Final Project Report

Darragh Hanley

Stanford ID 06001134

**Abstract**

The goal of a search engine is to return relevant URL documents for search queries that users enter. The target of this project is to return the relevant URLs for a given set of information on the queries within the search engine. We try ZZZ different classifiers : zzzzz. ZZZZ performs best based on the experiments… a bit further on checks.

1. **Introduction**

The goal of this project is to predict relevant URLs within a test data set using attributes for each query. The training data set contains 80046 records and 10 attributes which could be used to help predict whether the URL is relevant for the query. In addition it includes the query id, URL id and the relevance of each URL. The relevance is a binary value (1 for relevant, 0 for not relevant) with 43.7% of queries being relevant. Additionally, a test data set is provided which contains 30,001 observations.

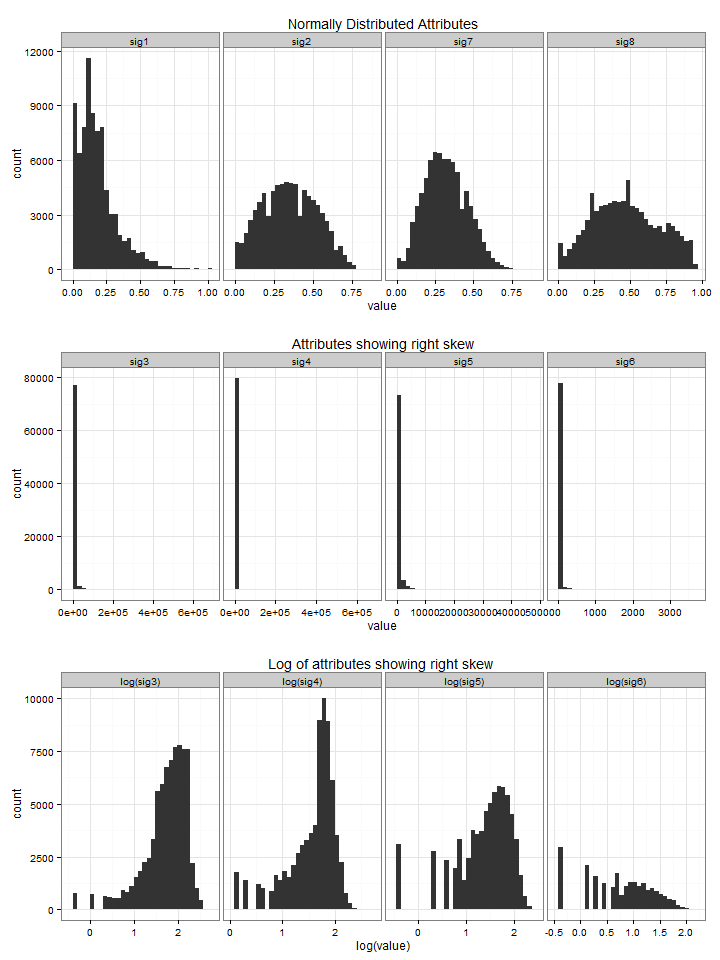
1. **Data Observation and Transformations.**

Both the training and testing data set have no missing values or duplicated rows. The data contains two nominal attributes, the query ID and the URL ID representing search engine query and URL data. URL ID is unlikely to be a significant predictor as over 94% of records have a unique value. query\_id is more common with only 12408 unique ids. Value query\_length is assumed to represent the number of terms in a query. Query length did impact the number of relevant URLs as can be seen in the mosaic plot below, with 2 being the median query length as well as achieving the highest proportion of relevant results. The value is\_homepage is a binary field and likely to represent whether the query is launched from the users homepage. The queries from homepages appear to have a higher number of relevant URLs (48.6%) than the queries not from homepages (41.9%). Using a binomial probability test, we have seen it is highly unlikely this is due to chance (p-value < 2.2e-16). One possible reason is that users which set the search engine as homepage are better at entering optimal search terms.

The remaining numerical signals have no indication to their exact meaning, and are labelled signal 1 through 8. A table of summary statistics can be seen in appendix A. The 8 signals were plotted on a histogram to check distribution. Four signals showed strong bias right bias. These signals were transformed to near normal distribution using logging, before logging the 0 values were changed to .01 to avoid the logs returning infinite values. A correlation matrix for the numeric variables was calculated both before and after logging and can be seen below. Although some signals showed a high correlation, particular signals 3 and 5 (correlation of 0.815 without logging, and 0.782 with logging) we judged none were high enough to leave out of the model. It can be noted that correlation among a number of signals, and with URL relevance, increased with logging. No single attribute showed a high correlation with URL relevance however correlation does increase slightly with logging on signals 3 through 6.

We tested aggregating attributes according to their query\_id using the aggregated max, min, mean and median values of the individual rows, see appendix B. We could boost the relevancy correlation of sig8, from .0287 to .0641, by taking the minimum value per query. Therefore this field will be added to the data set.

Add new transformation section, add scaling to transformations,



Correct the log charts, they have the log of the log

****



1. **Classifier Evaluation**

**We used 6 different classifiers in this project. In order to evaluate each, we use 5 fold cross validation. We randomly split the data into 5 distinct blocks of roughly equal size using the caret package from R which contains tools for data splitting. First, we leave out the ﬁrst block as a hold out block of data and ﬁt a model on the remaining data, this model is used to predict the held-out block. Then we continue this process until we’ve predicted all 5 held–out blocks. The ﬁnal performance is based on the average misclassification error of all blocks.**

1. **Candidate Solution and Data Selection**

**Give an initial blurb….**

* 1. **Decision Tree**

**The decision tree generally performed less effective however we include the results for benchmarking. We used function rpart() in the library ‘rpart’ to generate a decision tree in R. The default parameters using the transformed inputs generated a misclassification error of 36.3%, however improvements could be found on this by varying the max depth of the tree. Different split criterion varied results at larger max depths which may point to Gini being more prone to overfitting.**



**Appendix**

1. **Summary Statistics**

**Summary Statistics of original numeric variables, excluding factor column: is\_homepage.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | query\_length | sig1 | sig2 | sig3 | sig4 | sig5 | sig6 | sig7 | sig8 |
| NAs | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Minimum | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Maximum | 18.0000 | 1.0000 | 0.8600 | 673637.0000 | 660939.0000 | 46994.0000 | 3645.0000 | 0.8800 | 0.9400 |
| 1. Quartile | 2.0000 | 0.0800 | 0.2100 | 78.0000 | 24.0000 | 10.0000 | 0.0000 | 0.2200 | 0.2900 |
| 3. Quartile | 3.0000 | 0.2400 | 0.4800 | 2537.7500 | 591.0000 | 336.0000 | 2.0000 | 0.4200 | 0.6400 |
| Mean | 2.5858 | 0.1832 | 0.3469 | 4857.0786 | 742.3163 | 550.5276 | 14.0992 | 0.3195 | 0.4718 |
| Median | 2.0000 | 0.1500 | 0.3400 | 417.0000 | 220.0000 | 64.0000 | 0.0000 | 0.3100 | 0.4600 |
| Variance | 2.3168 | 0.0217 | 0.0298 | 553753762.6768 | 23216584.6626 | 3564294.6671 | 8112.3214 | 0.0192 | 0.0535 |

**Summary Statistics of numeric variables after transformations for use in models, excluding factor column: is\_homepage.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | query\_length | sig1 | sig2 | log\_sig3 | log\_sig4 | log\_sig5 | log\_sig6 | sig7 | min\_sig8 |
| NAs | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Minimum | 1.0000 | 0.0000 | 0.0000 | -6.9078 | -6.9078 | -6.9078 | -6.9078 | 0.0000 | 0.0000 |
| Maximum | 18.0000 | 1.0000 | 0.8600 | 13.4204 | 13.4014 | 10.7578 | 8.2011 | 0.8800 | 0.9300 |
| 1. Quartile | 2.0000 | 0.0800 | 0.2100 | 4.3567 | 3.1781 | 2.3026 | -6.9078 | 0.2200 | 0.1100 |
| 3. Quartile | 3.0000 | 0.2400 | 0.4800 | 7.8390 | 6.3818 | 5.8171 | 0.6931 | 0.4200 | 0.3100 |
| Mean | 2.5858 | 0.1832 | 0.3469 | 5.8232 | 4.1913 | 3.2406 | -3.7825 | 0.3195 | 0.2202 |
| Median | 2.0000 | 0.1500 | 0.3400 | 6.0331 | 5.3936 | 4.1589 | -6.9078 | 0.3100 | 0.2000 |
| Variance | 2.3168 | 0.0217 | 0.0298 | 8.5635 | 13.4253 | 16.8198 | 19.2046 | 0.0192 | 0.0210 |

1. **Correlation to Relevance field after aggregating numeric attributes over query\_id**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | original value | Max per query\_id | Min per query\_id | Median per query\_id | Mean per query\_id |
| query\_length | -0.00005431 | -0.00005431 | -0.00005431 | -0.00005431 | -0.00005431 |
| is\_homepage | 0.06028810 | 0.02696290 | 0.02051457 | 0.00549458 | 0.01817045 |
| sig1 | 0.16019789 | -0.00570228 | 0.02599883 | 0.03311323 | 0.02281130 |
| sig2 | 0.30581508 | 0.11946646 | 0.13694133 | 0.16077294 | 0.17057616 |
| sig3 | 0.07274379 | 0.02448695 | 0.05003876 | 0.04491426 | 0.03437706 |
| sig4 | 0.03425501 | 0.00988298 | 0.04032274 | 0.01875498 | 0.01804620 |
| sig5 | 0.10322486 | 0.02844414 | 0.05055365 | 0.05915561 | 0.04528149 |
| sig6 | 0.12451341 | 0.02026424 | 0.04411823 | 0.05842145 | 0.04081394 |
| sig7 | 0.16514143 | 0.06472197 | 0.11874826 | 0.10455677 | 0.10584065 |
| sig8 | 0.02873055 | 0.01310363 | 0.06413252 | 0.03939868 | 0.04451280 |