Real Estate Site Selection With Predictive Modeling in the Open-Source R Language

Case History

Category: *Retail*

Methods: *R Language, Predictive Modeling, Decision Tree, Linear Regression, Random Forest™, Real Estate Selection*

Summary

We explored the use of the open-source R Language to develop alternative types of predictive models for retail-site sales forecasting. Alternative model types included decision tree, linear regression, and Random Forest™. The study found that a straightforward linear regression—when developed using best-practice, cross-validation techniques to select variables from a large variety of store, customer, and trade-area predictor variables—can be developed rapidly and cost-effectively using open-source, free software.

After using the R Language for the entire process (from data processing to predictive modeling, cross-validation, GIS mapping, and linking to Google Maps), we found that the R Language is a powerful and cost-effective tool for predicting retail-site sales, with the only drawback being the ramp-up time required to learn to use a script-based matrix language and syntax.

Strategic Issues

Within the field of retail sales forecasting, a common approach today is an analog one, where trade area and performance data for similar retail locations are summarized and then used as an input to a judgmental evaluation of the sales opportunity for a proposed retail site. Another common approach is to apply a "gravity" model that employs the use of a spatial perspective on the store and trade-area characteristics (i.e., retail square footage, distance, and drive times) to estimate sales.

This case study explores the value and feasibility of using predictive models developed using best-practice, cross-validation techniques as a tool to be used in the process of retail-sales forecasting. The potential value of predictive modeling lies in its ability to harness the information contained within a large variety of data fused together, with the trade area as the linking concept.

Would sophisticated predictive modeling methods (e.g., linear regression or Random Forest™ with cross-validation for variable selection) deliver better predictions of retail sales for potential sites

Decision Analyst
strategic research ■ analytics ■ modeling ■ optimization

604 Avenue H East • Arlington, TX 76011-3100, USA
1.817.640.6166 or 1.800.ANALYSIS • www.decisionanalyst.com

Copyright © 2016 Decision Analyst. All rights reserved.

versus calculation of average sales based on a few predictor variables chosen by logic or intuition (decision tree model)? Would such sophisticated models require too costly a software investment or too long a time period for analysis, making the approach undesirable? To answer this latter question, we explored the use of the R Language open-source software for performing all of the analysis entailed.

Research Objectives

The purpose of this research was to determine the level of predictive accuracy that can be achieved with best-practice, predictive models that use a large variety of data variables, including store sales, store attributes such as square footage and staffing, trade-area characteristics such as census-derived sociodemographics, location of competitor and sister stores, customer surveys, and economic data such as employment rates. Within the context of this large variety of data, we set out to:

- Measure the relative predictive accuracy of alternative predictive models of varying complexity and the value of using cross-validation techniques for variable selection.
- Gauge the difficulty of implementing complex predictive models using the R Language, an open-source, free software.
- Explore the use of the R Language to link with Google Maps for the purpose of effectively communicating model results.

Research Design and Methods

A store and trade-area data set was built that included 167 potential predictor variables for approximately 1,000 retail stores for 54 weeks. Data reduction functions available within the R Language eliminated variables that were linear combinations or highly correlated with other variables or had variability too low to make them useful as predictors of store sales. The final reduced data contained 43 potential predictor variables.

The final data set was randomly split into training data and validation data, and the validation data was set aside and not used for model development. This action was taken to ensure a solid measure of predictive accuracy based on applying the predictive models to other data not used in model development.

Using the training data, cross-validation techniques were applied with two different models (Random Forest™ and linear regression)

Copyright © 2016 Decision Analyst. All rights reserved.

to pick both the number of variables and the particular variables that maximize predictive accuracy. The models were then tested for predictive accuracy using the validation data, and their predictive accuracy was compared to a simple decision tree model that used three preselected predictor variables based on logic and intuition.

The R Language was then used to plot the validation data predictions on a geographic map of the U.S. depicting state boundaries. Then the top-ten retail sites within the validation data were identified and plotted using an R program that interfaced with Google Maps. The delivery platform was a browser page that included a dashboard table and the Google Map.

Results

The Random Forest™ model with 20 variables had the greatest predictive accuracy within the validation data (R-squared statistic of 71%). The linear regression with 20 predictor variables performed almost as well with an R-squared statistic of 66%. The decision tree model with three predictor variables had a near-zero R-squared statistic, suggesting that best-practice, cross-validation techniques used for variable selection would greatly outperform a simple model based on logical or intuitive selection of predictor variables.

R Language packages were found to implement sophisticated and complex data reduction and cross-validation within the training data and to link model predictions to Google Maps for easy visualization of results.

604 Avenue H East • Arlington, TX 76011-3100, USA
1.817.640.6166 or 1.800.ANALYSIS • www.decisionanalyst.com

Copyright © 2016 Decision Analyst. All rights reserved.