Model-Building

Performing exploratory analysis helped build better intuition into the various features of the dataset and shed light on how diverse hops can be. Most notably, the shared characteristics of hops among a particular region or purpose showed potentially interesting relationships that could benefit from additional studies. To that extent, machine learning models were explored to delve deeper into these relationships and develop even stronger answers to our research questions. Taking this a step further, a predictive tool was also created to classify hops in order to output a practical application for this project.

As this is a classification problem, tree-based ensemble algorithms (Random-Forest & XG-Boost) were explored in an effort to predict the region or purpose of a hop. These methods are known for their robustness in the industry as both rely on multiple weak learners that are combined into a strong learner. However, bagging methods such as Random-Forest build these models in parallel while boosting methods such as XG-Boost build sequentially. Both techniques are proven to be viable choices for multi-level classification scenarios, as with this study.

While the exploratory analysis gave the proper insight into which attributes to classify, further feature-engineering was required to prepare the data for the model. The outcome variables for both studies (region and purpose) were encoded to transform the categorical variables into appropriate numeric form for the models. Dummy-vectors were created for the predictor categorical variables that included the various aroma tags. Afterwards, this data was partitioned to create a training and testing set to train and evaluate the models. Although the final accuracy rates fluctuated with splitting, both methods seemed to classify hops based on region, or purpose, at an approximate rate of 70%.

Although this was a respectable result, the exploratory work seemed to promise more. There are several reasons that could be contributing to this. Most importantly, although our dataset has a large amount of features, the total number of hops are only 304. While this data was scraped from the most reliable hops database in the world, there is an innate issue in the continuous data that is hard to overcome while training these models: the high variability of the brew values. As the database also claims, individual hop measurements can differ substantially depending on a lot of unaccounted factors.

To compensate for this in future studies, exploring statistical methods that are designed for smaller-level studies seems to be an attractive option at this juncture. Either way, modeling our dataset helped understand the data on a deeper level, and output a tool that can be strengthened in the future.