ECNS 460

Analyzing Cannabist Company Product Pricing and Recommending via Chemical Profiles

Fred Syata
Alex Westlund

Analyzing if chemical profile can predict pricing

The first portion of the project analyzed by category (flower, concentrate, edible, vape). The models were trained on all chemical predictors that were frequently present in the data. We excluded chemicals that had more than 90% NA values from the models due to sparsity. The linear models were not particularly successful, except in the case of predicting edible prices where it achieved an adjusted R-squared value of 0.649. The random forest models, however, consistently have lower RMSE values and in every case except for Flower perform extremely well. This exploration was inspired by the visualization stage and due to interest from the parent company.

Categorizing product type based on terpene data

We wanted to see if we could accurately predict what category a set of terpenes data might belong to based on the terpenes data. We used a multinomial logistic regression which is very similar to a logistic regression but instead of the dependent variable being 1 or 0 it is now between 4 categories; flower, edible, vape, and concentrate. The errors are very low at 5.014% error rate. When testing this regression model we can confidently say that we can categorize the product type by the terpenes data.

User dataset

Next we wanted to create a recommender that analyzes the users past terpenes data and finds a product that is similar to their preference. However our data does not contain user information so we decided to randomly assign 90 rows to each user without replacement. Obviously user data is not random but we wanted to establish a working process for this recommender which will work once we replace the fake users with real users. This is our main dataset that will be used in your recommending process. We took a simple approach at reducing the user data for each terpene down to a single value which we could analyze. We decided to take the average percentage of each terpene to get the users average preference across all terpenes.

Recommender function/process

Next we created the recommender function which takes in two arguments a vector of terpenes and the product type for a single product. This function outputs the top 100 users that have the most similar average preferences of each terpene and the predicted price of that product using both linear regression and random forest methods. To calculate how close a person preference

was with the imputed terpene data we use Euclidean distance which sums the row of terpene data and subtracts it by the imputed sum of terpene data and then squares it to get a distance from how different they are from each other, which we then select the 100 lowest distances and their user id. With this function created a producer or seller could input the data of one of their products and get 100 people that would be the most likely to purchase it as well as a reference of how to price it based on other products similar to it. This can be incredibly helpful for a business by being able to market a new product to the right audience and how to price it within the market.

AI (ChatGPT) was used to assist in the creation of non-ugly plots and figures, and for debugging issues when implementing random forest modeling. Also helped with Euclidean distances.