Stat E-139 Project Proposal

Stephen Camera-Murray, Desirée Koh,
Jennifer Le Hégaret, Elizabeth Wilson-Milne

November 29, 2015

Overview

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance in order to schedule staff effectively. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied. To address this challenge, Rossmann decided to run a competition on Kaggle. For our class project, we have formed a team to compete in this contest.

The Data

The Rossmann data consists of two files. The first file contains store-level data such as:

- Type of store (in terms of its "model" and its overall selection)
- Information about the closest competitor (distance and longetivity)
- Information about that store's particular promotion cycle

The second file contains daily data for each store, such as:

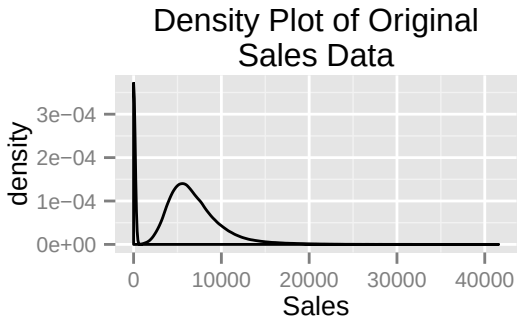
- date (also noting that date's exact weekday, holiday status, and school holiday status)
- whether or not that particular store was open that particular day
- whether or not the store was running a promotion that day
- sales amount and number of customers

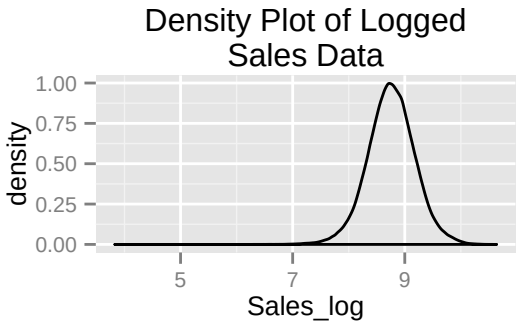
The Plan

1. Stitch together the two data files so we have full store information for each sales data point.
2. Analyze all variables to check for normality.
3. Add additional derived variables that may help with prediction, such as season and weekend dummy variables.
4. Check for multicolinearity to ensure we do not have any highly correlated variables.
5. Establish a baseline by performing a naive regression to ensure that our final model is an improvement upon the baseline.
6. Perform any transformations to meet the assumptions of regression.
7. Perform stepwise regressions until we have a final model that improves upon the baseline.
8. Collect our prize money!

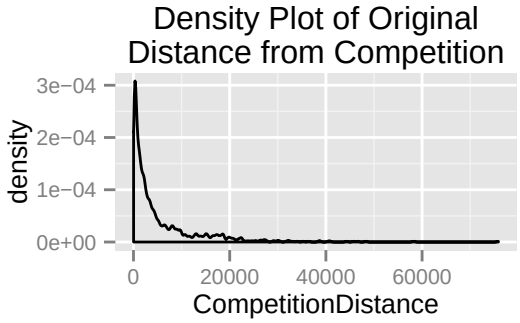
The Progress

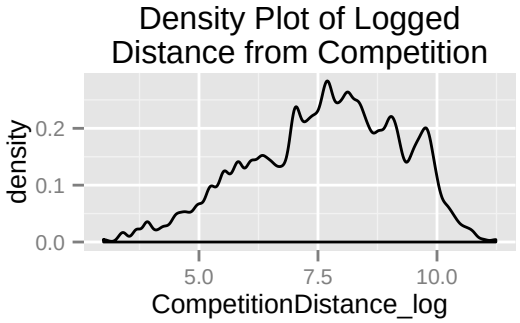
We have already merged our two data files and investigated the normality of the variables within the given data. While some of the sampling distributions were far from normal to start, some log transformations attained remarkably bell-shaped curves, such as for the Sales figures:





while other sample distributions stay rather interesting looking even after log transformation:





We look forward to further explorations of the data!

Challenges so far

R coding syntax can be tricky when manipulating string variables. Our current challege is to determine whether or not a second promotion is active by comparing a given date with the list of months comprising a particular store's promotion interval.

Milestones

Week	Goal
11/30 - 12/4	Explore and supplement data in R
12/5 - 12/11	Refine model in R
12/12 - 12/16	Write up results in .rmd
12/17 - 12/19	Final edits and submission