**Market Research: Quality Control Automation**

**1 Abstract**

**2 Introduction**

*About the company and Market Research in general*

The market research industry is being transformed by the demands of multichannel marketing, changes in customer behavior, and the availability of new technologies and services.  
  
The digitalization of research methods and the emergence of analytics tools to extract insight from new sources of structured and unstructured data can provide unprecedented insight into customer behavior and calls to action. InContext is one such company, offering category managers and retailers tools to quickly ideate and iterate on new business ventures. By creating virtual, shoppable store environments, they are able to assess the efficacy of different category assortments, shelf arrangements, new signs/displays, and store layouts without having to spend the time and money setting these up in store. Not only is this a powerful tool for gauging customer behavior, but it also allows the business to visualize ideas before investing additional time and resources towards new and risky ventures.

When it comes to the research component of InContext Solutions’ product offerings, there is a lot of time and effort invested in ensuring that the quality of the survey/virtual experience data is top tier. Each study is head up by an analyst on the Insights Team who manages the project from beginning to end. 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 analyst either decides to keep the respondent or disqualifies them.

The disqualification (DQ) process takes 4-5 hours per 1000 respondents. Given that studies contain 4000 respondents on average, this amounts to 16-20 hours per study. A lot of this time is spent manually pulling/joining data and sifting through open ended responses for signs of difficulties. The aim of this project is to train a classifier to fully automate this process.

**3 Data Processing**

*Data Description - Data size: 44,235 labeled respondents, 10 features*

This data comes from a company called InContext Solutions and is used for market research. It is generated using surveys and online panels. It is key for the company to be certain that their survey responses are reliable in order to provide insight into making valuable marketing and business decisions. Quality controls are performed every thousand completed surveys. The following features are considered qualifiers and disqualifiers for survey respondents and are utilized for quality control.

Usually InContext Solutions can afford to lose a given number of tests and currently their disqualification rate is approximately 10%. Respondents are disqualified based upon a combination of the following 11 predictors: top 3% time and speeding (respondents who complete the survey faster than is reasonable), navigation rating and frame score (respondents are identified as having had a poor experience in the virtual store), issue key word and issues (respondents' feedback to their experience), straightline (identifies unengaged respondents such as those who select the same response for all questions), bad open ends (another measure to identify unengaged respondents), no pick ups (identifies respondents who fail to pick up any product), outlier sales (when respondents do not engage seriously in the virtual store including engaging in unrealistic purchasing behaviors), and not in database (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 considered a bad respondent and is disqualified (although this measure varies depending on the number of open ended responses in the survey).

Most of these scores can be automated by standardizing a procedure and writing a script to combine these. However, issue key words, issues, and bad open ends are all based on the judgement of the analyst. Additionally, the scoring penalty for these columns are adjusted according to the analyst’s discretion (usually according to the number of open ended questions asked in the survey). For this reason, we also add in a predictor for the number of open ended responses in the survey.

We compiled previously scored and labeled data from the QC process from a single analyst from 13 separate studies from the past year amounting to 44,235 respondents. In order to use the raw text data, we pooled each of the respondents’ open-ended responses into one column. We combined the scores for the aforementioned columns, applied the term frequency–inverse document frequency algorithm to the pooled text column to extract word vectors, scaled the predictors, and then used a linear support vector classifier to classify it as one of 0, 1, 2, or 3.

Using this score in tandem with the other metrics, we trained multiple binary classifiers using an 80/20 train test split in order to predict whether or not respondents were disqualified. Their respective performances are listed in the Results section.

**4 Methodology**

A number of methods were used in examining the data in this project. K-nearest neighbors provides a simple method of classifying any given observation. This involves identifying the K points closest to a given observation and estimating the conditional probability for a certain class as the fraction of K points whose explanatory values equal that class. The size of K has a large effect on the classification of a given observation. When K = 1, the decision boundary can be overly flexible, and results in a classifier that has low bias, but very high variance. As K increases, the method becomes less flexible giving it a lower variance, but a higher bias. [relate to the data]

Random forests provide a second method of examining our data. Decision trees segment or stratify the predictor space into a number of regions, allowing us to classify each observation. Random forests are a way to improve accuracy by producing multiple trees that are then combined to yield a single consensus prediction. Random forests decorrelate the trees by not considering a majority of the available predictors at every split in the tree. Only a number of predictors are considered at each split. This difference from bagged trees prevents highly correlated predictions. Averaging many uncorrelated quantities leads to a larger reduction in variation among the trees, and therefore increases reliability. [relate to data]

Discuss:

* accuracy isn’t the best metric here as dq’s represent ~10% of the data. So, if we predicted all not-dq’d, we would have 90% accuracy
* Precision vs Recall vs F1-score
* Confusion Matrices

Models considered:

* Random Forest
* SVM (both linear and non-linear)
* K-Nearest Neighbors

**5 Results**• Non-linear techniques slightly outperformed linear ones.

The models ranked according to their cross-validated f1-scores, are as follows:

1. MLP Neural Network - leave out no need to add to methods lets just say that it will be further work and mention preliminary results
2. Random Forest
3. KNN
4. Linear SVM

**6 Conclusion**• A statement of the implications of your study.  
• A discussion of further questions raised by your study.

This process is prone to human error due to boredom and frustration.

to conclude, we are empowering analysts by freeing up their time to address client needs and enabling them to focus on more meaningful work.