IOWA DOCUMENT [Joseph Pugh]

March 29, 2022

1 Market expansion with trust and liquid courage

Exploring liquor store buying habits to find new consumers By Joseph Pugh

1.1 Overview

Alcohol is a difficult commercial product. In the United States alcohol is heavily regulated. These regulations even change from state to state. The ability to advertise is limited. The locations where alcohol is sold also restricted. When a new product enters the market adoption can be slow. How do you find consumers?

This report utilizes the state of Iowa's liquor database. This database contains all invoices between

retailers and distribution centers since 2009. The database is updated monthly. We will be focusing specifcally on a type of alcohol called mezcal.

Mezcal is a type of alcohol produced in Mexico made from any species of agave plants. Each species having their own characteristics. If you are unfamiliar with mezcal, think of it as artisanal tequila. Mezcal has been in the United States market for nearly 30 years, with popularity dramatically rising in the past decade.

Through this report we have created a method to find populations of customers with similiar buy habits as those who already consume mezcal. Customers we feel would be likely adopters of mezcal.

1.2 1. Method

The data is from the Iowa Liquor Sales Database. This website is run my their state government and contains a bunch of datasets on different subjects and some analysis.

This dataset contains over 22 million rows, 24 features, and covers the entire state for over a decade. We are going to focus on the greater metropolitan area of the largest city, Des Moines. After cleaning the data set we are going to rank stores by their sales of mezcal products. The ranking system will be comprised of their total sales and variety of offerings. Once all stores have a ranking associated with them we will remove the invoices for mezcal products.

The remaining dataset will be encoded and run through an unsupervised clustering algorithm. This is to see if there is any underlying similiarities between stores based on their liquopr purchasing. We will store any meaningful result from the clustering as a new feature to be included in a classification model.

Once the classification model fits the data we will look specifically at what is traditionally Type 1 error. Type 1 error are false positives. In this case it would be stores which the model believes sell mezcal even though they currently do not.

We spend the time creating and tuning models until we trust them just to say it is wrong some of the time but in acceptable ways. What if the model is not wrong and human error is responsible for the discrepancy? That is the underlying idea. Create a model to predict whether or not a liquor store sells mezcal based on their other invoices.

If the model believes a store should be selling mezcal then that store should be a good candidate to be approached with mezcal products.

1.3 2. Data Cleaning

Part One #### A) Are the missing values grouped together?: Would dropping them under represent a certain area? The dataset contains multiple features which explain the location of a store: Address, City, Zip Code, County, and GPS location. Multiple of these have over 80,000 missing values, with GPS data missing over 2 million or around 11%.

I started by cleaning the zip code column. Which had missing values, miss labeled data, and incorrect data types. Once the zip codes were ready to be used as reference. Entries with missing GPS data were grouped by zip code.

The missing GPS data was correlated with location. Store Location GPS information was dropped from the dataset.

B) Drop unneeded columns: Invoice number, pack size, volume sold in gallons

Invoice and pack columns are important for the distributers with warehouse inventory, not for store habits. store sell the individual items not cases most often. Volume in gallons is redundant. We also have volume in liters.

C) Correct data types:

Convert numbers to int or float. Date to Datetime

D) Fix mismatched entries.:

Store Name, County, City, Liquor type all had matching issues. A function was defined to correct mismatched entries based on their most popular value.

Correct spelling inconsistencies. Some values were in all caps, others were capitalized, some not at all.

I mapped the correct city spellings to fix replacements.

Liquor Category was compressed to represent slightly broader categories

Part Two

E) Missing and incorrect values: Missing values were handled in a few ways. In the case of GPS I just dropped the column. When the missing values were only a few, I attempted to use other data values to discern the real, intended value.

There were also a bunch of incorrect entries. Bottle price, bottle volume, bottle cost, and bottles sold all contained questionable entries that were dealt with by comparing them with similiar entries. This includes google searches, using other columns and math such as bottle price * bottles sold = sale in dollars

1.4 3. Exploratory Data Analysis

EDA Part 1

A) Feature Engineering: I created columns for profit/ item, profit/ invoice, profit/ ml, profit/ invoice/ liter, and retail price/ ml.

Then I created a feature called 'bottle price category' which is a combination of price/ ml and bottle price. The idea behind using this combination was that: if you buy a large quantity of cheap alcohol the price is still high, and needs to be omitted due to price/ ml. If you buy a tiny bottle of expensive alcohol is it still expensive?

Using the two features: price/ ml and bottle price. I found the mean and std of those features and decided that:

Very expensive: 2 standard deviations above the mean in both categories Expensive: 1 standard deviation above the mean in both categories Inexpensive: below the mean in both categories Normal: anything that did not fit into those categories

B) Feature Exploration The features in this dataset have a lot of different values. I chose to explore features by grouping features into tables.

Whisky has the most unique stores at: 425 Of ubiquitous categories tequila has the highest average retail price Mezcal is only represented in 87 different stores, while has 30 avg per bottle. Scotch has avg is 35.95 and is in 332 stores

Insights like these lead me to choose mezcal. I needed to select a product that is not in every store.

EDA Part 2

1.5 4. Pre Processing

Pre Processing

A) Feature Engineering I expanded the date column creating features for year, month, day.

I created a ranking system for mezcal sales. I ranked stores twice. Once was based on the variety of mezcal products offered and again by number of mezcal bottles sold.

I took these rankings and multiplied them together in order to penalize poor performance.

B) Liquor Search Function Next I created a liquor search function. The function takes in an input of the liquor you are hoping to buy or searching for and returns the top ten ranked stores and their address for that liquor type.

C) Encoding One Hot Encoded 'City', 'County', 'Zip Code', 'Category Name'.

Ordinal Encoding for 'Bottle Price Category' and 'mezcal ranking'

D) Scale reduction I removed the rows for mezcal invoices.

At this point I needed to decrease the size of my data so the computer could handle fitting a model. Ultimately I grouped by Store Name, Bottle Price Category, Category Name, and the one hot features. then aggregated the rest of the columns by either sum or mean.

E) Scaling Now the dataframe is the right shape. I used standard scaler to scale the continuous categories

1.6 5. Modeling

Modeling Part 1

A) Unsupervised learning. Now we have our dataset trimmed down to our area of interest, the mezcal rows have been removed, our stores have been ranked, and features encoded. Our data is currently 225,053 rows and 95 columns.

I performed both PCA and TruncatedSVD. The reason for both was truncatedsvd should perform better on sparse data like we have after using One Hot Encoding.

Top image is PCA, Bottom image is TruncatedSVD Explained variance ratio [0.27899861 0.24220742 0.1141782 0.08882338 0.0545745 0.02389255 0.01397982 0.00872217 0.00821632 0.00753246]

Explained variance ratio [0.27895604 0.24167532 0.11364992 0.08866879 0.02675612 0.05005119 0.01445442 0.00875513 0.00827792 0.00759433]

Having similiar results I moved forward with truncatedSVD

Birch hierarchical clustering was selected for it's ability to handle large datasets. After Adjusting the hyper parameters: n_clusters and branching_factor. Ultimately 100 clusters and branch factor of 15 was chosen.

Here is a graphical representation of the clusters. Reminder there are 100 different clusters so coloring has some overlaps

These are sample counts for each cluster. The parameters of 100 clusters and branch factor of 15 gave the most even distribution. less clusters and different branch factors lead to one giant cluster. cluster 30 is still dominate, but dramatically less so. [5 1995] [9 70] [11 578] [13 3282] [14 34] [15 74] [19 176] [20 1058] [21 1680] [22 159] [24 116] [25 92] [26 42] [28 99] [30 164339] [36 64] [39 57] [40 27] [41 59] [42 37] [43 20] [44 4369] [47 401] [52 4073] [53 6331] [56 344] [57 82] [58 19] [60 3291] [61 267] [68 1090] [76 28] [77 112] [78 168] [79 29581] [81 129] [89 146] [99 54]

After clustering I examined how stores were treated. I grouped by store name, and applied the mean, min, max, mode, standard deviation, variance, and skew to the cluster attribute.

Most stores were placed in multiple clusters. A few were all the same cluster. These tended to be general stores/ gas stations

A trend emerged. Stores with higher (better) mezcal rankings had higher variance in cluster distribution.

I stored the variance in clusters as a feature for classification to consider.

B) Supervised learning Modeling Part 2

Train Test Split was applied to our dataframe with 20% test set. And then our new feature variance was scaled with standardscaler.

Currently our ranking is listed as a continuous variable. I tested random forest regressor on the data to predict mezcal rankings for stores. The model was 'perfect' r-squared was 0.9994556783414396. The model was too good to use. It didn't offer any error of stores to approach for sales.

Next I created a column for mezcal ranking classification. Store rankings ranged from 2-66049. Rankings between 2-1500 became labeled 2, 1500-66068 became 1, and the non-mezcal selling stores ranked 66049 became 0.

A Random forest classifier was created and run through Repeated Stratified KFold cross validation. The result was: Average accuracy: 0.9792252173323898 Average STD: 0.001042200508734727

1.7 6. Results

Here are the results from tuning the random forest classifier

The initial model was rather good. However I needed to cater my model to fit my needs. Originally I set out to make my model have a worse fit. I wanted to create a model that was still accurate enough to be trustworthy, but also miss label enough stores from the 0 and 1 categories in a pro-mezcal or mezcal leaning favoritism.

Ultimately I chose the model which had the best r-squared or best fit. (class weighting 0:20,1:1,2:1) The accuracy across all models was rather consistent and precision/ recall/ f1 did not vary much either.

The predicted labels were stored as a column. We again grouped by store name, this time filtering for stores who currently fit into: label 0 (no mezcal) and predicted NOT 0.

This gave us a list of stores who do not currently sell mezcal, but contain characteristics such that the model thought they would be mezcal sellers.

The same grouping was done for those labeled 1, but predicted as 2.

The highest scoring model provided us with 53 stores to approach about beginning to sell mezcal and 16 stores who the model felt should be selling a greater variety/ or more.

If those stores adopt mezcal as a product, we also have results from the 'worst' or most aggressive model. The most aggressive model gave us an additional 32 stores who are not yet selling mezcal and 12 stores who could expand their offerings.

Included in the dataset is the store name and address of these locations around the greater Des Moines area.

The next course of action would be for distributors who carry mezcal products, to send their liquor represenatives to these stores.

1.8 7. Taking it further

Originally I hoped to include the entire state. That dataset was too large for my computer to run EDA let alone a machine learning algorithm. If I were to choose another alcohol type for exploration it would have been the RTD/ cocktail group. This stands for Ready to Drink cocktails. This category is also increasing in popularity as their availability increases.

The Iowa database website also has census information. If you were trying to specifically target cohorts, you could run statistical analysis in an attempt to find out who is really drinking your product.

1.9 8. Acknowledgements

Thank you for reading.

Thank you to Iowa for providing the data.

Thank you to Ricardo Alanis, for mentoring me through this project. They had to manage my expectations when kernals died, and models told me I didn't have TiB of memory.

Thank you to Aaron D'Souza who also gave me advice along the way.

[ ]: