Which Variables help in predicting supermarket revenue? Evidence from Chicago

Introduction

Location choice is critical to retail organization, due to its effect on supermarket success (Clarkson, Clarke-Hill, and Robinson 1996). To choose the outlet location, previous studies have suggested that the location of supermarkets largely depends on the local demographic profile (Baviera-Puig, Buitrago-Vera, and Escriba-Perez 2016). Our study considers 45 demographic variables and tries to answer the research question: what demographic variables are important for predicting the revenue of supermarkets? We examine which variables have the biggest impact through an elastic net. This model is chosen because it has shown to work better than other regularized regressions, such as the least absolute shrinkage method (LASSO), when several variables are highly correlated (Zou and Hastie 2005). This is the case in our dataset. We assess the best parameters for the model using k-fold cross validation to further prevent overfitting (Friedman, Hastie, and Tibshirani 2001). We find that the % of households with children under 9 years old, % unemployed, and % of households with a mortgage matter most. ## Data Our research works with a dataset of 45 demographic variables related to 77 supermarkets located around the Chicago area from the year of 1996. We define demographic data as data that reflects a profile of the customers; examples of such data included in our data include age, sex, income level, race, employment, homeownership, and level of education. The variables are all continuous. A full description of each variable included can be found in table A of the appendix, and table B includes the summary statistics. Past studies have indicated that these variables likely affect the supermarket turnover. For example, the variable of income has a relation to food consumption in supermarkets (Jones 1997); the variable of gender may influence the choice of supermarket (Beynon, Moutinho, and Veloutsou 2010); and the variable of race composition may have an impact on supermarket location(Lamichhane et al. 2013). Our response variable is yearly total turnover ($). The demographic data were derived from the U.S. government's (1990) census for the Chicago metropolitan area. We scale both the independent and dependent variables as follows:

$$\tilde{\mathbf{X}} = \frac{\mathbf{X} - \mu}{\sqrt{\frac{\sum \mathbf{x^2}}{n-1}}}$$

Where $\tilde{X}$ is the scaled variable, X is the unscaled variable, μ is the mean of $\boldsymbol{X}$, and n denotes the sample size. This scaling is done to ensure a fair interpretation of the model coefficients, as the range of each variable is different. Table C of the appendix shows that several variables are highly correlated. It is to be expected that characteristics such as income, home ownership, age etc. are to a large extent correlated.

Method

In short, our method is using an elastic net, which deals with overfitting by using a weighted average of the penalty terms applied in the LASSO and ridge regression. Before we explain the elastic net, let's consider the standard linear regression model:

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

In which $\hat{\boldsymbol{y}}$ is an n x 1 column vector of the predicted response variable, $\mathbf{X}$ is an n x (p+1) matrix with n observations of p predictor variables and the intercept. $\hat{\boldsymbol{\beta}}$ is an (1 + p) x 1 column vector of the estimated

coefficients for true parameter $\boldsymbol{\beta}$ of the intercept and p predictor variables. The standard method for finding the optimal $\boldsymbol{\beta}$ is the ordinary least squared (OLS) method that minimizes the sum of squared error between the predicted and observed response variables $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$. Thus the following loss function is minimized:

$$L(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).$$

The problem with using this method is that it is prone to overfitting (Friedman, Hastie, and Tibshirani 2001). This is because when applied to a training set, it often overestimates the effect of certain variables. To solve this problem, regularized regression methods can be used. These apply penalty terms to the above loss function in order to shrink the $\boldsymbol{\beta}$. Elastic net is a regularized regression that combines two penalty terms:

$$L_1(\boldsymbol{\beta}) = \lambda \sum_{i=1}^{p} |\beta_i| \qquad L_2(\boldsymbol{\beta}) = \lambda.\boldsymbol{\beta}^T\boldsymbol{\beta}$$

In which L_1, L_2 are the two penalty terms, and λ is a constant parameter that determines their size. The P_1 is the sum of the absolute value of the coefficients, and is used in the so-called LASSO method (Tibshirani 1996). When this penalty term is used, coefficients are continuously reduced, or removed completely. But using just this penalty term has several downsides; it performs less well when $p \geq n$, and when several variables are highly correlated (Zou and Hastie 2005) . Zhou and Hastie (2005) show that in these circumstances, the elastic net performs better than the Lasso, by adding a second penalty term, P_2. This is the same penalty used in a ridge regression (Hoerl and Kennard 1970). Elastic net then determines the weight between the two penalty terms, using the constant parameter α. Given the high correlation between several of the variables in our dataset, it seems appropriate to use the elastic net. This gives the following loss function:

$$L(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\alpha \sum_{i=1}^{p} |\beta_i| + \lambda(1-\alpha)\boldsymbol{\beta}^T\boldsymbol{\beta}$$

In order to find the $\boldsymbol{\beta}$ at which $L(\boldsymbol{\beta})$ is minimized, we use the majorize-minimization algorithm (MM). We use the following majorization function:

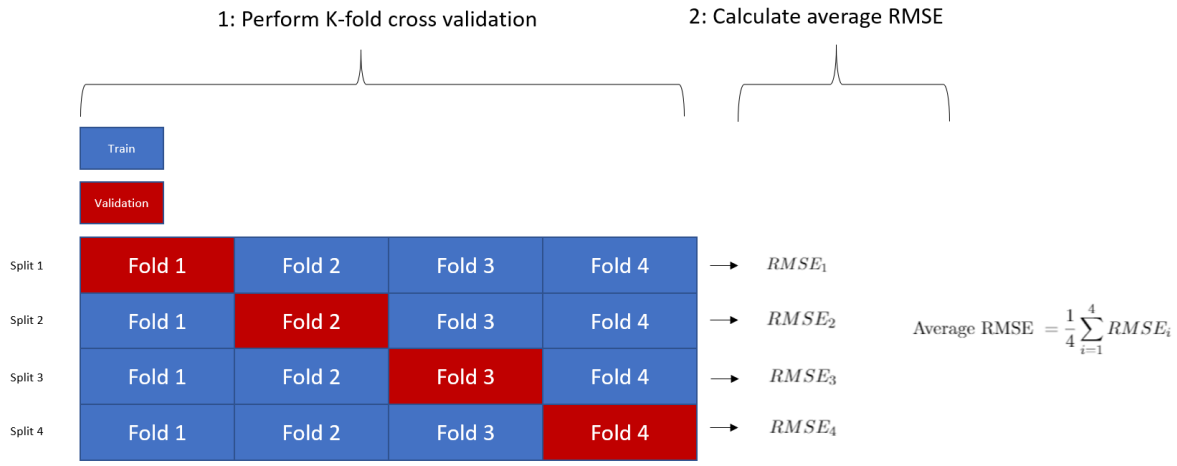
$$L(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^T(\mathbf{A})\boldsymbol{\beta} - n^{-1}\boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{y} + c$$

$$\mathbf{A} = n^{-1}\mathbf{X}^T\mathbf{X} + \lambda(1+\alpha)I + \lambda\alpha\mathbf{D} \qquad \mathbf{D} = \begin{bmatrix} \frac{1}{max(\beta_1,\epsilon)} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{max(\beta_p,\epsilon)} \end{bmatrix} \qquad c = \frac{1}{2n}\boldsymbol{y}^T\boldsymbol{y} + \frac{1}{2}\lambda\alpha\sum_{i=1}^{p}|\beta_i|$$

We then find the $\boldsymbol{\beta}$ for which this function is minimized through stepwise updating the $\boldsymbol{\beta}$. This happens by solving the following: $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} = (\mathbf{A})^{-1}n^{-1}\mathbf{X}^T\boldsymbol{y}$. Here, $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ is the estimated $\boldsymbol{\beta}$ at step k. This stepwise updating continues until the next set of $\boldsymbol{\beta}$ does not improve the loss functionby more than ϵ. To find the λ and α for our elastic net, we use K-fold cross validation. In this method, the dataset is split into K random samples. One of the samples is picked as the validation set, and the $\boldsymbol{\beta}$ are determined based on the remaining K-1 samples. This then happens K times, until each of the samples is used as a validation set. For each iteration, we calculate the root mean squared error (RMSE), and then the average RMSE across the K validations.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{y}} - \boldsymbol{y})^2} \qquad \text{Average RMSE} = \frac{1}{K}\sum_{i=1}^{K} RMSE_i$$

In which $RMSE_i$ is the RMSE of the validation on sample i. Our metric for picking the λ and α is the lowest average RMSE. The process of finding this metric process is visualized below for K = 4.



Using K-fold cross validation to pick the ideal λ, α reduces the likelihood of overfitting, since the parameters are not just based on one random sample, but on multiple, reducing the likelihood that features from one sample dominate when predicting. We fold our sample 10 times - this allows us to use a lot of different folds, without becoming computationally too expensive. We use RMSE to measure the error between our prediction and the actual values. The squaring of these errors ensures that larger errors contribute more to the loss function, which decreases the chance of our model making large mistakes. When searching for λ, α, we evaluate all combinations between two sets. The first set containing all possible α values is defined as follows: $\alpha = \{0.1, 0.2, ..., 1\}$. The set containing the λ values is $\lambda = \{10^{x_1}, 10^{x_2}, ..., 10^{x_{50}}\}$ with x increasing from -2 to 10 in 50 steps.

Results

By minimizing the RMSE of all α and λ combinations, we find that an α of 0.2 and λ of 0.1 produces the best fitted model. This indicates that in our problem more emphasis should be put on the L_2 penalty. This result is consistent with the findings in (Marquardt and Snee 1975), which found that in problems with highly correlated explanatory variables ridge regression performs best.

To answer our research question on what variables are most important for the prediction of supermarket turnover we analyse the estimated coefficients obtained by training on the scaled dataset. These are found in table . In this table only the coefficients with an absolute value higher than 0.01 are displayed, because any coefficients below this threshold are at least an order of magnitude smaller and thus have extremely little effect on our dependent variable. The coefficients represent change in our standardised dependent variable when the independent variable changes with one standard deviation. Thus the absolute size of the coefficient can be interpreted as the contribution of the variable for the prediction. From table we find that the three most influential variables are the % of households with children under nine years old, % unemployed and percentage of households with a mortgage. Here the first two are positive in their influence on supermarket turnover and the last one is negative.

Baviera-Puig, Amparo, Juan Buitrago-Vera, and Carmen Escriba-Perez. 2016. "Geomarketing Models in Supermarket Location Strategies." *Journal of Business Economics and Management* 17 (6): 1205–21.

Beynon, Malcolm J, Luiz Moutinho, and Cleopatra Veloutsou. 2010. "Gender Differences in Supermarket Choice." *European Journal of Marketing*.

Clarkson, Richard M, Colin M Clarke-Hill, and Terry Robinson. 1996. "UK Supermarket Location Assessment." *International Journal of Retail & Distribution Management*.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Vol. 1. 10. Springer series in statistics New York.

Hoerl, Arthur E, and Robert W Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12 (1): 55–67.

Jones, Eugene. 1997. "An Analysis of Consumer Food Shopping Behavior Using Supermarket Scanner Data: Differences by Income and Location." *American Journal of Agricultural Economics* 79 (5): 1437–43.

Lamichhane, Archana P, Joshua Warren, Robin Puett, Dwayne E Porter, Matteo Bottai, Elizabeth J Mayer-Davis, and Angela D Liese. 2013. "Spatial Patterning of Supermarkets and Fast Food Outlets with Respect to Neighborhood Characteristics." *Health & Place* 23: 157–64.

Marquardt, Donald W, and Ronald D Snee. 1975. "Ridge Regression in Practice." *The American Statistician* 29 (1): 3–20.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–20.