COMP 488 - Machine Learning
Applied Machine Learning: Quality Control Workflow

Jose Luis Rodriguez

Loyola University Chicago

December 9, 2017

1 Abstract

In any business application where data is collected, companies must implement thorough and well-defined quality control work flows. All parties involved should strive for the highest quality and highest performing results in the data collection process. In the process of data collection, it is necessary to establish a quality control protocol (including benchmarks for process and metrics for given features as well as testing) to diminish as much as possible the chance for human error. In this report, a dataset was used containing hand curated survey respondent data including metrics for respondent engagement, network performance, and responses to open-ended questions, among others. The feature that this report aims to automate is the qualification rate feature (whether or not a survey respondent?s data is kept in the report). There were two primary applications where machine learning was proven to be superior to previous manual methods. The data contained some respondent feedback pertaining to their experience/issues that needed to be scored and classified as well, which required the use of a text classifier. The results of the text classifier were merged with a subset of the data, and three different classification algorithms were utilized (KNN, Linear SVC, Random Forest). The preliminary results of this project indicate that it is possible to automate much of the quality control work flow.

2 Introduction

Advances in technology and computational power have had repercussions throughout economies around the world. An industry that has been deeply impacted by this evolution is market research. With the advent of big data, advanced research techniques and data analytics tools have been crucial to gaining insight into the ways in which consumers behave in online as well as real-world store environments (Ingelbrecht, 2016). This new reality requires companies to take action in real time. The company that provided the data for this report creates virtual stores where participants can experience an online store (just like what an average grocery store may look like) and experience scenarios (such as shelf rearrangement and product displays/locations). By creating these virtual stores, the company is able to assess the efficacy of different choices a real store may make (such as pertaining to aisle layout, for example) without the need to spend the time and money testing said choices in real life.

As it stands, the company spends a significant amount of time and manpower curating and assuring the quality of their work (including survey reliability and consumer experience). Quality controls are performed every 1000 completed surveys. Based upon the quality of the data (as determined by the metrics outlined in the next section), the a respondent?s data is either kept or disqualified. The disqualification (DQ) process takes 4-5 hours per 1000 respondents and on average takes 16-20 hours per study. Time is devoted to manually pulling/joining data and sifting through open ended responses for signs of difficulties. The aim of this project is to automate the quality control process using machine learning techniques.

3 Data Description

The data used in this project consists of 44,235 responses to online surveys and ten features. Generally during the data collection process, a market research company will over sample to compensate for the disqualification of participants for a variety of reasons. The following features are considered qualifiers and disqualifiers for survey respondents. According to the company?s quality control documentation, top 3% time and speeding refers to respondents who complete the survey faster than is reasonable. Navigation rating and frame score refer to respondents identified as having had a poor experience in the virtual store. Issue key word and issues refer to respondents? feedback to their experience. Straight-line identifies unengaged respondents such as those who select the same response for all questions. Bad open ends is another measure to identify unengaged respondents. No pick ups identifies respondents who fail to pick up any product. Outlier sales refer to respondents who do not engage seriously in the virtual store including engaging in unrealistic purchasing behaviors. Lastly, not in database refers to data that is missing in the database such as incomplete responses. These eleven measures together create another measure called total points. Any respondent with a score greater than four is disqualified (although this measure varies depending on the number of open ended responses in the survey).

For the preliminary data exploration, a correlation analysis was performed highlighting seven features that reflect signs of poor user experience. Top 3% time appears to be correlated with speeding, straight-line, and bad open ends. Additionally, straight-line is correlated with outlier sales. These characteristics tend to illustrate respondents whom did not engage seriously with the survey. Finally, navigation rating is positively correlated with open-ended issues, no pick ups, and frame score which reflect poor user experience.

The errors are at the discretion of the individual quality control analyst, and the process is very prone to error. Moreover, as different analysts apply different criteria to how they score metrics, there are additional inconsistencies in the quality control work flow.

4 Methodology

To minimize the aforementioned flaws, one year of data was obtained from thirteen different studies conducted by an individual analyst. Overall, there are a total of 44,235 unique respondents in the dataset. The features contained in text were compiled into a single column and the text classification algorithm (linear SVC) was applied generating a feature that contains four different categories or levels of satisfaction. After classifying the text, three different classification algorithms were applied to 70% of the data as a training set. Then, the models were tweaked on 15% of the data on a development set. Finally, the final model was tested on 15% of the data.

The first approach used in this report is K-nearest neighbors (KNN), which provides a simple method of classifying any given observation. Given its simplicity, KNN was ideal as it identifies the K points closest to an observation while estimating the conditional probability for a certain class as the fraction of K points whose explanatory values equal that class. The classification of an observation is affected by the size of K. KNN has a low bias and very flexible variance. The second approach used in this report are random forests. Random forests use decision trees to classify each observation, and as such they provide a mechanism to improve accuracy, as multiple trees are produced and combined to generate a single prediction. By considering uncorrelated trees, a limited number of predictors are considered at each split. Finally, random forests average uncorrelated trees, reducing the variation among the trees and yielding high reliability (James, 2017). A third approach used in this report is a linear support vector classifier (linear SVC). This type of classifier uses a margin which separates observations into different classes. The support vector classifier?s decision rule is based on a subset of the training data unlike other methods. Only the observations that lie on the margin or on the wrong side of the margin affect the support vector classifier and are known as support vectors (James, 2017).

The metrics used to measure performance of the models were the f1 score, the recall score, the precision score, and the ROC AUC score. For this report, the measure that was favored was the f1 score as it provides the harmonic mean of the precision and recall (Geron). In many cases, market research companies are seeking to maximize recall as it will detect bad responses while maintaining as much of the sample as possible. In cases of quality control where most likely a greater percentage of the data is deemed good, utilizing the accuracy metric will be inappropriate as there will be an imbalance between disqualified and qualified respondents.

5 Results

In table 1 the models? performance is compared. The random forest classifier performed the strongest with an f1-score of 0.88522. KNN resulted in a score of 0.88525. Lastly, the linear SVC scored 0.85794.

Given the results of the different models, it is possible to apply each model to different business use cases. If the goal is to maintain the majority of the sample, following the precision score of the models (linear SVC and random forest) is recommended. If, on the other hand, the objective is to assess the ability of the quality control method, it is recommended to use a classifier that maximizes the recall (KNN model). Confusion matrices for each of the models are available in the appendix.

6 Conclusion

When assessing quality control work flows, it is imperative to consider different aspects of the data collection process as well as the relevant benchmarks and metrics. In a nutshell, nonlinear techniques outperform linear techniques making a strong argument for standardization of data collection work flows. By having different models that result in close f1 scores, it is possible to use each model in different use cases (such as trying to keep or eliminate more respondents). Multiple data manipulations and assessments by different parties will exponentially increase the data inconsistencies; therefore, there is a strong argument to create standardized work flows and data pipelines throughout the data collection process in any field.

7 Reference

James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R. 7th ed., Springer, 2017.

1.4. Support Vector Machines. Retrieved December 08, 2017, from http://scikit-learn.org/stable/modules/svm.html#support-vector-machines

1.6. Nearest Neighbors. Retrieved December 08, 2017, from http://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification

1.11. Ensemble methods. Retrieved December 07, 2017, from http://scikit-learn.org/stable/modules/ensemble.html#random-forests

3.3. Model evaluation: quantifying the quality of predictions. Retrieved December 08, 2017, from http://scikit-learn.org/stable/modules/model_evaluation.html#precision-recall-a

3.3. Model evaluation: quantifying the quality of predictions. Retrieved December 08, 2017, from http://scikit-learn.org/stable/modules/model_evaluation.html#confusion-matrix

Ingelbrecht, Nick. (2016, July 08). Hype Cycle for Market Research, 2016 (ID: G00290893). Retrieved from Gartner database.

Geron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow. O'Reilly.

8 Appendix

Table 1: Overall Model Performance Results

	F1-Score	**Recall Score**	**Precision Score**
KNeighbors	0.88947	0.76782	0.83123
Random Forest	0.88323	0.71475	0.87246
LinearSVC	0.85351	0.64013	0.85022

Figure 1: Issues Dataset - Metrics

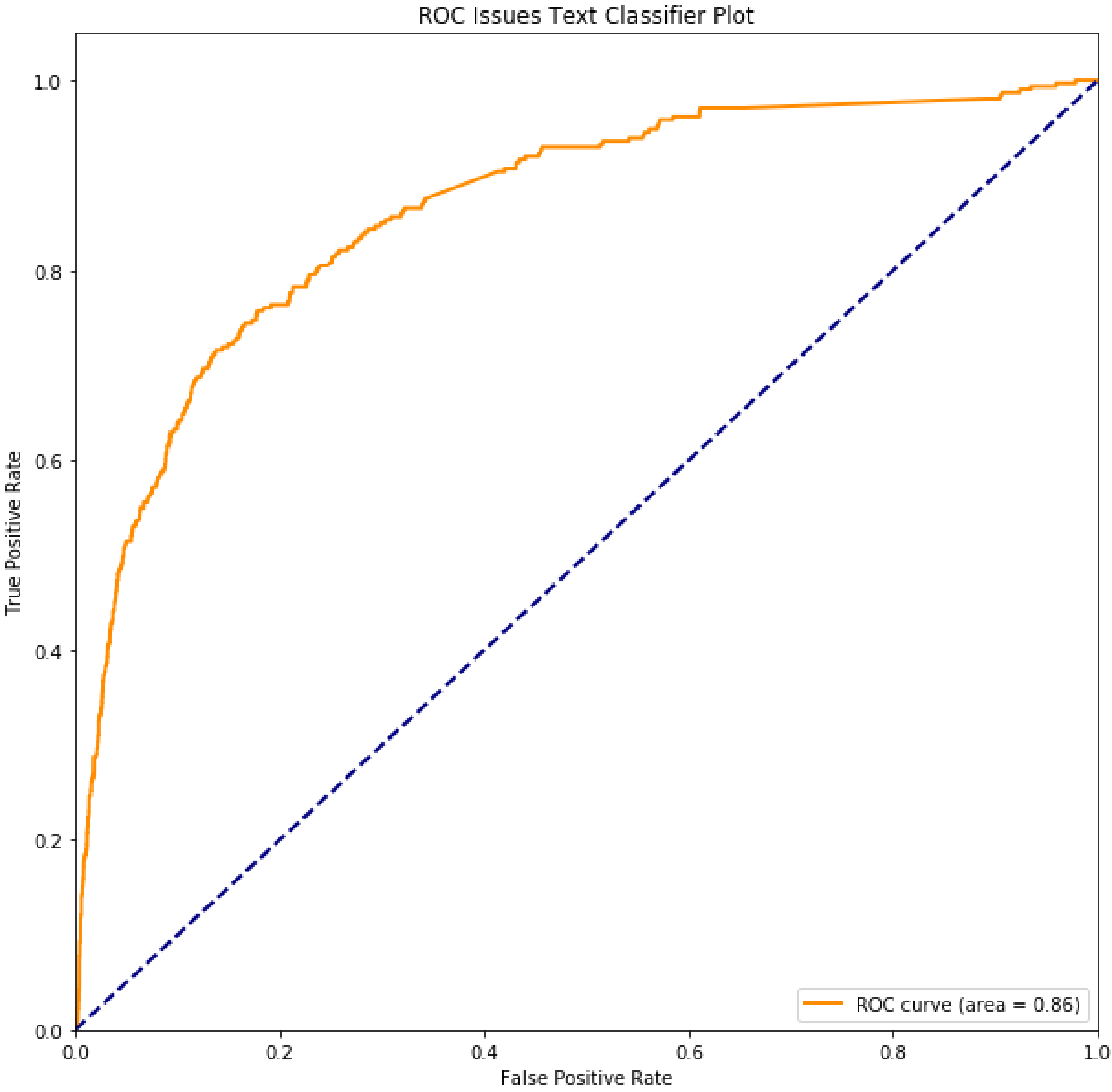


Figure 2: Issues Dataset - Metrics

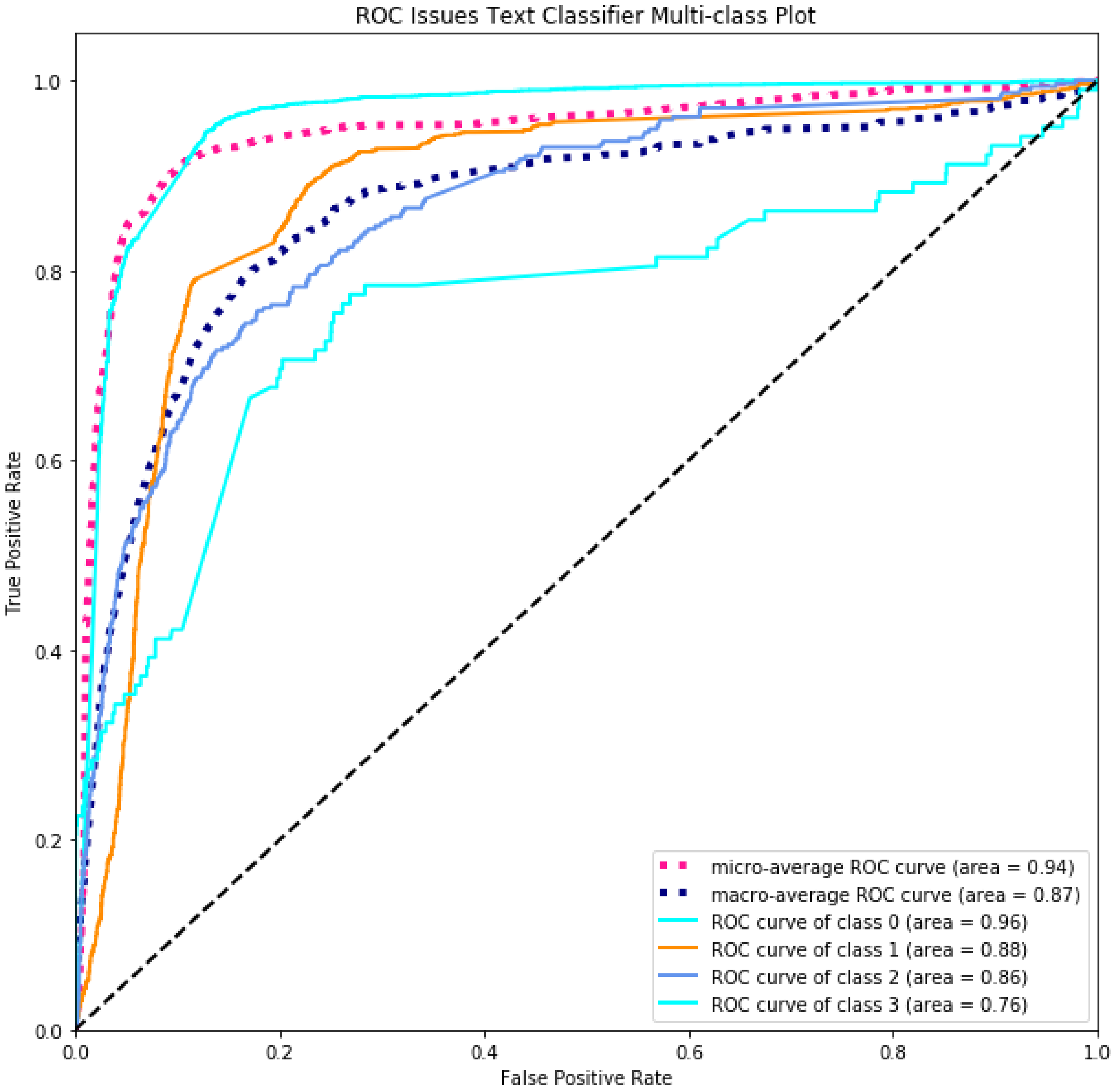


Figure 3: Issues Dataset - Metrics

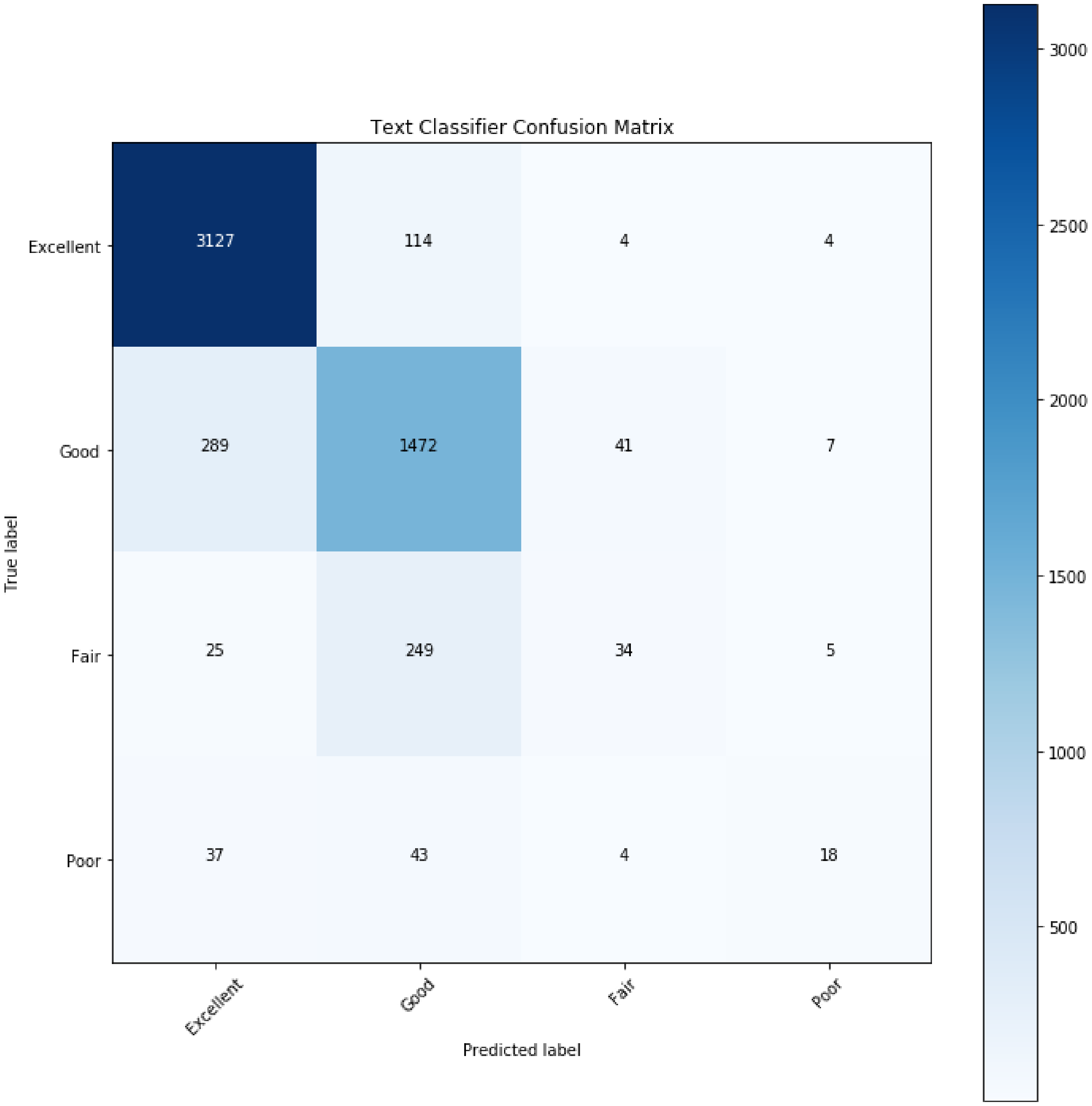


Figure 4: Issues Dataset - Metrics

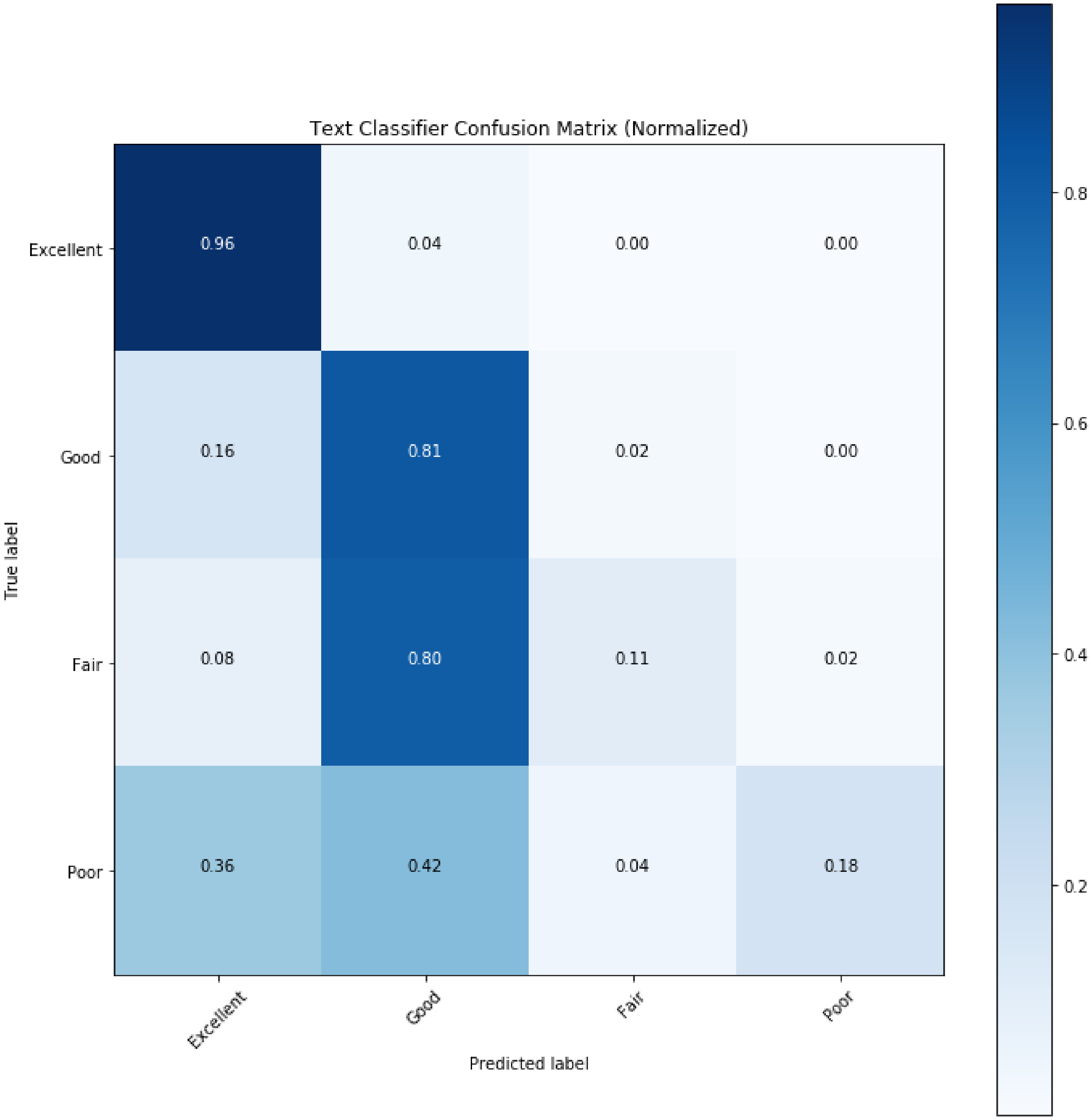


Figure 5: Responses Dataset - Metrics

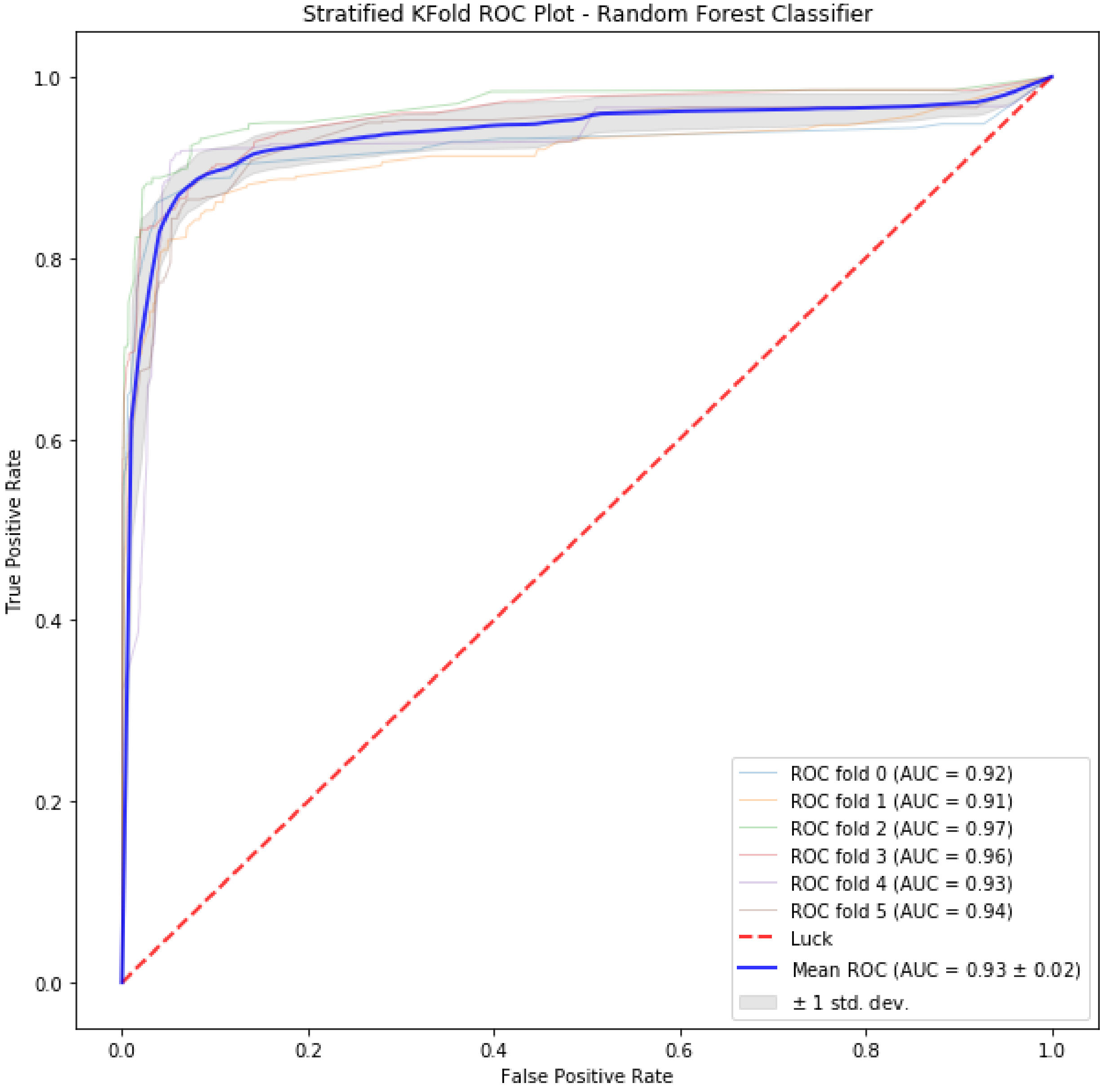


Figure 6: Responses Dataset - Metrics

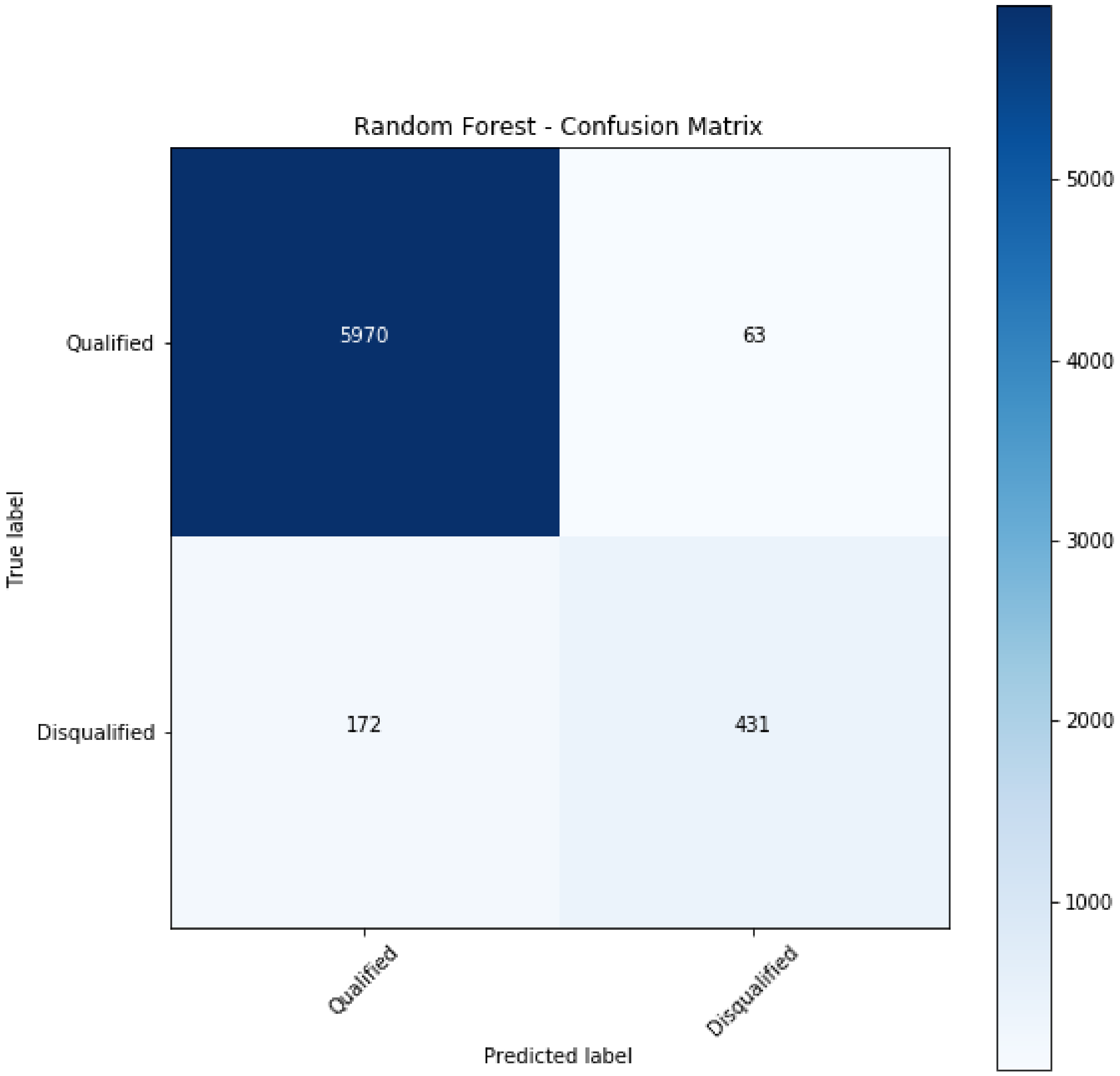


Figure 7: Responses Dataset - Metrics

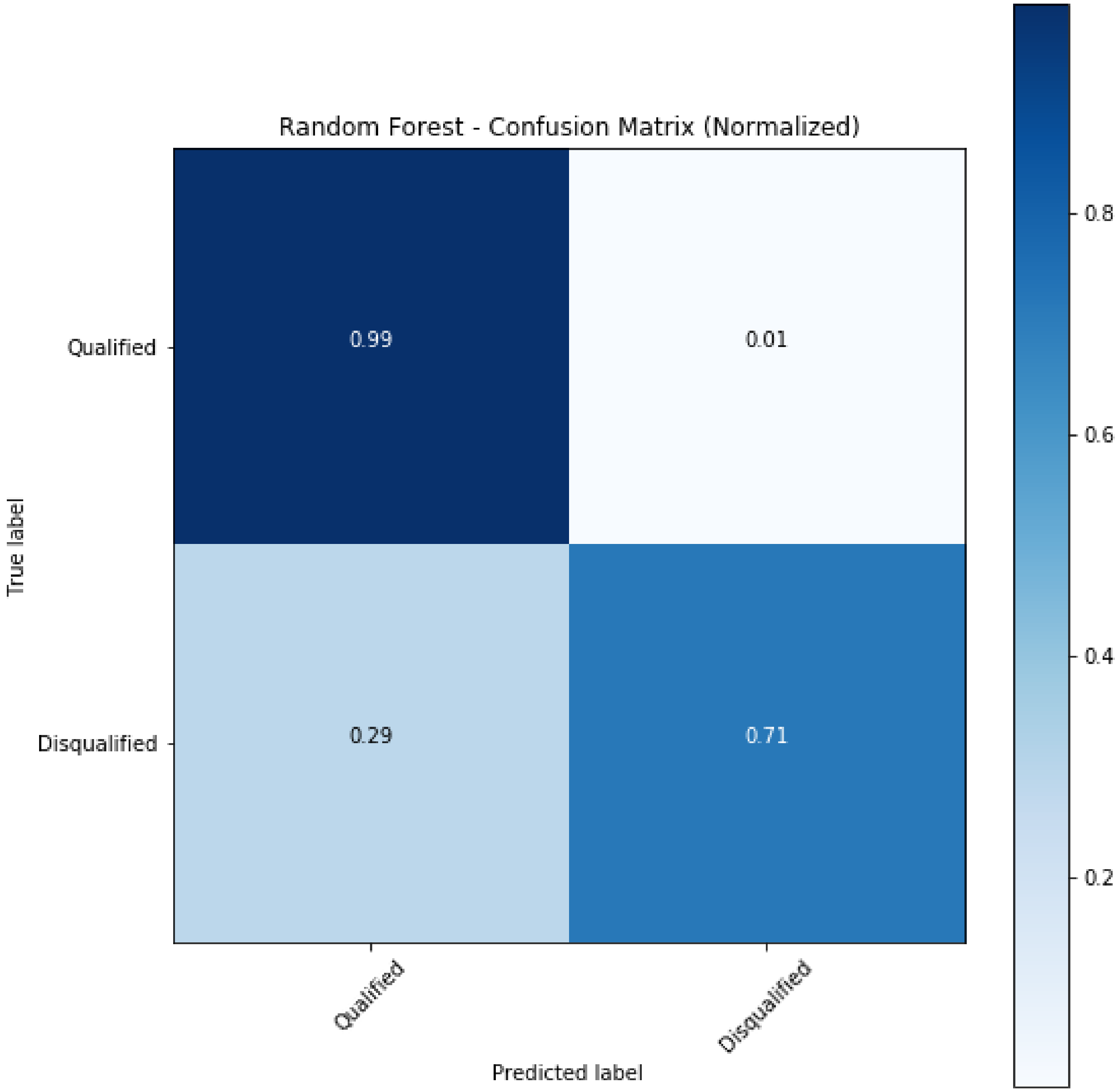


Figure 8: Responses Dataset - Metrics

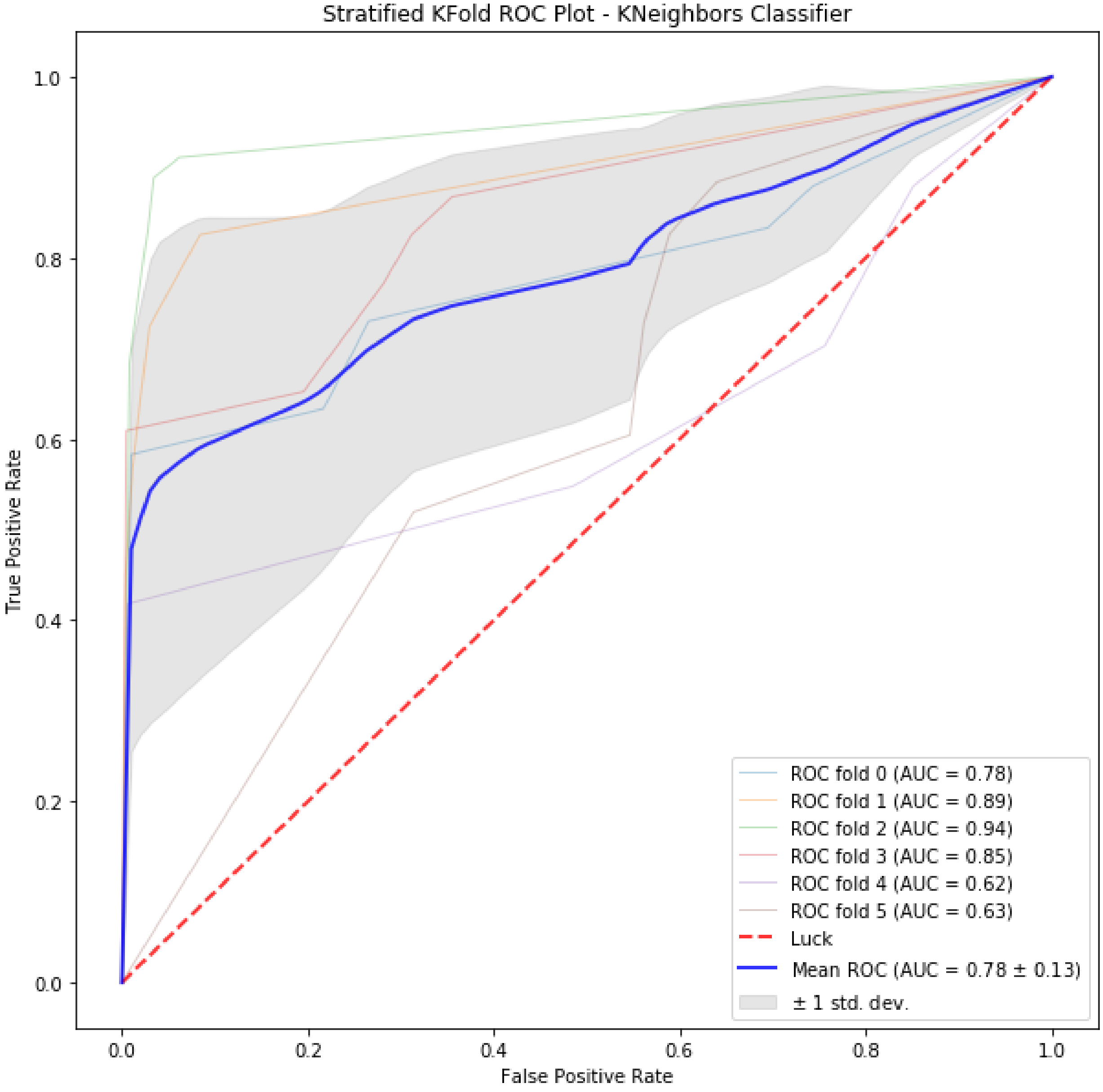


Figure 9: Responses Dataset - Metrics

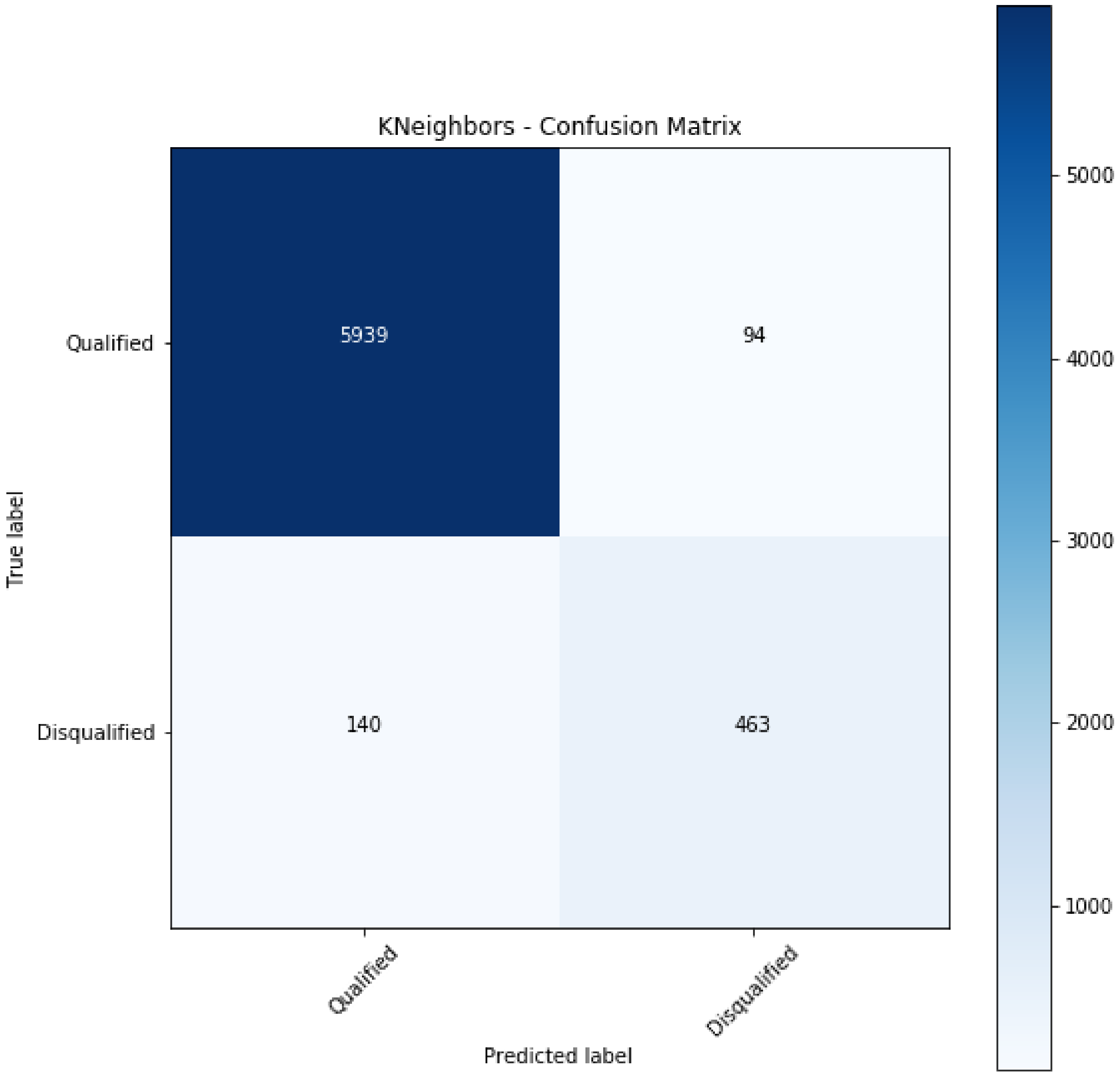


Figure 10: Responses Dataset - Metrics

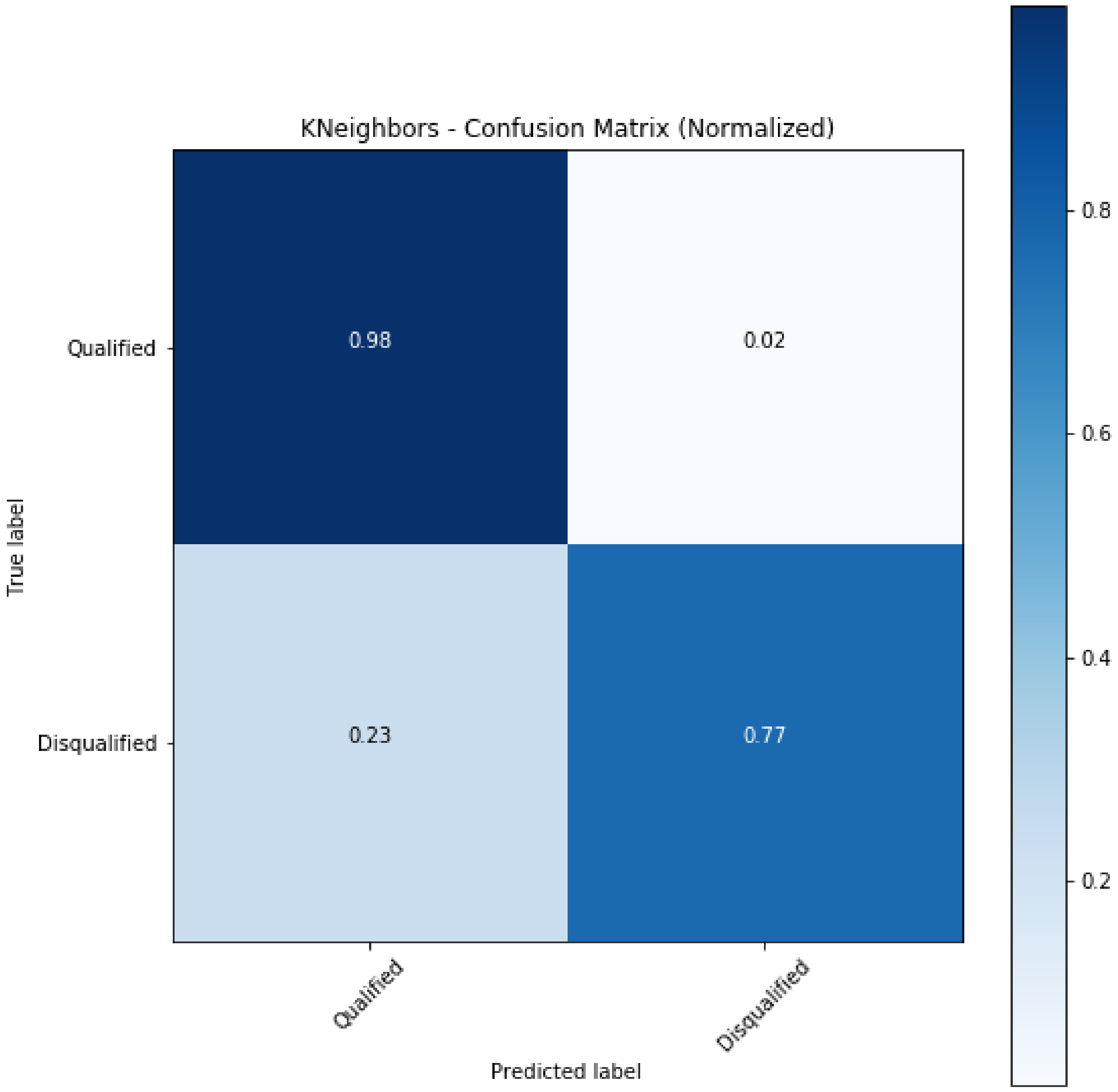


Figure 11: Responses Dataset - Metrics

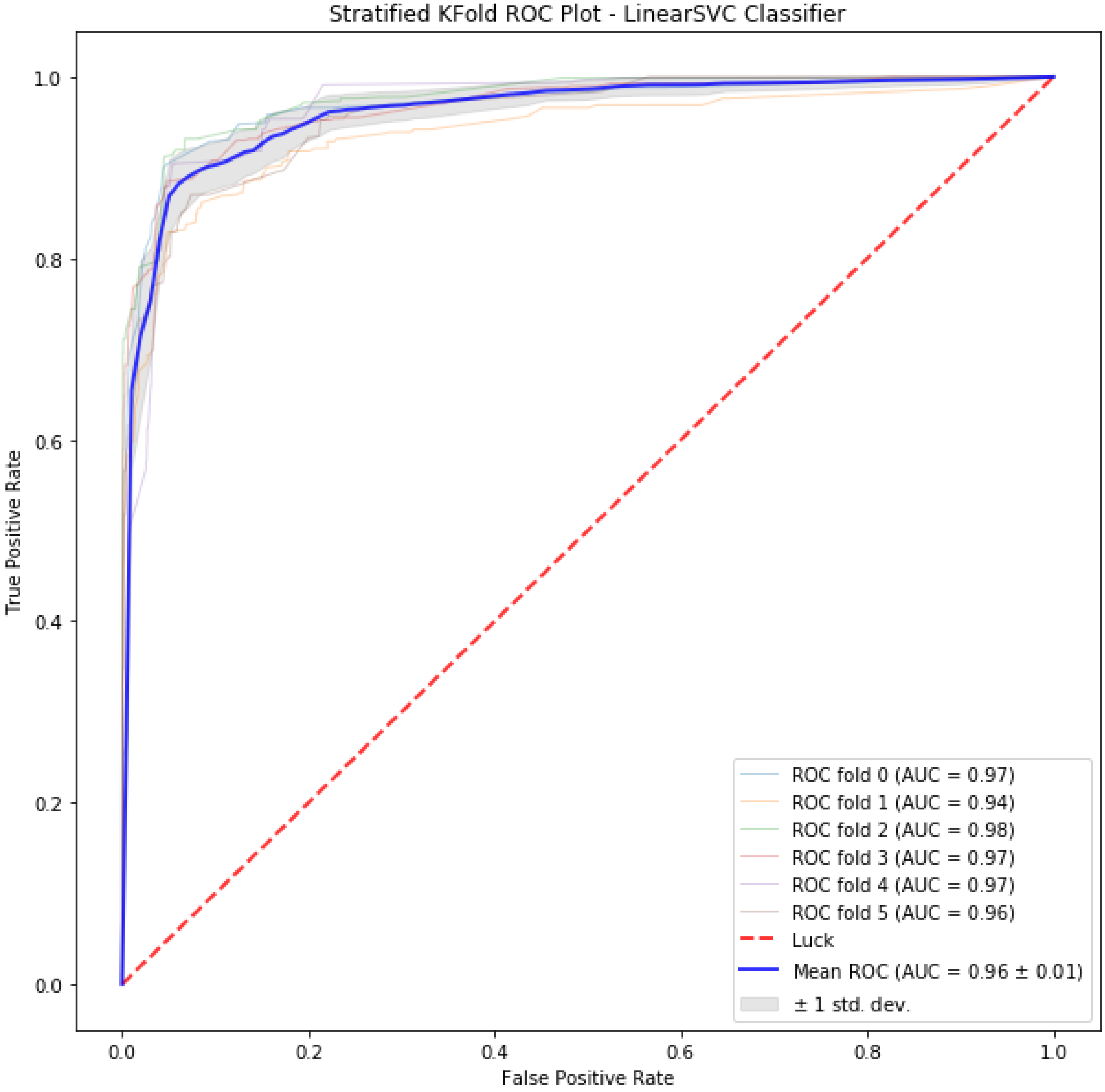


Figure 12: Responses Dataset - Metrics

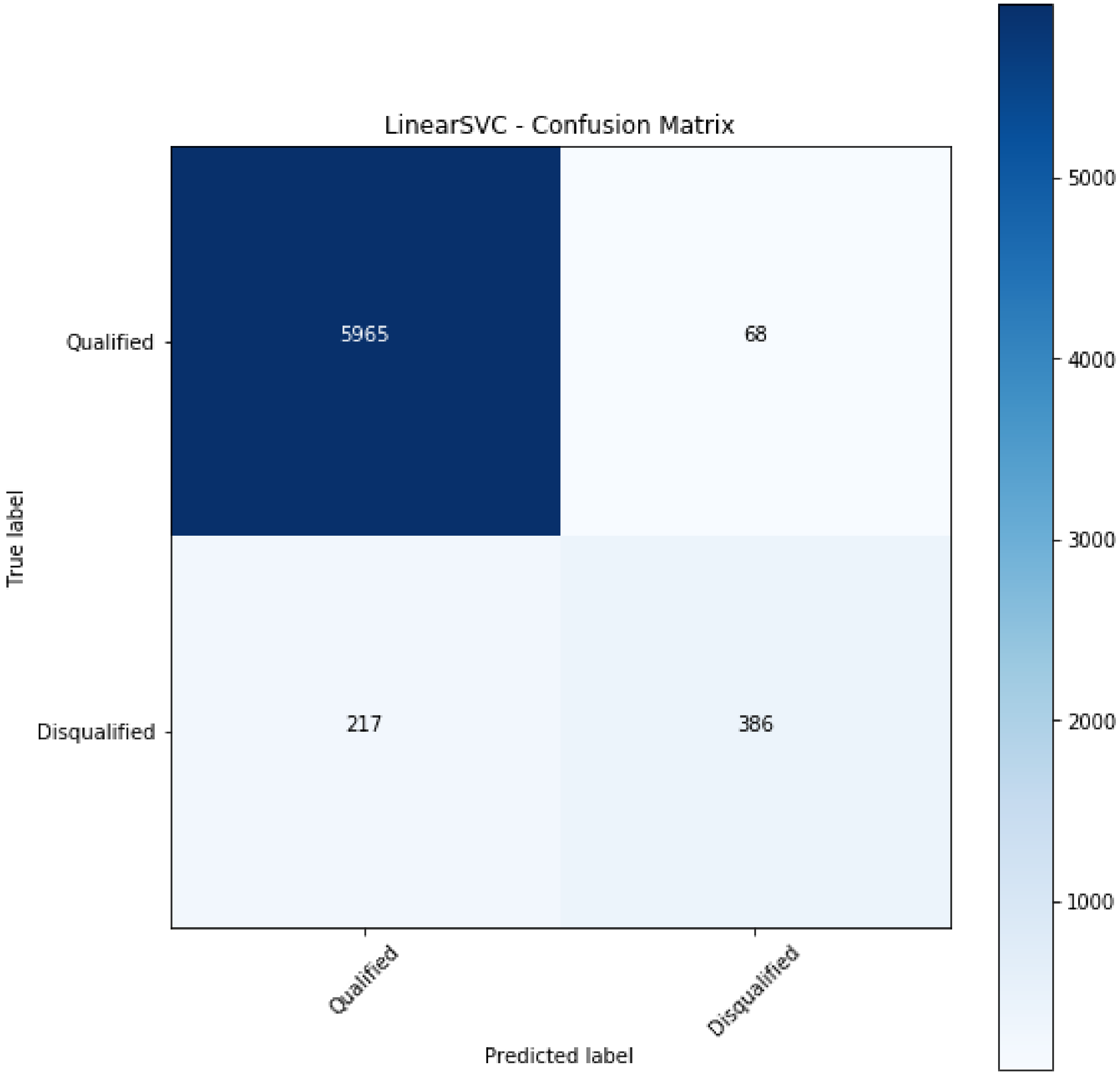


Figure 13: Responses Dataset - Metrics

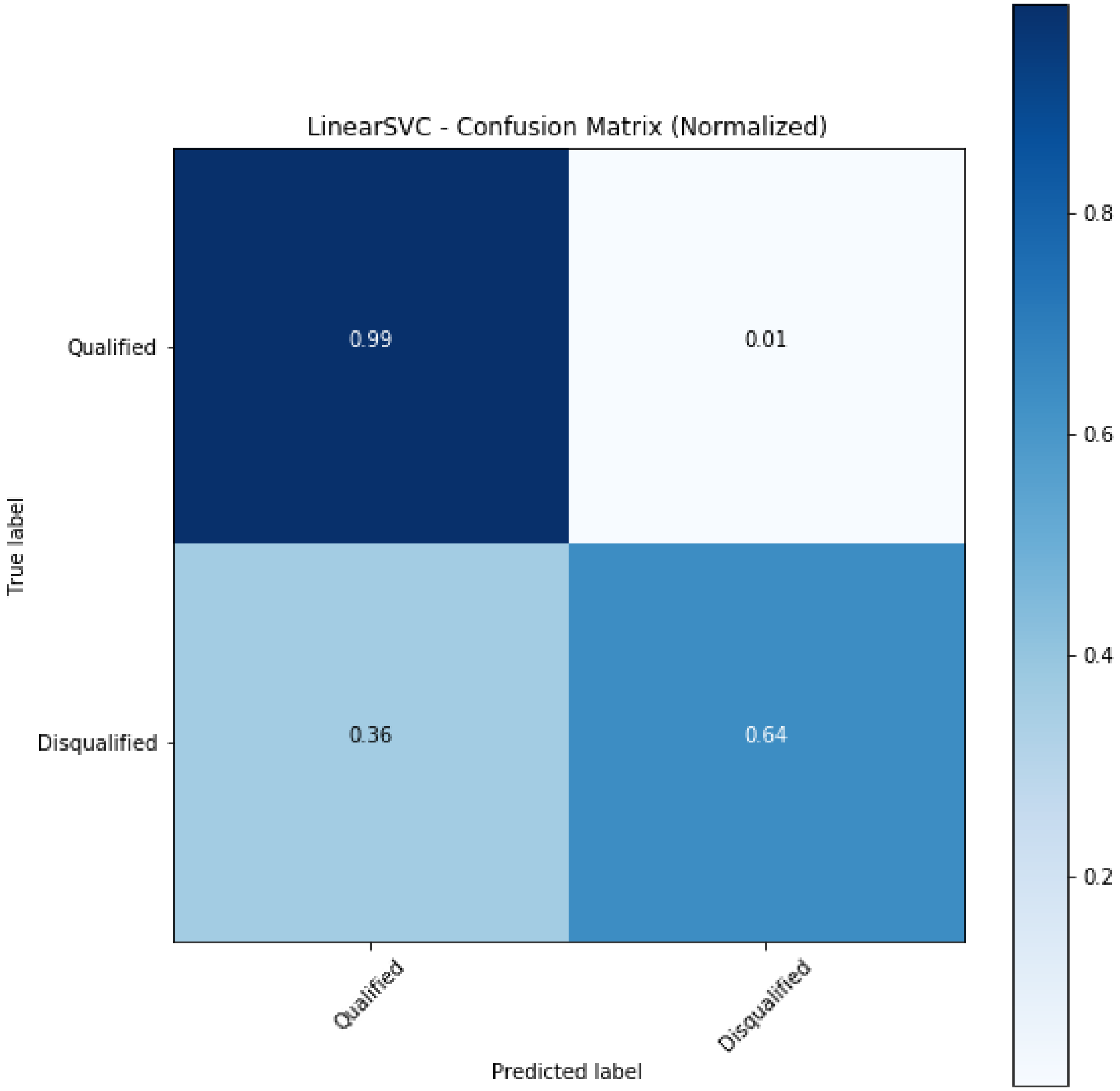
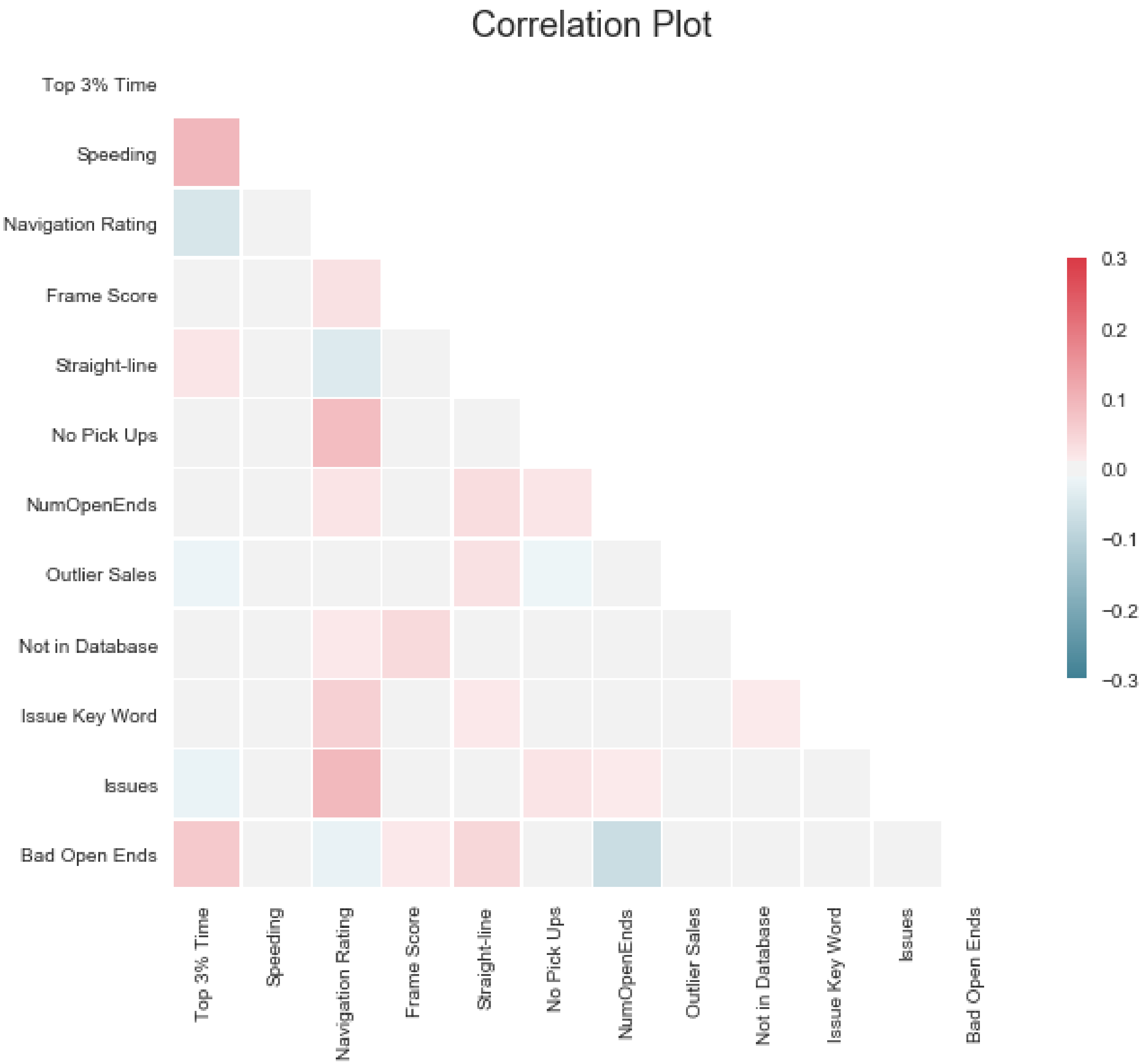


Figure 14: Raw Dataset


Correlation Plot
Top 3% Time
Speeding
Navigation Rating
Frame Score
Straight-line
No Pick Ups
NumOpenEnds
Outlier Sales
Not in Database
Issue Key Word
Issues
Bad Open Ends
Top 3% Time
Speeding
Navigation Rating
Frame Score
Straight-line
No Pick Ups
NumOpenEnds
Outlier Sales
Not in Database
Issue Key Word
Issues
Bad Open Ends