Dataset Characterization

Benchmarking Machine Learning Methods for Expert Recommendation

Group Members

Fidelia Orji, fao583, fao583@mail.usask.ca
Amirabbas Jalali, amj301, amj301@mail.usask.ca
Seyedeh Mina Mousavifar, sem311, sem311@mail.usask.ca, will hand in the document

Dataset Overview

The dataset we are using is the Stack Overflow[1] (SO) dataset. Stack Overflow has millions of active users that create content on this platform. The dataset has the following files: Posts – 14.3G, PostLinks – 84.7m, Tags – 797.9k, Users – 504.8m, Votes – 1.1G, Badges – 242.7m, and Comments – 4.2G. It consists of 19m questions, 29m answers, 73m comments, and 57k tags.
This dataset is gigantic, and it is impossible to store it on PC RAMs. The only solution to deal with this enormous dataset is using Database to store them on HDDs, but these datasets are more massive than our HDDs. Consequently, we must use cloud base Database platforms like Google Big Query[2] and Stack Exchange Data Explorer[3] (SEDE). We choose SEDE because it is free. This platform enables us to retrieve chunks of datasets using SQL queries, however, it limits data exports to 50,000 rows per query. So, we try to collect approximately 1 million questions, by first selecting users who have significant contributions in this platform, and collecting their related questions.
We performed the following operations ~~using python scripts~~.

1. We rank users by their reputation and select top users with a reputation of more than 320,000. This threshold leads to approximately 1 million questions.
2. Then we selected the answers that these users have provided for questions from the Posts table.
3. Afterward, we collected the questions for these answers from the Posts table.
4. Finally, we selected the Badges of these users from the Badges table.

The Posts file consists of an attribute named PostTypeId which is 1 if the Post is a question and 2 if the Post is an answer to a question. Answers are the activities of our selected users. The data is a mixture of answers and questions that belong to the selected users. In order to separate questions and answers, we have to find the items that have PostTypeId = 1 and concatenate them to the questions dataset. We retrieve 922,199 questions, 930,668 answers, 108 users, and 219,725 badges based on the criteria below.

[1] https://archive.org/details/stackexchange
[2] https://cloud.google.com/bigquery
[3] https://data.stackexchange.com/stackoverflow/query/new

Below we describe some of the files contained in the SO dataset by using some exploratory data analysis method to examine and quantify their distribution.

The users' data after our filtering consists of 108 users and fourteen features comprising of reputation, upvotes, downvotes, creation date, etc. Figures 1 to 4 show users' reputation distribution, users' view count distribution, users' upvotes distribution, and users downvotes distribution respectively.

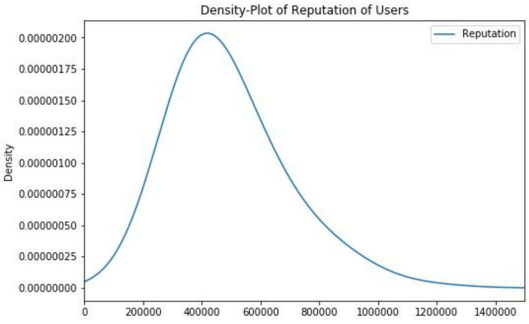


Fig. 1: Users' reputation distribution

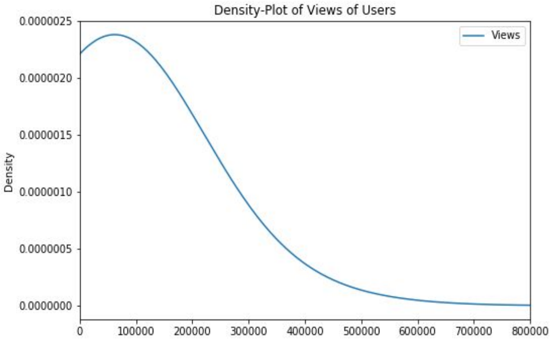


Fig. 2: Users' view count distribution

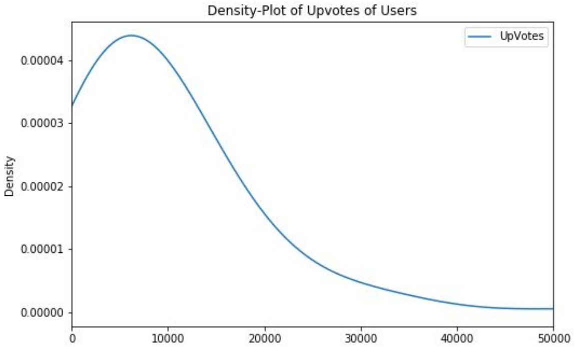


Fig. 3: Users' upvotes distribution

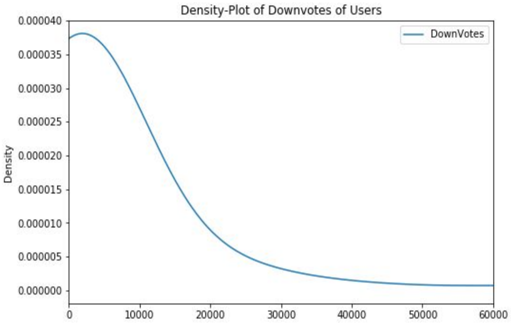
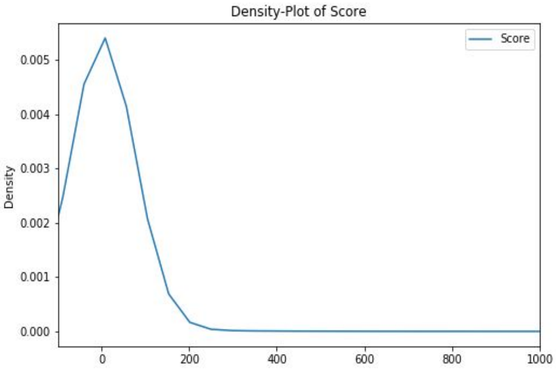


Fig. 4: Users downvotes distribution

The questions data consists of 922,199 questions with 22 features comprising ofthe score, view count, accepted answer id, title, body, tags, last activity, etc. Figures 5 to 7 show the density plot of questions per user, questions score distribution, and questions answer count distribution respectively.



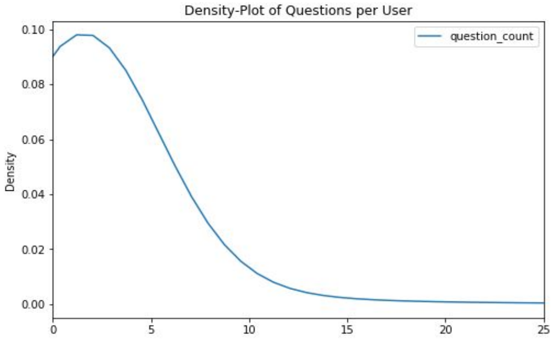


Fig. 5: Questions per user

Fig. 6: Questions score distribution

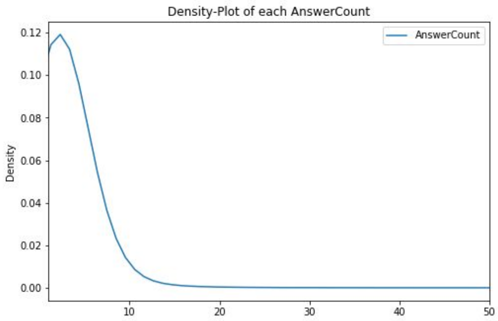
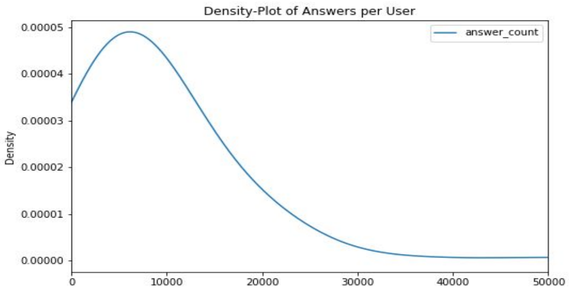


Fig. 7: Questions answer count distribution

The answer data consists of 930,668 answers with 22 features comprising of the score, view count, owner user id, title, body, tags, etc. Figures 8 to 9 show answers per user distribution and answers score distribution respectively.



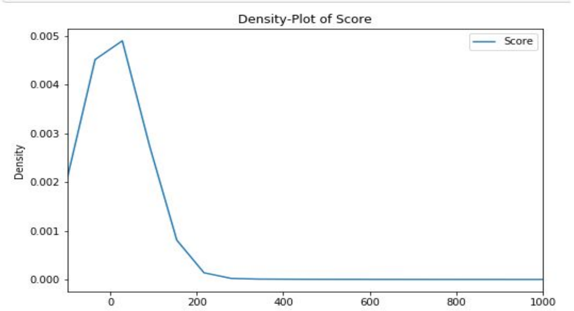
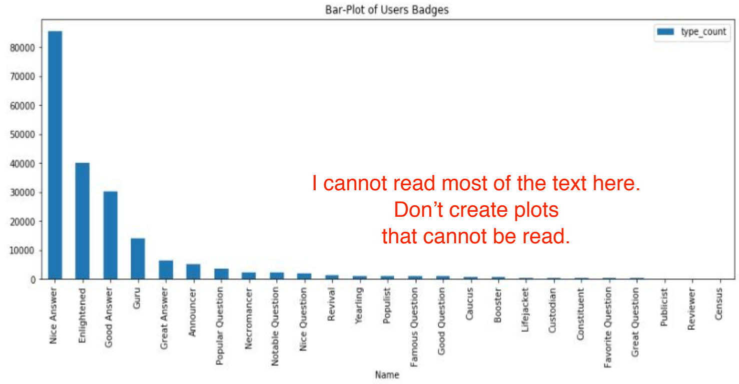


Fig. 8: Answers per user distribution

Fig. 9: Answers comment count distribution

For the badges data, there are 219,725 badges and six features. Figure 10 shows users' badges distribution.



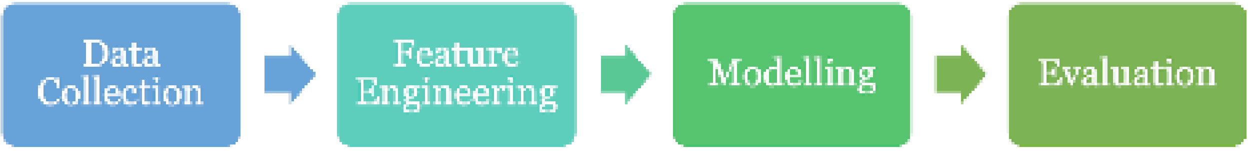
I cannot read most of the text here.
Don't create plots
that cannot be read.

Fig. 10: Users' badges distribution

As we can see in the plots above, most of our distributions follow the Power-law distribution because they are related to human behaviour. But users' reputation, questions score, and answer score seem to follow the Beta distribution.

Approach

In this project, we aim to recommend the experts who are most likely going to answer a given question. The project will have the following pipeline:



Providing timely and quality answers to questions is important in improving community questions answering (CQA) websites. An appropriate recommendation algorithm technique is needed in such websites that have millions of users with diverse skills. The purpose of our recommender system is to suggest relevant experts for a given question. Ranking experts based on experts' features, questions' features and their interaction require effectively adjusting the importance of these features and discovering proper ranking. However, due to the complexity of this problem, we will cast our recommendation problem into a classification problem, in which we will predict whether a question will be answered by a user or not. We will use previous answers provided by users to questions to predict users who will possibly give a quality answer to a new question posted on the website. In order to do this, we have to determine the attributes that define a user, a question, and a quality answer. In addition, we have to find out what users consider when choosing a question to answer. Our task will be to predict active users that will give quality answers to newly posted questions. We are going to classify the active users based on how they engage in answering questions. The tag-based badges for answers to questions will help us in this classification. We will examine users that provide a higher percentage of quality answers and build profiles for them. Quality answers would be judged based on the number of upvotes by peers. The users in each of our categories will be ranked based on the quality of answers that they provided. Moreover, for the questions, we will create different categories for the questions based on their quality and rank each category.

The ability to extract relevant features from our SO dataset is very relevant in this project as these features will determine the quality of our machine learning prediction. We will investigate and compare the performance of the following machine learning algorithms: Naive Bayes, Support Vector Machine(Cortes and Vapnik, 2020), Logistic Regression, Decision Tree(Quinlan, 1986), and Random Forest(Tin Kam Ho, 1995) for predicting the best users that will provide quality answers to a given question. Performance modelling of these algorithms for SO can help community questions answering websites in determining algorithms that will help them to improve. We will assess the impact of feature engineering, size of the training set, training and inference times, and accuracy among the models.

Schedule and Plan of Work

Activities	Date
Cleaning of Dataset and Researching for Approach	17th to 29th Feb 2020
Feature Engineering	1st to 20th March 2020
Model building	21st March to 4th April
Models Testing	5th to 11th April
Paper Writing and Presentation Preparation	11th to 26th April

Metrics of Success

Evaluating the five algorithms mentioned above will be part of our project as it will help us in determining the best performing algorithm for the problem we are trying to solve. In evaluating the performance of the algorithms, we are going to use the F1 score and confusion matrix. The F1 score will help us to measure the robustness of the algorithms. The F1 score tries to find a balance between precision and recall. Thus, a high F1 score reveals a better performance of a model. The confusion matrix will give us an idea of the type of error each model has and the things the model is getting correct. We will use ten-fold cross-validation (comprising training set and validation set) as part of our evaluation method. To measure the success of our metrics we will use an independent dataset that has not been used for training nor validation to run the models and compare the predicted users' category to already actual users that have been voted up for producing quality answers.

Prof Stavness and I are concerned about over-lap between projects in CMPT 898 and CMPT 820. For my part, I think it is fine for you to use the same dataset in this group project as in CMPT 898. I have emphasized the importance of separate learning algorithms, and you have satisfied my requirements.

You should make sure Prof Stavness is okay with it too.

You absolutely must not submit a project report for these courses that:
1. Consists of copy/paste from the project report(s) for other courses. Each report is written separately.
2. Has copy paste from websites or articles or books.

The main learning objective of this project is to get some practical experience in all the tasks of research. Copy/paste evades this learning objective. Copy/paste will be dealt with harshly, because it is a violation of trust. Broken trust cannot be repaired in Science. If your work in Science is found to be falsified in any way (including copy/paste), your research career is over. You can even damage the career of your collaborator(s), including your supervisor.

Make sure you understand that! If it’s not clear, please ask.