, and post comments on individual questions and answers. The website serves as a platform for users to ask and answer questions, and, through membership and active participation, to vote questions and answers up or down and edit questions and answers.

• Motivation

We first had come up with GRE problem statement to solve the problem of getting an admit in a University based on GRE score. However, we could not find enough data to proceed in a productive manner. Changing the approach to first finding dataset helped us in looking for a more interesting problem.

While we were browsing over internet for dataset, we came across https://archive.ics.uci.edu/ml/datasets.html. UCI data repository contains dataset for various categories. We listed down many topics including Airbnb, Yelp and Restaurant-consumer dataset along with Stack overflow dataset. Finally, we came up with restaurant-consumer data set and stack overflow dataset. After some thoughts, discussions and research we found stack overflow dataset to be effective enough to implement data mining algorithms and solve a real-time problem of finding time that a user needs to wait for to get an answer for their question on the website. Hence, we decided to move ahead with the stack overflow data set and directed our efforts towards it.

• Objective

In this project, we are trying to predict the time until a user gets an answer to question asked on StackOverflow. We are using different data mining algorithms and analyzing by comparing their results. The main tasks to perform for this project are parsing of the questions tags, train the system based on tags to predict the time and test the system with remaining data. We predicted accuracy of an algorithm by comparing the predicted time to get a response against the existing time from test data. This way we can decide the accuracy, precision, recall and the other metrics for an algorithm. The overall aim of a project is to have an idea of how long it might take a question to be answered on stackoverflow.

• Market Review

As a result of being interested in Stack Overflow data , the need arose to track other Stack Overflow-based research. Latest study shows that stackoverflow dataset analysis has huge demand.

2017

- Azad, Shams, Peter C. Rigby, and Latifa Guerrouj. **Generating API Call Rules from Version History and Stack Overflow Posts.** ACM Transactions on Software Engineering and Methodology (TOSEM) 25.4 (2017): 29. [PDF]
- Abdalkareem, Rabe, Emad Shihab, and Juergen Rilling. **On Code Reuse from StackOverflow: An Exploratory Study on Android Apps.** Information and Software Technology (2017).[website]
- Abdalkareem, Rabe, Emad Shihab, and Juergen Rilling. **What Do Developers Use the Crowd For? A Study Using Stack Overflow.** IEEE Software 34.2 (2017): 53-60.[PDF]

- Campbell, Brock Angus, and Christoph Treude. **NLP2Code: Code Snippet Content Assist via Natural Language Tasks.** arXiv preprint arXiv:1701.05648 (2017).[PDF]
- Convertino, Gregorio, et al. **Toward a mixed-initiative QA system: from studying predictors in Stack Exchange to building a mixed-initiative tool.** International Journal of Human-Computer Studies 99 (2017): 1-20. [website]
- Vinayakarao, Venkatesh, et al. **Anne: Improving source code search using entity retrieval approach.** Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017. [website]

Section 2: System Design and Implementation Details

Algorithms Considered

Initially started with the goal of comparing results of Clustering with Regression. We explored algorithms such as Linear regression, Logistic Regression and SVR(Support vector Regression). We did the implementation in scikit learn in Python.

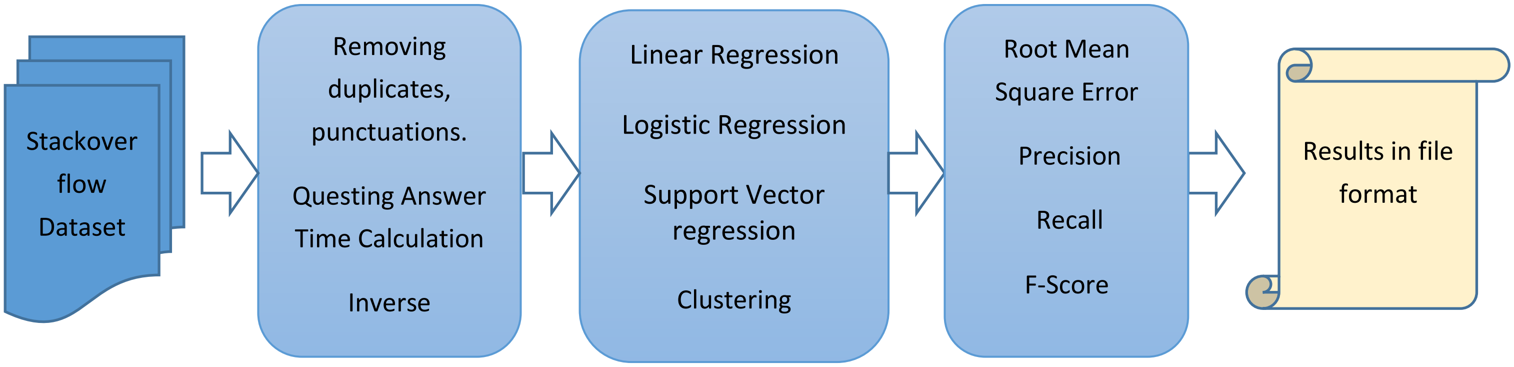
Technology and Tools Used

Familiarity with Python due to its use in our class activities and the rich library support by Python for data mining, were the two factors that contributed to our choice of Python as programming language. We used jupyter notebook to develop the code. Github has been used as a versioning and source control system by the team.

Architecture Related Decisions

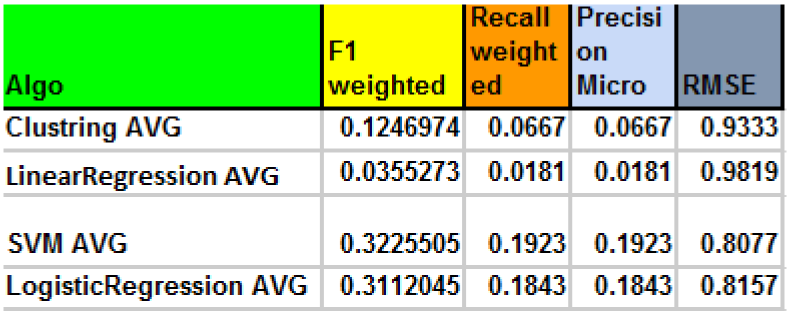
As a part of data preprocessing, we removed the duplicates and punctuations from dataset. We then formed an inverse documentation matrix from tags information. The system is then trained on this matrix using regression and clustering algorithms. Regression model was validated by RMSE , precision, recall and F-Score. Finally, we compared the prediction results of different regression algorithms with clustering.

- ## System Data Flow:



Above diagram describes the entire flow of the prediction project. The data is split in ratio of 9:1 for training and test. Both datasets are preprocessed to remove question duplicates. For removing duplicates entire data is sorted on answer times and only first entry for each unique question id(qid) is stored. The labels are generated for training data by calculating time difference of “qt”(question posting time) and “at”(answer time). The train and test data then undergo different regression algorithms and predict labels for test data.

- ## Screenshots



Algo	F1 weighted	Recall weighted	Precision Micro	RMSE
Clustring AVG	0.1246974	0.0667	0.0667	0.9333
LinearRegression AVG	0.0355273	0.0181	0.0181	0.9819
SVM AVG	0.3225505	0.1923	0.1923	0.8077
LogisticRegression AVG	0.3112045	0.1843	0.1843	0.8157

Average of all the algorithms over 10 Iterations

Section 3: Experiments

DataSet Used

Name: Stackoverflow Dataset (answes.csv)

Source: https://www.ics.uci.edu/~duboisc/stackoverflow/

The dataset consists of 3162491 records stored in csv file format. Each record is has following attributes:

qid: Unique question id
i: User id of questioner
qs: Score of the question
qt: Time of the question (in epoch time)
tags: a comma-separated list of the tags associated with the question. Examples of tags are html'',R'', mysql'',python'', and so on; often between two and six tags are used on each question.
qvc: Number of views of this question (at the time of the datadump)
qac: Number of answers for this question (at the time of the datadump)
aid: Unique answer id
j: User id of answerer
as: Score of the answer
at: Time of the answer

Data Preprocessing:

- Cleaning data :
  In data preprocessing, we are selecting tags information. Removing duplicates and punctuations from the data is one of the crucial cleaning step to get the accuracy of the algorithms. We have removed such words and characters [,.‘].

- Generating Labels for Training Data:
  The labels are generated for training data by calculating time difference of “qt”(question posting time) and “at”(answer time)

- Inverse Document Frequency Matrix:
  Once the data is cleaned, now we create inverse document matrix over which regression and clustering algorithms are implemented. In information retrieval, tf–idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

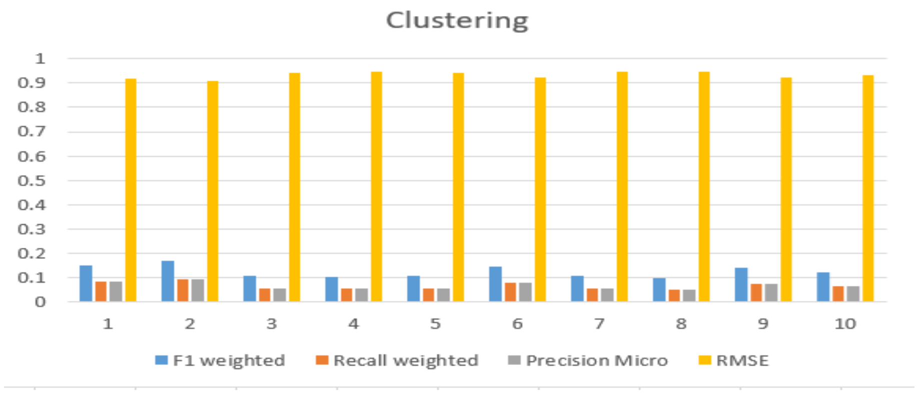
Methodology Followed

We have used Simple Linear regression, Logistic and Support Vector regression algorithms on the dataset model created.   Along with regression algorithm results, we are using clustering to compare the results. We have divided complete dataset into 90/10 train and test data. Model is generated using train data and tested on test data. Accuracy is validated using Root mean squared error, precision, recall and F1-Score.

- Linear Regression: In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. Tags variable is used as a predictor (X) and time difference is used as a prediction(y). [ref link 4]
- Logistic Regression: Logistic regression was developed by statistician David Cox in 1958.[2][3] The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features).We have implemented this algorithm on one variable called tags and predicted a time result.[ref link 5]
- Support Vector Machine Regression: Support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. [ref link 6]
- Clustering: k-means clustering is a method of vector quantization, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. We have applied this algorithm on train data. We are developing a model based on estimating nearest cluster and nearest point in the cluster to predict the time for a specific test data.[ref link 8]

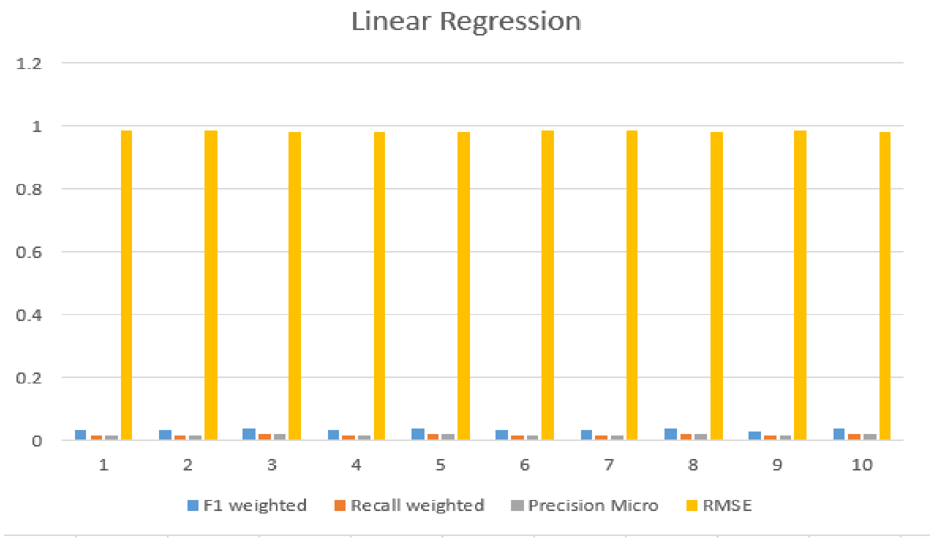
● Graphs showing comparative results

Clustering



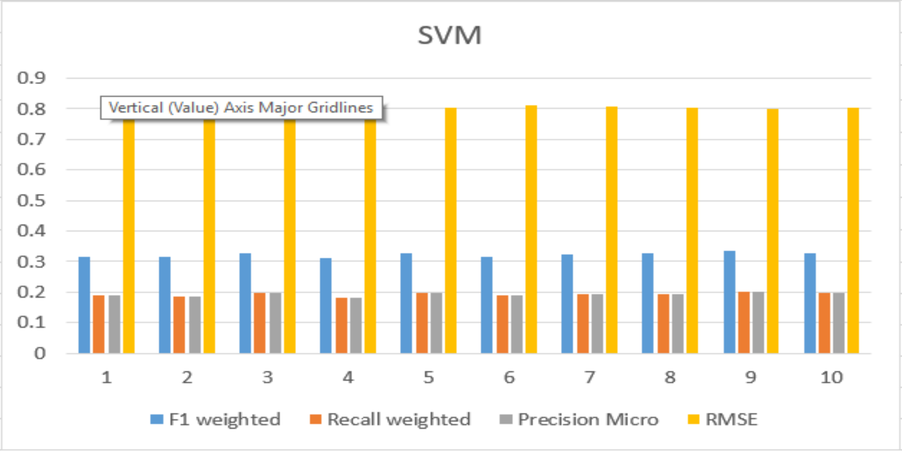
10 Iterations

Linear Regression



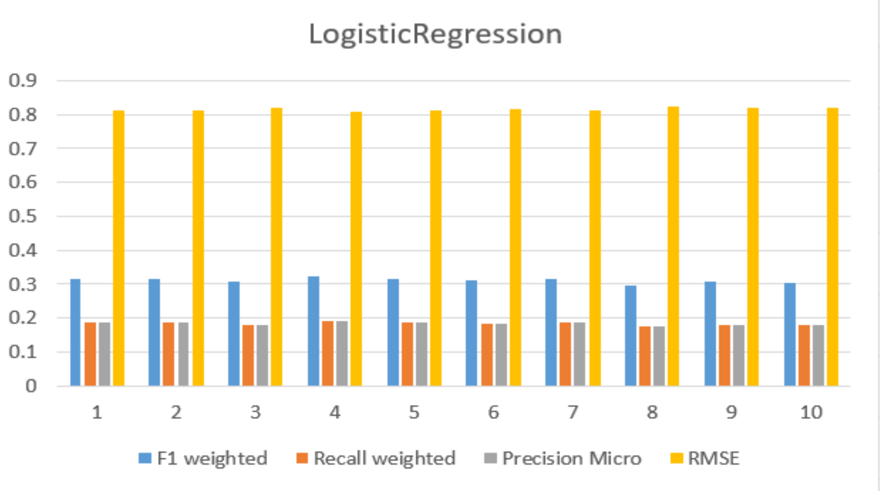
10 Iterations

SVM



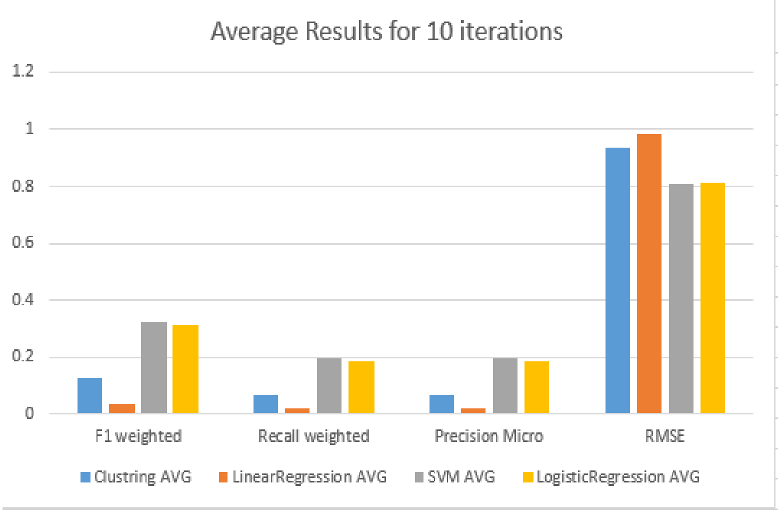
10 Iterations

Logistic Regression



10 Iterations

Average of all 10 Readings



- ## Analysis of Results

It is evident from the above chart that SVM Regression is the best algorithm that suits the dataset under study as it has all four parameters better than the other four algorithms. It is also clear that out of the four algorithms that were under study, Linear Regression is the algorithm that is the least suitable for Stackoverflow dataset, as it produces the least F-1 score and Precision.

Section 4: Discussion & Conclusions

- ## Decision Made

First decision to be made was the project topic. Approach to choose the topic was choosing a right and efficient dataset. After a couple of rounds of discussions with professor, we finally decided to go with Stackoverflow dataset

Second decision was to choose appropriate programming language. We selected Python because of its existing data mining related libraries.

Third decision was to choose algorithms for prediction. As its a continuous dataset, we concluded that problem falls in regression category. Out of available regression algorithms we tried 4- linear, ridge, SVM, logistic. Along with regression, we wanted to try out clustering approach and we successfully able to implement k-means clustering.

- ## Difficulties Faced

Finding out correct dataset for the project. Choosing algorithms for effective and accurate results.

- ## Things that worked

Python libraries for data mining. Dataset understanding as stackoverflow dataset was clean and simple.

- ## Things that didn’t work

For project topic selection, we initially followed approach of finding problem first and dataset later which did not work. We learnt to get to the problem statement from dataset. Using all attributes from the dataset for analysis.

- ## Conclusion

Working on this project was a great experience in terms of understanding actual application of data mining algorithms on real world dataset.

Project Plan

	Team	Time/Date
Project Topic/dataset Discussion	All	24 Feb – 16 March
Project analysis	All	16 March – 11 April
Project implementation	Gaurang :Clustering , Metrics Generation Parth :SVR, Logistic Amruta :Linear, Logistic	11 April - 9 May

References :

1. https://www.ics.uci.edu/~duboisc/stackoverflow/
2. https://en.wikipedia.org/wiki/Tf%E2%80%93idf
3. https://en.wikipedia.org/wiki/Stack_Overflow
4. https://en.wikipedia.org/wiki/Linear_regression
5. https://en.wikipedia.org/wiki/Support_vector_machine
6. https://en.wikipedia.org/wiki/Logistic_regression
7. https://en.wikipedia.org/wiki/Ridge
8. https://en.wikipedia.org/wiki/K-means_clustering

Libraries Used:

-numpy , scipy, sklearn , pandas, matlib