Data Exploration and Preparation

The raw dataset required significant processing prior to analysis. The raw dataframe build directly from the data downloaded from the Ravelry API was in the following format:

|  |  |
| --- | --- |
| Attribute name | Type |
| discontinued | Boolean |
| gauge\_divisor | Integer |
| grams | Integer |
| id | Integer |
| machine\_washable | Boolean |
| max\_gauge | Integer |
| max\_hook\_size | list |
| max\_needle\_size | list |
| min\_gauge | Integer |
| min\_hook\_size | list |
| min\_needle\_size | list |
| name | String |
| notes\_html | String |
| organic | Boolean |
| permalink | String |
| rating\_average | Float |
| rating\_count | Integer |
| rating\_total | Integer |
| texture | String |
| thread\_size | list |
| wpi | Integer |
| yardage | Integer |
| yarn\_company | [string](http://www.ravelry.com/api#YarnCompany_full_result) |
| yarn\_fibers | list |
| yarn\_weight | [list](http://www.ravelry.com/api#YarnWeight_full_result) |

Processing the dataset required two steps. First, the data for all attributes was coerced to type numeric in order to simplify application of machine learning methods. Secondly, attributes which were considered redundant or not relevant to the analysis were removed.

Coercing all attributes to numeric type was carried out first as this gave an expanded list of attributes which could then be selectively removed. The attributes which were of type ‘list’ required the most extensive processing. The most extensive of these was the ‘yarn\_fibers’ attribute which contained the physical makeup of each yarn. For each yarn in the database, this attribute consisted of a list of fiber types. Each fiber type was itself a list in the following format:

|  |  |
| --- | --- |
| List element | Type |
| Fiber\_type$name | String |
| Fiber\_type$animal\_fiber | Boolean |
| Fiber\_type$vegetable\_fiber | Boolean |
| Fiber\_type$synthetic | Boolean |
| Fiber\_type$id | Integer |
| $id | Integer |
| $percentage | Integer |

*Table xx – elements present in a Fiber\_type list. The yarn\_fibers attribute for each yarn in the database consisted of a list of Fiber\_type elements.*

The ‘data\_processing’ script (Appendix XX) converted the yarn\_fibers attribute into a more useful form by extracting the percentage of each fiber type in each yarn. This required the creation of 26 new attributes – one for each fiber type, with each representing the percentage of that fiber type present in each yarn instance.

The other attributes of type ‘list’ in the raw data were processed in the same manner to numeric attributes. Expanding the lists in this manner resulted in a dataset of 91 attributes. Those of type ‘boolean’ were converted to numeric (TRUE=1, FALSE=0).

The second step of data processing was then carried out by removing attributes considered redundant or outside of the scope of this investigation. All text-only attributes were removed (yarn description, company name etc). All attributes related to hook or needle size were removed since these are dependent on the yarn weight and therefore carried no additional information. There were several attributes for yarn weight since the raw dataset included the yarn weight in several different units of measurement. Therefore, all attributes relating to yarn weight were removed with the exception of ‘yarn\_weight\_ply’.

The ‘id’ attribute is simply a numeric identifier for each yarn. It was decided to leave this present in the dataset since the id is assigned sequentially as new yarns are added to the database. The ‘id’ attribute therefore gives an indication of the relative age of yarn, with older yarns having smaller ids.

Finally, three new attributes were created from the ‘yarn\_fibers’ attributes. These new attributes named ‘vegetable’, ‘animal’ and ‘synthetic’ recorded the percentage of vegetable, animal and synthetic fiber in each yarn. These give a more general overview of each yarn’s composition and it was believed that including them may improve the quality of any correlations.

After processing, the dataset consisted of 43 independent variables, along with the dependant ‘rating\_average’ variable (henceforth referred to as ‘yarn rating’).

**PCA**

**References**

**http://www.ravelry.com/api**