Understanding the Developer Community

An analysis of the 2017 Stack Overflow Survey Results

Louis Michael
Virgina Polytechnic Institute and State University
Blacksburg, VA
louism@vt.edu

Shannon R. Serrao
Virgina Polytechnic Institute and State University
Blacksburg, VA
shann87@vt.edu

ABSTRACT

This paper is the class project for CS/STAT 5525. We worked with with the Stack Overflow Developer Survey in an attempt to provide insights about the developer community at large and specifically how people use Stack Overflow (SO). Stack Overflow is a website that provides assistance to developers in the form of a question and answer forum. They also conduct an annual survey asking questions on a series of different topics. We choose to look at the Stack Overflow engagement part of this survey. We divide our project into a supervised classification problem aimed at helping Stack Overflow to identify general users dissatisfied with Stack Overflow based on several other Stack Overflow usage factors and an unsupervised clustering problem which reviled some of the different core groups of users.

CCS CONCEPTS

• **Information systems → Clustering and classification**;

KEYWORDS

Clustering, Classification, Data Mining, Data Analysis

1 PROBLEM STATEMENT

Our project centered around two objectives:

(1) Classification problem on the attributes relevant to user engagement with Stack Overflow (SO), in relation to their satisfaction with the product.
(2) Clustering problem on the dataset. With an emphasis on features related to respondents usage of Stack Overflow.

The overall goal of our analysis is provide a greater understanding of groups of Stack Overflow [7] users and if how they use individual features can be used as a predictor of their overall satisfaction with the product.

2 ALTERATIONS FROM THE PROPOSAL TO THE PROJECT

The largest single shift from the proposal to the project was in the form of the subset of features we chose to investigate. When we initially proposed the project we were hoping to investigate some of the factors that effect the developer community at large, but as we started to conduct analysis we chose to narrow the project focus onto how users of Stack Overflow interact with the product. This was initially part of the intent of our project but shifted into the lime light when we realized that we wanted to have both our classification and clustering on similar features in order to provide a better insight on the relations of this information. This also allows our project to more clearly focus our real world insights onto insights for Stack Overflow for it's users.

Some of the smaller alternations ended being in the specifics of the classification algorithms utilized, during the proposal we had suggest decision trees and SVM, but in the final project discuss results of decision table, SVM, as well as regression analysis in the form of linear regression and SVM applied to regression. These classification methods provided stronger results, and regression we felt was more application given the class label that we investigated was a satisfaction value on a scale form 0 - 10.

3 DATA DESCRIPTION

The dataset in our analysis comes from the Stack Overflow survey of 64,000 developers across 213 countries around the globe; 51,392 of the total respondents are considered analyzable due to incomplete responses in some cases. The survey, conducted from January 12^{th} to February 6^{th}, considered only those respondents usable if they completed the developer section of the survey. An extensive analysis of this data set is published by Stack Overflow [5](It contains a detailed description of methodology of the survey). In our project, we aim to provide additional analysis/insight of the survey. We propose a study of the dataset by defining a classification problem and clustering analysis problem. We highlight the survey blocks below:

(1) Developer attitudes (Q250)
(2) Job-seeking and compensation (Q410 to Q350)
(3) Education and professional development (Q520 to Q550)
(4) Job and career satisfaction (Q220, Q225)
(5) Stack Overflow (Q910 -Q980)

The subsection of these attributes that we chose to focus on the most is the questions asking about Stack Overflow engagement. These features are described in detail in the combination of Table 1 and Table 2. These features are refereed to by number, listed as the far left column of Table 1 where listing the full feature name was not practical.

4 DATA PRE-PROCESSING

- The Stack Overflow(SO) dataset has 155 attributes, of which we decided to develop insights only on the SO engagement attributes listed in 1. So the first decision in preprocessing stage was feature selection of only SO engagement attributes. The other attributes were deemed incidental to the the our problem.

S.N	Feature Name	Question Asked	Response Type
1	StackOverflowDescribes	Which of the following best describes you?	1
2	StackOverflowDevices	Which of the following devices have you used to connect to Stack Overflow over the last three months?	3
3	StackOverflowFoundAnswer	Over the last three months, approximately how often have you done each of the following on Stack Overflow? Found an answer that solved my coding problem	4
4	StackOverflowCopiedCode	Over the last three months, approximately how often have you done each of the following on Stack Overflow? Copied a code example and pasted it into my codebase	4
5	StackOverflowJobListing	Over the last three months, approximately how often have you done each of the following on Stack Overflow? Seen a job listing I was interested in	4
6	StackOverflowCompanyPage	Over the last three months, approximately how often have you done each of the following on Stack Overflow? Researched a potential employer by visiting its company page	4
7	StackOverflowJobSearch	Over the last three months, approximately how often have you done each of the following on Stack Overflow? Searched for jobs	4
8	StackOverflowNewQuestion	Over the last three months, approximately how often have you done each of the following on Stack Overflow? Asked a new question	4
9	StackOverflowAnswer	Over the last three months, approximately how often have you done each of the following on Stack Overflow? Written a new answer to someone else's question	4
10	StackOverflowMetaChat	Over the last three months, approximately how often have you done each of the following on Stack Overflow? Participated in community discussions on meta or in chat	4
11	StackOverflowAdsRelevant	The ads on Stack Overflow are relevant to me	5
12	StackOverflowAdsDistracting	The ads on Stack Overflow are distracting	5
13	StackOverflowModeration	The moderation on Stack Overflow is unfair	5
14	StackOverflowCommunity	I feel like a member of the Stack Overflow community	5
15	StackOverflowHelpful	The answers and code examples I get on Stack Overflow are helpful	5
16	StackOverflowBetter	Stack Overflow makes the Internet a better place	5
17	StackOverflowWhatDo	I don't know what I'd do without Stack Overflow	5
18	StackOverflowMakeMoney	The people who run Stack Overflow are just in it for the money	5
19	StackOverflowSatisfaction	Stack Overflow satisfaction	2

Table 1: Questions from survey regarding the usage of Stack Overflow

Response Type	Options Provided
1	1. I'd never heard of Stack Overflow before today ,2. I've heard of Stack Overflow, but have never visited ,3. I've visited Stack Overflow, but haven't logged in/created an account, 4. I have a login for Stack Overflow, but haven't created a CV or Developer Story, 5. I have created a CV or Developer Story on Stack Overflow
2	0 to 10,Anchor 0 = "Not at all satisfied", Anchor 10 = "Extremely well satisfied"
3	A. A desktop or laptop computer; B. An iPhone or iPad, using a mobile web browser; C. An iPhone or iPad, using the Stack Exchange iOS app; D. An Android smartphone or tablet, using the mobile web browser; E. An Android smartphone or tablet, using the Stack Exchange Android app; F. Some other phone, using the mobile web browser
4	1 = "Haven't done at all", 2 = "Once or twice", 3 = "Several times", 4 = "At least weekly", 5 = "At least once each day"
5	1 = "Strongly disagree", 2 = "Disagree", 3 = "Somewhat agree", 4 = "Agree", 5 = "Strongly agree"

Table 2

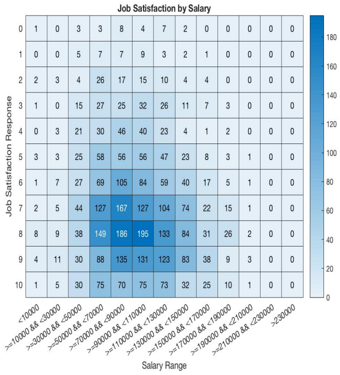
- All attributes except StackOverflowDevices have an ordinal nature to them. Hence, attributes 3-18 were transformed to a equispaced scale, using the order as described in 2. The related nature of these features made them the focus of the clustering problem.
- The attribute for classification program in the project StackOverflow satisfaction has an ordinal nature to it as well. However, we treat in the initial attempts this attribute as a nominal type and implement in order for a classification goal. A later attempt at regression of this same problem was made

by considering this attribute to be numeric in nature. For the classification section of this program, this field was transformed into three bins of equal width [0, 3.33), [3.33, 6.66), [6.67, 1]. So our Satisfaction field was discretized into one group having very low satisfaction, medium satisfaction and high satisfaction in their usage of SO. This feeds into our goal of estimating the customer satisfaction of SO based on the SO experience detailed in the survey.

- The StackOverflowDescribes attribute has two attribute values (*1. I'd never heard of Stack Overflow before today ,2. I've heard of Stack Overflow, but have never visited*), which corresponds to those survey participants who have no SO experience and hence gauging their SO satisfaction is meaningless. Thus, in the record data, these records show up as missing values and instances corresponding to them were outright eliminated.
- The nature of the survey results was such that most StackOverflow have a high satisfaction range [6.67, 1]. Thus the lower satisfaction classes are rare[10]and need to be treated. To remedy for class unbalance in the dataset, we re-sampled the data[13], to provide for equal frequency in all the three nominal bins. This prevents our classifier from being biased against the low and medium SO satisfaction users. We reason against using outright equal frequency binning of the dataset without re-sampling because it makes the bin structure [0, 7.5), [7.5, 8.5), [8.5, 1] extremely skewed and interpreting of the satisfaction of a SO user becomes intractable.

5 DATA EXPLORATION

Initially we had started to explore some of the different factors that might impact job satisfaction and created a heat map of the unnormalized Job Satisfaction data (0 is the most unsatisfied and 10 is the most satisfied) in relation to buckets of salary (given in USD) in Figure 1. For this specific figure only respondents from the United States were considered. The figure shows a overall trend that most survey respondents tend to be relatively well paid compared to the national mean in 2016 of $46,640.94 [4]. This may suggest further interesting insights to be gained from this dataset but as we shifted our interest to the Stack Overflow specific features one of the first steps we took was to search for correlation between any of the features to see if any interesting trends arose. For this only Features 3-19 where considered as they where preprocessed into meaningful ordinal values that made them appropriate to preform correlation against. We examined the correlation by plotting a correlation matrix using the preprocessed data and the MATLAB economics toolbox function `corrplot` cor [6]. The results can be seen in Figure 2. Many of the features are loosely correlated in a relevantly predictable sense, if you liked one thing about SO you were generally more likely then not to say good things about the product in regards to another feature and less likely to believe other aspects like ads were annoying. The only more strongly correlated features were features 5, 6, and 7 corresponding to StackOverflowJobListing, StackOverflowCompanyPage, and StackOverflowJobSearch respectively which can be seen in the repeated cutout of Figure 2. As a result these features peaked our interests and were clustered separately. This correlation suggests that there is a group of users distinct from general users that leverage more of the business features of SO.



	<10000	>=10000 && <30000	>=30000 && <50000	>=50000 && <70000	>=70000 && <90000	>=90000 && <110000	>=110000 && <130000	>=130000 && <150000	>=150000 && <170000	>=170000 && <190000	>=190000 && <210000	>=210000 && <230000	>230000
0	1	0	3	3	8	4	7	2	0	0	0	0	0
1	0	0	5	7	7	9	3	2	1	0	0	0	0
2	2	3	4	26	17	15	10	4	4	0	0	0	0
3	1	0	15	27	25	32	26	11	7	3	0	0	0
4	0	3	21	30	46	40	23	4	1	2	0	0	0
5	3	3	25	58	56	56	47	23	8	3	1	0	0
6	1	7	27	69	105	84	59	40	17	5	1	0	0
7	2	5	44	127	167	127	104	74	22	15	1	0	0
8	8	9	38	149	186	195	133	84	31	26	2	0	0
9	4	11	30	88	135	131	123	83	38	9	3	0	0
10	1	5	30	75	70	75	73	32	25	10	1	0	0

Figure 1: A heat map of salary sorted into buckets compared to respondents job satisfaction rating

6 CLASSIFICATION/REGRESSION TASK

6.1 Introduction

We had to choose how to estimate the SO Satisfaction attribute based on the user information or the other measurable attributes available to SO. Thus SO satisfaction could be discretized into bins and made nominal; this corresponds to the classification task. The second option is to retain its original ordinal character use regression to estimate the satisfaction. In our project we first attempt Classification on our discretized SO Satisfaction class; we then follow it by attempting regression on the original attribute after binarizing the nominal attribute and normalizing the rest of the numeric attribute datasets.

In the classification problem, we have three bins or classes which have been resampled to give us equal frequency over all classes to redress class imbalance. To classify three classes, we have to use either *one against all* treatment or *one against one* on our attempts[9].

6.2 Classification/Regression Models

We use the following classification and regression methods.

- Decision Table: We use a simple Decision Table which uses area under ROC curve[8] as the evaluation measure for classification on the training set. This best first algorithm searches the space of attribute subsets by greedy hill climbing augmented using backtracking facility[12]. This model uses one against all.
- Decision tree: This uses cost based classification which implements decision trees using J48 algorithm[15]. We decide to use one against one because it is important to penalize misclassifying the lower SO satisfaction classes as opposed to the topmost SO satisfaction class. This can be appropriately done only by using a one on one classification scheme.
- SVM using C-SVC method and kernel is exponential function [1]: From our trials on using SVM, SVM does not perform well on the multi-class classifier when we implement the one against all method. Hence we choose the one against

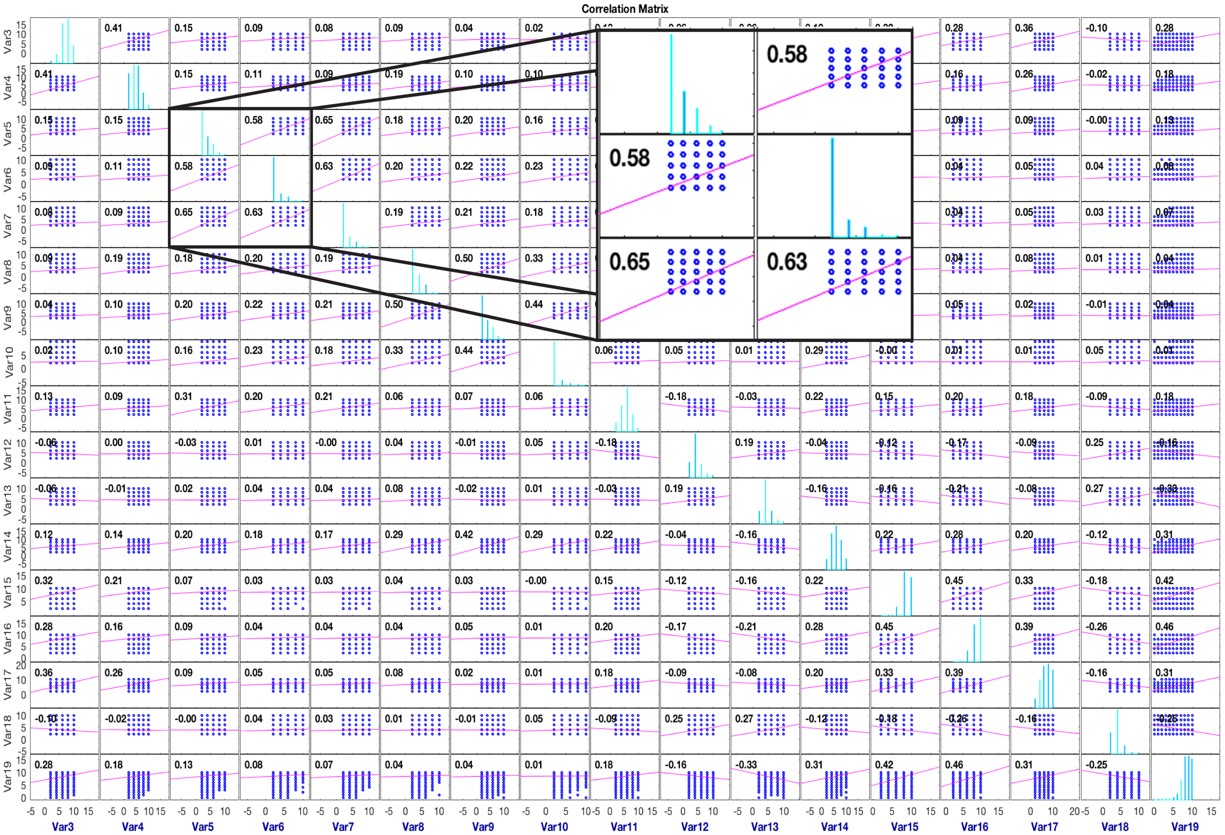


Figure 2: A correlation matrix of 17 of the Stack Overflow specific features

all method, again using an explicit cost matrix to penalize misclassying the rarer class. The kernel is function is as $exp(-\gamma|u-v|^2)$, where γ is a tunable parameter which controls the influence of a training point at a distance from the margin.

- Linear Regression: Linear regression is applied on the dataset after using M5 method[16] for attribute preselection.
- nu-SVM Regression: We use a nu-type SVM regression with the same kernel as before, with the cost parameter kept as low as possible with a view to prioritizing the training points closer to the margin. The nu-SVM does much better than the epsilon type SVM, which is not reported.

6.3 Evaluating the Classifier

In all of our classification/regression tasks, we perform 10-fold Cross Validation to ensure good predictive power of our model and independence of the test data and training set. For classification, we recall that we have a class imbalance problem as well as we have a multiple classes. So we pick Area under ROC curve which is a good evaluation measure for imbalance in the class population. Also, the confusion matrix is a allows us to interpret multiple classes and look explicitly at the breakdown of misclassified points versus class. We then impose a cost-matrix, where we penalize the FP's to FN's and compute the cost (see 6). The justification for this imposing such a cost penalty is that for Stack Overflow, a dissatisfied user is more critical than a satisfied user. We see that the Decision Table and the J48 Decision tree give us very good results(See table 4,4,5) Approximately 91 and 98 percent of the classes are classified correctly and in all the three algorithms do a very good job of not misclassifying the low satisfaction class with the high satisfaction class(See table-7). We report the total cost of the three models, we see that cost per data point of the model is very low. Among the three models, the J48 Decision tree does the best. For the regression task, compute the correlation coefficient and RMS error (See table-8). The adjusted R2 statistic for the linear regression is 0.318. To this, we make two points: although adjusted R2 is low, we note that this is a high dimensional data set and keeping the variability of the model low is not trivial. Also, we remark predictability of the model for low values of Satisfaction is more important that overall predictability. These features were not optimized in this project,

Predicted −>	A	B	C
A	10895	0	0
B	317	10159	419
C	1272	876	8747

Table 3: Confusion matrix Decision Table

Predicted −>	A	B	C
A	10895	0	0
B	64	10569	262
C	217	1503	9175

Table 4: Confusion matrix Decision Tree J48

Predicted −>	A	B	C
A	10864	31	0
B	1168	9271	456
C	166	4168	6561

Table 5: Confusion matrix SVM method

Predicted −>	class A	class B	class C
A	0	3	6
B	1	0	3
C	1	1	0

Table 6: Cost matrix

since regression was not the focus of course. Moreover, for a rough insight into SO customer satisfaction, the classification problem gives us good results.

Model	Decision Table	Decision Tree J48	SVM using C-SVC
AUC ROC	0.984	0.966	0.901
Cost	3722	2570	6963

Table 7: Area under ROC curve

Model	Linear Regression	nu-SVR
Correlation Coefficient	0.5603	0.54
RMS error	1.1329	1.1504

Table 8: Model Evaluation: Linear Regression and nu-SVR

7 CLUSTERING TASK

7.1 Model Building

For the clustering portion of the model building we employed K Means Clustering in order to group survey responses to see if any interesting patterns arise. This provided the potential for the data set to show some insights that we had yet to fully grasp. A series of feature sets were explored in this manner. The strongest coordinations that were identified from Figure 2 where in the features StackOverflowJobListing, StackOverflowCompanyPage, StackOverflowJobSearch. These features were then investigated at a series of different K - Values and a final versions of the clustering were chosen based on a combination of high relative silhouette values and a point after the "elbow point" of the mean of distances of points from their representative cluster centroid squared. These metrics were evaluated on a variety of different Ks the results of this analysis are shown in Figure 7. This resulted in two interesting clusters displayed in Figure 3 and Figure 4. Figure 3 shows clustering for K = 2 which had a relatively high silhouette value and Figure 4 shows a plot when K = 19 which had both a relatively high silhouette value and was also after the "elbow point"[14].

We also conducted similar evaluations in a set of attributes we believed would be interesting from a domain expert perspective, StackOverflowSatisfaction, StackOverflowFoundAnswer and StackOverflowCopiedCode. The idea being that satisfied users where those that were finding questions they had solved such that they could simply copy a solution to the question. This is an extremely common use case of SO. The clustering for these features was most interesting at K = 24 from a similar evaluation as stated above, displayed in Figure 6. This clustering is plotted in Figure 5. Lastly we also conducted this process on all 17 of the features used to create the correlation matrix, features 3-19, which had a interesting clustering at K=15 as determined from Figure 8. The relevant centroids are listed in Table 9. The insights gained from this clustering is discussed in the section titled Real World Insights.

7.2 Model Evaluation

To evaluate the performance of the K-means clustering that is preformed we used silhouette measure [11] in addition to the mean of squared distance which follows the same principles of Sum of Squared Errors (SSE) [2] of an individual run. Silhouette measure evaluates cluster elements both on how similar the elements are to the cluster that they are placed in and how dissimilar they are to clusters they are not a part of. This gives an idea of how well

Cluster	19	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	8.4510	7.4054	6.4628	6.5777	6.2838	6.2905	6.1014	6.7297	5.9223	6.5236	5.6959	5.7939	7.6892	8.1959	8.0811	6.5203	5.2297
2	7.5730	5.9551	3.6958	2.4840	2.1867	2.3129	2.6188	2.7433	2.2282	3.8894	7.6768	4.7640	4.8764	7.6439	7.3189	4.8176	4.7139
3	5.7840	6.2534	3.9107	2.8764	2.3718	2.5794	3.2170	3.2316	2.3406	5.0322	4.7103	8.3904	3.6926	7.3229	6.7580	5.1547	4.9865
4	8.6985	7.8347	5.6113	2.7360	2.2540	2.3642	3.1354	3.0130	2.2877	4.4591	7.3619	4.6764	6.2999	8.9487	8.9518	8.4637	4.0765
5	9.2597	7.3274	3.8347	3.0780	2.3298	2.4680	2.8211	3.1135	2.2915	6.6726	3.5367	3.5908	7.9000	9.2882	9.6319	8.8192	3.0126
6	9.0206	8.1680	6.4771	3.0447	2.3937	2.5276	5.1022	5.9083	3.4148	5.7321	3.8778	3.9788	8.0435	9.2444	9.3866	9.0223	3.3302
7	8.5392	7.0796	4.0914	6.0392	4.6271	5.5998	2.8254	3.0380	2.2743	6.3848	4.1663	4.4074	5.5546	8.4133	8.5166	6.5321	3.8337
8	8.3176	6.1777	3.3122	3.1265	2.5532	2.7456	3.7066	6.8223	3.8075	5.1575	4.0619	3.8371	7.7362	8.2638	8.4374	5.7941	3.7416
9	9.2752	8.3578	6.2141	6.5367	5.0795	5.6743	3.7798	4.6743	2.7171	7.2508	3.6881	3.9021	7.9557	9.3777	9.5963	8.8670	3.2508
10	7.0542	5.0542	2.8509	2.3685	2.1552	2.2691	2.3294	2.4668	2.1284	4.4601	4.0402	4.2200	3.8571	7.1133	6.8465	4.0514	4.1764
11	8.4228	7.5307	6.0050	2.4266	2.1414	2.1698	2.4408	2.3739	2.0987	4.7294	3.9699	4.1556	4.3488	8.6073	8.3237	6.0552	3.9557
12	9.0390	8.7458	7.3026	3.2480	2.3100	2.4756	2.5927	2.5162	2.1367	6.1749	3.7462	4.1226	5.4401	9.3057	9.4260	9.1355	3.4479
13	8.6564	6.1855	3.0828	2.6730	2.1958	2.2878	2.4376	2.5216	2.1332	5.8851	3.8779	3.9984	5.7860	8.5509	8.7594	4.9766	3.6060
14	8.6197	7.4861	6.1017	3.6393	2.5614	2.7702	4.3688	4.5573	2.6102	6.0461	4.2522	4.3010	7.0780	8.7308	8.5803	4.6603	3.8129
15	8.6000	7.3812	3.2114	2.5273	2.1510	2.2563	2.4506	2.3984	2.0751	5.2147	3.8424	4.3020	4.3886	8.7706	8.9502	8.6702	3.6531

Table 9: The centroids created from clustering features 3-19 into 15 clusters, the feature that the value corresponds to is listed as columns and the cluster numbers are listed as rows

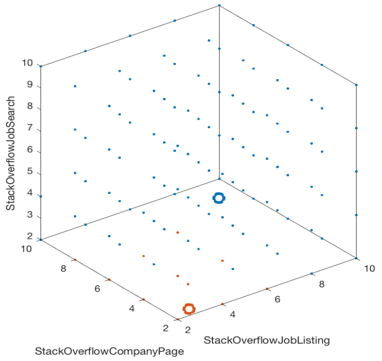


Figure 3: A three dimensional visualization of the clustering of StackOverflowJobListing, StackOverflowCompanyPage and StackOverflowJobSearch into 2 clusters using K-Means

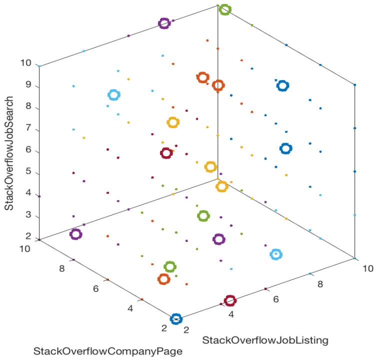


Figure 4: A three dimensional visualization of the clustering of StackOverflowJobListing, StackOverflowCompanyPage and StackOverflowJobSearch into 19 clusters using K-Means

defined a cluster is. The value of the Silhouette measure varies between 1 and -1 where 1 indicates well clustered and -1 indicates poorly clustered. We used this metric to evaluate the quality of the clustering that we preform. We also used these metrics to select K for our K-means as discussed in the Model Building Section.

Silhouette is calculated as

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Where $s(i)$ is the silouette of a data point i, $a(i)$ is the average distance from i to all other data points in the assigned cluster of i, and $b(i)$ is the average distance of i to all points in i's neighboring cluster. A points neighboring cluster is the cluster such that the average distance between a point and all points in another cluster is minimized.

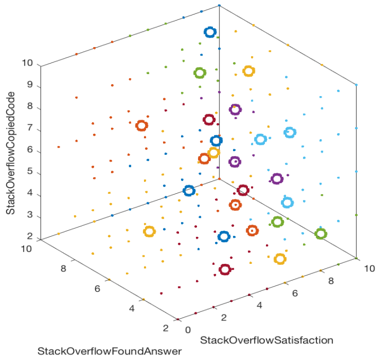


Figure 5: A three dimensional visualization of the clustering of StackOverflowSatisfaction, StackOverflowFoundAnswer and StackOverflowCopiedCode into 24 clusters using K-Means

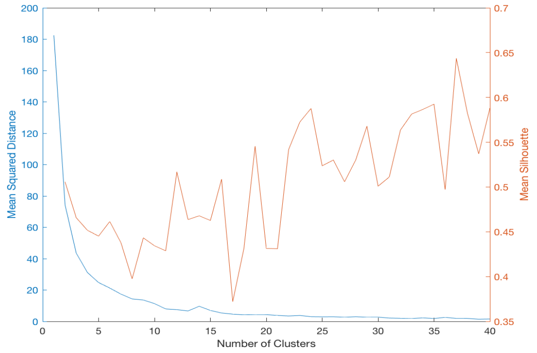


Figure 6: Mean Squared Distance and mean Silhouette across 1 - 40 clusters of the features StackOverflowSatisfaction, StackOverflowFoundAnswer and StackOverflowCopiedCode

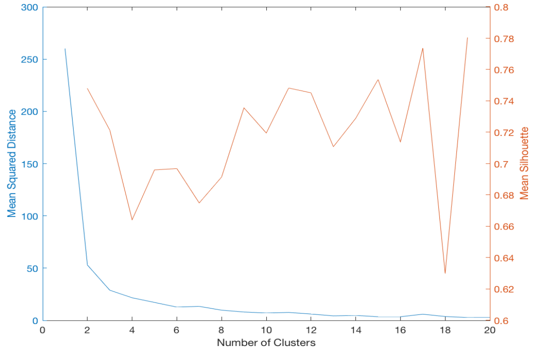


Figure 7: Mean Squared Distance and mean Silhouette across 1 - 20 clusters of the features StackOverflowJobListing, StackOverflowCompanyPage and StackOverflowJobSearch

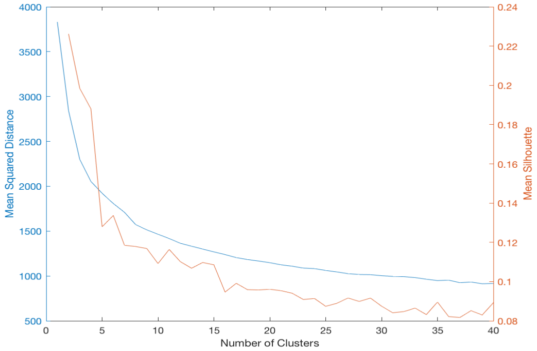


Figure 8: Mean Squared Distance and mean Silhouette across 1 - 40 clusters of the 17 features used in clustering

The 17 feature clustering overall was relatively weak, having silhouette values under 0.3 while the 3 feature clustering of StackOverflowJobListing, StackOverflowCompanyPage and StackOverflowJobSearch was fairly strong having silhouette values above 0.7. The three feature set of StackOverflowSatisfaction, StackOverflowFoundAnswer and StackOverflowCopiedCode had overall fairly strong results with a silhouettes around 0.7.

8 REAL-WORLD INSIGHTS

Stack Overflow has user data about most of the other eighteen non-class attributes(exception being Stack OverflowMakeMoney) for general users not necessarily part of the survey. The survey however, captures a good sample of all users of Stack Overflow. This survey can be used by Stack Overflow to build on the data of a general user on the Stack Overflow website to predict his/her general satisfaction. This could aid in making business decisions on how to improve the product and what factors do users generally like/dislike about Stack Overflow. Thus a classification built on the SO satisfaction could aid a predicting the satisfaction of a general user. This class is highly unbalanced towards satisfied participants which suggests that most users are satisfied with Stack Overflow. However, by emphasizing the prediction of dissatisfied SO survey participants, we can improve the prediction of the classifier on live data. Classification helps Stack Overflow build a One of interesting insights to be gained form clustering is in looking at the clusters formed when evaluating with all 17 features, most of the relatively hight overall satisfaction clusters with Feature 19 > 9, had some interesting overlap in that they had higher values for features 3 and 4 and lower values for features 5, 6, and 10. Feature 3 and 4 are, frequency of finding answers and copying code respectively. Features 5, 6 and 10 are frequency of job listing, company page and meta chat. This grouping suggests that some of the generally most satisfied users also tended to not be using SO for job hunting and were not engaging in much of the community, instead they seemed to be more objective focused only visiting to find solutions to coding questions they had at hand. This suggests that these may not be high priority features in regards to future development and resources should be utilized in providing more accurate answers and more easily allowing for the incision of code.

The shaping of the clusters in Figure 4 suggests that there groups overall tend to lineally increase their involvement across many of the job and company features and broadly divide into high engagement users of these features causal users and different types of users that have very low engagement with these features using none only one of these features. This suggests that Stack Overflow may benefit from pushing casual users into exploring these features. An alternative action that they could take from this information is to narrow their focus onto core hardcore users and looking to fulfill more advanced feature requests from this group in specific.

ACKNOWLEDGMENTS

The authors would like to thank Dr. B. Aditya Prakash for his time and insights provided through CS 5525.

The authors would also like to thank Stack Overflow for conducting this survey annually and providing a place for developers to help each other solve problems big and small. Their services are frequently invaluable even when doing the work for this paper.

The authors would like to extend our thanks to Pang-Ning Tan, Michael Steinbach and Vipin Kumar, authors of "Introduction to Data Mining" [2] and Trevor Hastie, Robert Tibshirani and Jerome Friedman, authors of "Elements of Statistical Learning" [3]. These texts helped expand our understanding of almost every topic discussed in this paper and was an invaluable resource.

REFERENCES

[1] 2000. *An introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press.

[2] 2006. *Introduction to Data Mining*. Pearson/Addison Wesley.

[3] 2009. *The Elements of Statistical Learning*. Springer.

[4] 2016. Measures Of Central Tendency For Wage Data. (2016). https://www.ssa.gov/oact/cola/central.html

[5] 2017. Developer Survey Results 2017. (2017). https://insights.stackoverflow.com/survey/2017/#methodology

[6] 2018. corrplot. (May 2018). Retrieved May 2, 2018 from https://www.mathworks.com/help/econ/corrplot.html#btbc5t7_seealso

[7] 2018. Stack Overflow. (March 2018). Retrieved March 27, 2018 from https://stackoverflow.com/

[8] P. Flach C. Ferri and J. Hernandez-Orallo. 2002. Learning Decision Trees Using the Area Under ROC Curve In Proc. of *19th Intl. Conf. on Machine Learning*. (2002).

[9] T Hastie and R. Tibshirani. 1998. *Annals of Statistics*. (1998).

[10] M.V Joshi. 2002. On Evaluating the performance of Classifiers for Rare Classes. *Proc. of 2002 IEEE Intl. Conf. on Data Mining, Japan* (2002).

[11] Peter J.Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 1 (Nov. 1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[12] Ron Kohavi. 1995. The Power of Decision Tables. In *8th European Conference on Machine Learning*. Springer, 174–189.

[13] M. Kubat and S. Matwin. 1997. Addressing the Curse of Imbalanced Training sets: One sided selection. In Proc. of *14th Intl. Conf. on Machine Learning*. (1997).

[14] District Data Labs. 2016. Elbow Method. (May 2016). Retrieved May 2, 2018 from http://www.scikit-yb.org/en/latest/api/cluster/elbow.html

[15] Marian Cristian Mihăescu, Paul Ştefan Popescu, and Dumitru Dan Burdescu. 2015. J48 List Ranker Based on Advanced Classifier Decision Tree Induction. *Int. J. Comput. Intell. Stud.* 4, 3/4 (Nov. 2015), 313–324. https://doi.org/10.1504/IJCISTUDIES.2015.072879

[16] Y. Wang and I. H. Witten. 1997. Induction of model trees for predicting continuous classes. In *Poster papers of the 9th European Conference on Machine Learning*. Springer.