Drinking with Iowans: An exploration of predictive and interpretive regression models on Iowa liquor sales data

Markelle Kelly
California Polytechnic State University, San Luis Obispo
mkelly23@calpoly.edu

David Lutze
California Polytechnic State University, San Luis Obispo
dlutze@calpoly.edu

Nathan Philliber
California Polytechnic State University, San Luis Obispo
nphillib@calpoly.edu

Abstract. In order to draw conclusions from data, statisticians creating models can choose to prioritize either accuracy or interpretability. While higher-accuracy predictive models fit the data more closely and predict new observations more effectively, interpretable models boast simpler, more straightforward conclusions. To explore the differences between these two strategies, we develop two models of Iowa liquor sales volume per capita, one aiming for the highest accuracy and one prioritizing interpretability. We then compare these models, noting that each has its strengths and weaknesses and is worth considering when modeling a new dataset. For the purposes of the Iowa liquor sales data, we choose our more accurate predictive model as superior.

Introduction

As outlined in Leo Breiman's paper "Statistical Modeling: The Two Cultures," statisticians have historically focused on models that can be easily interpreted, even at the cost of lower accuracy. While it makes sense to choose models from which we can obtain actionable conclusions, Breiman argues that this information has little meaning when pulled from a model that does not closely match the data. Therefore, he recommends that statisticians consider a more accurate predictive approach when modeling datasets, despite the increased complexity of results.

With Breiman's paper in mind, we examine the differences between these two strategies in the context of data on liquor sales in Iowa. Our goal is to predict the volume of alcohol purchases of Iowa liquor licensees per capita using information about each sale's contents, vendor and buyer, and geographic location. We construct two different models of purchased liquor volume per capita—one by prioritizing interpretability, the other, accuracy. Finally, we evaluate and compare both models in terms of predictive accuracy and the quality of conclusions that may be drawn. Through the process of analyzing this data and developing our models, we gain insight about the tradeoff between interpretable and predictive modeling techniques as well as the importance of feature engineering.

Dataset Creation

Data Cleaning

Before any analysis can be performed, the original dataset requires some cleaning and processing. One concern is the presence of null values. Specifically, since the data will be aggregated by county, any observations for which the county could not be determined were unusable and thus were dropped. This applied to less than 1% of the dataset, so data loss is not a critical issue.

To effectively predict sales volume per capita, it was also necessary to aggregate the data. We aggregate biweekly (in two-week periods) and by county. Biweekly aggregation is desirable because we aim to predict at a relatively high level of granularity, but need to balance this with manageability of the dataset. Furthermore, we aggregate by county in order to obtain more meaningful conclusions than grouping by zip code or geographic coordinates. Differences between counties seem more significant and actionable, especially because regulations impacting liquor purchases and licensing are likely to differ by county rather than zip code.

Feature Engineering

After cleaning, we extract useful information from the original dataset with the addition of several new features. Among these were, for each county and two-week period, the number of unique purchasers of alcohol, the total retail cost of purchases, the number of weeks since the start date of the dataset, and the average bottle price.

To understand what types of alcohol were being purchased, we divide observations into the groups “vodka”, “tequila”, “whiskey”, “rum”, “gin”, and “other”. The group an observation belongs to is determined by searching for key words in the category name field. We then aggregate the number of bottles sold from each category for each county and time period.

Although there are hundreds of unique vendors in the dataset (that is, it would be impractical to include each individual vendor in a model) we retain some useful details about the vendor of each purchase. First, we categorize “big vendors” as vendors with over 1000 entries in the full dataset. This list of 12 vendors is comprised of large companies like Bacardi, Diageo Americas, E&J Gallo, and Sazerac. Many vendors are listed under several different spellings, so we perform some additional cleaning to group all variations as one vendor. We also divide vendors by their average volume of sales, categorized either as “big volume vendors” or not. During aggregation, the average value of these variables is taken to get the proportion of “big” and “big volume” vendors across purchases.

Instead of including dummy variables for all counties in the dataset, we separate counties into five groups by their average volume of sales over the full dataset. The final group includes only one county, Dickinson, which has significantly higher per capita volume than all other counties. These groups are then dummified, allowing us to avoid using county itself as a feature.

Finally, we include the season (fall, winter, spring, or summer) of the two-week period as a dummy variable. This enables our model to reflect seasonal changes in liquor purchases. Date itself is not included as a feature; only the number of weeks since the start and the season of the year are used to represent time.

Use of External Data

In addition to extracting variables from the original dataset, we employ a few external data

sources, most notably Iowa's 2010 census data for each county. From this dataset, we obtain population and land area, which are used to compute population density for each county and transform volume of liquor purchased into volume per capita. Demographic variables, especially information on age, gender, household makeup, and race are also extracted. The populations by age are transformed into proportions for ages thought to be particularly relevant, including children under five, teenagers and people in their twenties, people over 21, and people over 65. The proportion of the population over 21 and median age are also divided by gender. Further, we compute proportions of people in each county who are white, who are the head of their household, and who have children under 18. Lastly, additional variables are added for average household size and whether housing units are rented or owned. These variables are included in the dataset because we expect that populations with different demographic compositions purchase different quantities of alcohol.

A dataset containing records of car accidents in Iowa is also used to supplement our feature engineering. We extract the number of accidents in which alcohol was involved for each county and time period. It seems likely that an area's alcohol purchases over a certain time period are correlated with drunk driving accidents during that period.

Finally, a simple dataset from pandas that contains federal holidays is used to determine whether each two-week period includes any days people have off from work. Since it makes sense that people might buy more alcohol if they have a long weekend to drink it, this binary variable is also included in our dataset.

Final Shape

After data cleaning and feature engineering, including the incorporation of external data, we arrived at a final dataset with 17,208 observations and 42 different columns.

Linear Regression Modeling

To model this dataset, we implement linear regression, estimating coefficients using the traditional closed form solution for ß. This formula, which involves matrix inversion, was practical for this project because of the nature of our dataset, which is relatively small in size, does not have multicollinearity issues, and has more rows than columns. If any of these characteristics were missing, an alternative implementation such as QR-decomposition or gradient descent would have been necessary.

Model Evaluation

Interpretable Model

To develop the most interpretable model, we implement forward-stepwise variable selection based on the minimization of BIC. Starting with no variables in the model, the feature that improves the BIC the most is added, and this process is continued until the BIC ceases to drop significantly. For the sake of simpler coefficient interpretations, our two sets of dummy variables are not considered in this feature selection. To avoid overfitting, the model is checked by running it on a simple 80-20 train-test split, ensuring that our coefficients and metrics are not the result of overfitting.

Furthermore, since this model prioritizes sound conclusions, we check the assumptions of linear regression (normality of residuals, linearity, equal variance, and independence of observations), which do not show reason for concern.

Accurate Model

Similarly, we obtain the most accurate model with forward-stepwise selection that minimizes mean square error. At each step, the variable that lowers the MSE the most is added until it is not possible to improve the MSE any more. Within variable selection, we use five-fold cross validation to ensure that our model actually has improved accuracy for new predictions and is not overfitting the data.

Results

Interpretable Model

Our final interpretable model contains seven features: average vendor size, week number, average household size, bottles of rum sold, county population, proportion of men over 21, and average bottle price. These seven factors are the seven most important variables from our dataset in predicting volume of sales per capita. Adding these features drops our BIC from over 10,000 to 6,557 and achieves an R^2 of 0.623.

Of the variables included, average vendor size, week number, number of bottles of rum sold, and population of county all have a positive coefficient (and thus a positive association with volume per capita). Having more purchases from big-name vendors, taking place later in the timeline of the dataset, including more bottles of rum, and taking place in a larger county are all associated with increases in the volume per capita of purchases made. On the other hand, average household size, proportion of males over 21, and average bottle price are negatively associated with volume of sales per capita. It makes sense that average household size and average bottle size might have this relationship with volume per capita; buying more expensive alcohol or having several children to care for are logical reasons to purchase lower quantities of alcohol. More surprising, however, is that the proportion of men over 21 has a negative association with volume per capita. Perhaps this is evidence of underage drinking, or maybe women are buying more alcohol than we might expect.

The benefit of this model is that we can pinpoint a small number of features that are most important in predicting volume of sales per capita. Because of this, it is more practical to consider the implications of each coefficient individually and develop actionable conclusions about liquor sales.

Accurate Model

Our final predictive model is much more complex than our interpretable model, containing all but five of our variables. The five variables not included are: number of alcohol related accidents, proportion of population aged 15-19, proportion of population aged 20-24, number of gin bottles sold, and proportion of big-volume vendors. This model has an MSE of 0.002 and an R^2 of 0.810.

Based on the variables included, we can conclude that the proportion of the population aged 25-28 is more important for prediction than the proportion of teenagers or those in their early twenties. It appears that the slightly older, potentially more established Iowans are purchasing more alcohol. The number of alcohol-related accidents is also not important in predicting volume of liquor sold per capita, so it seems that those purchasing more alcohol are not necessarily getting into more drunk driving accidents.

Notably, the seven variables selected in the interpretable model are all selected within the first 12 iterations of the forward stepwise selection process for this model. In other words, a similar group of variables comes up as important whether we are prioritizing accuracy

or interpretability.

Comparative Analysis

While the accurate model takes a bit longer to run and provides a more overwhelming final list of features, it reveals many of the same overall patterns in the data while predicting new observations with much higher accuracy. For example, the proportion of men over 21 has a coefficient of -1.48 in the interpretable model and -1.33 in the predictive model, so both models demonstrate that the proportion of men over 21 is negatively associated with liquor sales (at about the same level). Of course, the interpretable model gives us the straightforward summary that, after adjusting for the other variables in the model, "an increase of 1 in the proportion of men over 21 is associated with a decrease of 1.48 liters of alcohol sold per capita," while the dummy variables in the predictive model make the conclusion more complex. However, the underlying idea - having a higher proportion of men over 21 is generally associated with lower per capita liquor sales - is present in both models.

While the interpretable model allows us to create a simple summary of the variables at play, this may oversimplify reality, offering a false illusion that Iowa liquor sales can be thoroughly explained by seven basic factors. With higher accuracy and more predictors, it seems that our predictive model matches the complexities of reality more closely.

Our interpretive and predictive models differ notably in accuracy - the interpretive model achieving an R^2 of only 0.623 and MSE of 0.004 and the predictive model achieving an R^2 of 0.81 and MSE of 0.002.

In general, therefore, we prefer our predictive model. It has a much higher accuracy, increasing our confidence in how closely we are fitting the data and thus in the validity of our conclusions, while still providing us enough information to draw meaningful insight from the data.

General Observations

Maximizing Interpretability

Interpretability is an important consideration in statistical modeling, as it allows people to analyze data in ways that are meaningful to non-statisticians. However, interpretability is quite subjective and more difficult to quantify than other criteria such as accuracy. To create an interpretable model, we aim to use as few features as possible while still maintaining enough accuracy to be useful, a balance that can be challenging to achieve. For this project, interpretability was optimized with BIC, a metric which penalizes models for including more features. It is worth noting, though, that this is not the only metric used to guide model creation for interpretability. Using metrics with stricter or looser penalties for the number of variables will yield a variety of models with different numbers of features. Furthermore, it can be argued that not all variables are equally interpretable (for example, adding dummy variables changes the interpretation of all other variables in the model) and may require more thought on whether they are worth including. Therefore, when building a model with the goal of interpretability, it is important to consider and experiment with different metrics and variables rather than blindly minimizing BIC.

Feature Engineering

We also observe that feature engineering is an incredibly important part of data modeling, with dramatic implications on model selection itself. Our first sets of models were based on a smaller group of original features; these models had many more variables and were less accurate than our final models. Having more features available

in the variable selection process enables more curated subsets of variables, dramatically decreasing the complexity of each model. Overall, then, feature engineering is an important step for both predictive and interpretable models.

Since feature selection removes unimportant variables, it is worth trying out a wide range of variables, whether or not we are confident they might improve the model. For example, we expected the number of alcohol-related accidents to be a useful predictor for volume of liquor sold, but it is one of the five features not included in our predictive model. On the other hand, we did not anticipate the number of weeks since start date being an important feature, but it is included in both models.

Model Exploration

Even though most tasks and projects in data science are centered around the construction of a single final model, it is informative to explore a variety of models first. For the purpose of this project, the model maximizing accuracy is our preferred model, but the creation of the interpretable model helped inform our understanding of it. Based on the interpretable model, we substantiated ideas about which core variables are the most important. Since these appear in our interpretive model with similar coefficients, we were able to draw more valuable intuition from the predictive model. Additionally, testing out a variety of models gives us reference points for metrics like R^2.

Multiple models should also be considered because there are cases in which a more interpretable model would be preferred. If differences in accuracy between models were negligible, it would have made more sense to opt for interpretability, avoiding an unnecessarily complex predictive model. The complexity of the ideal model chosen, then, may give us some insight into the nature of the dataset.

Conclusion

After developing two different models for volume of Iowa liquor sales, maximizing both interpretability and predictive accuracy in turn, and comparing them, we feel we have gained significant understanding of the differences between these two strategies. While we conclude that the model focused on predictive accuracy is superior, this decision was made with the knowledge we gained from our creation of the interpretable model. If we were to move forward, we would iterate on the predictive model, but forming both models was a valuable exercise that provided additional layers of insight into our data.