Predicting Rossmann Store Sales

Sophie Guo, Kevin Legein, Arthur Shikhaleev, Justin Zhang

Duke University

Abstract

Predicting sales is important for store managers in many aspects, such as price stability, inventory controls, demand forecasting, supply chain management and marketing. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation.

Rossmann, Germany's second-largest drug store, operates over 3,000 drug stores in 7 European countries. Their store managers are tasked with predicting their daily sales for up to six weeks in advance. 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. In this dataset, we have 1,115 stores located across Germany. The goal is to help Rossmann create a robust prediction model for sales.

Data Summary

The variables of interests:

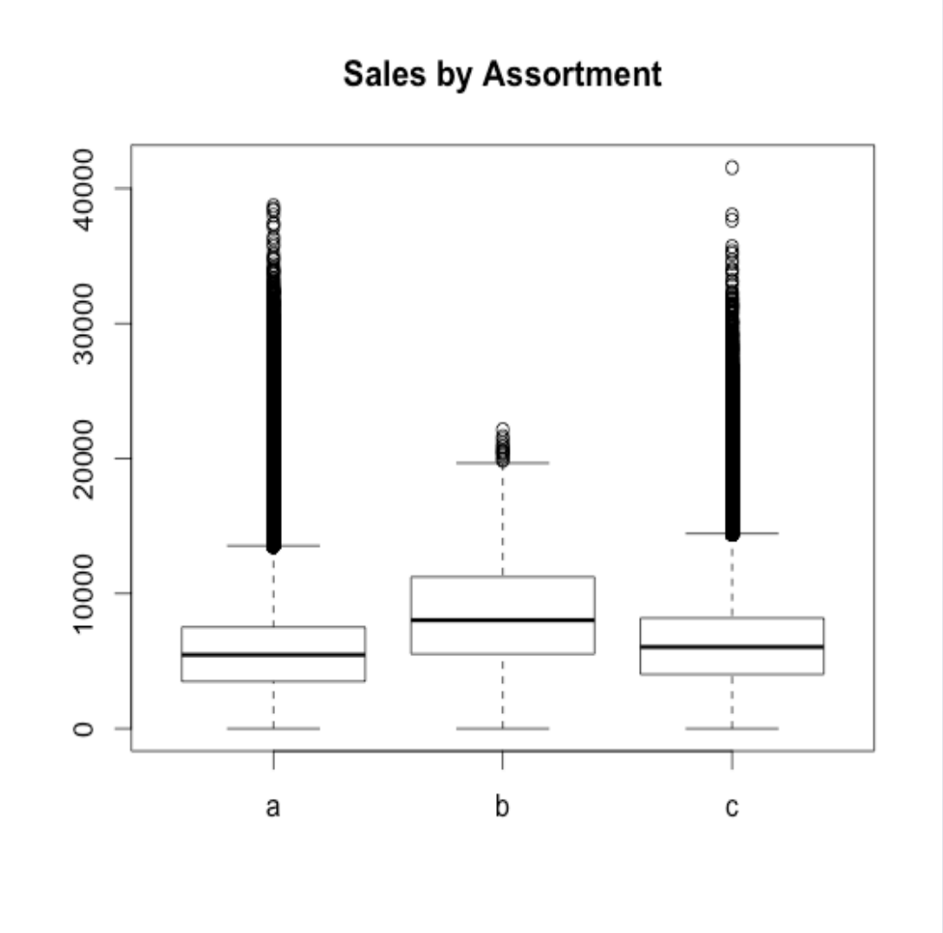
- Day of the week
- Number of customers on a given day
- Whether the store was open on that day
- Whether the day was a state holiday
- Whether the day was a school holiday
- Store type
- Assortment level
- Distance to the nearest competitor store
- The date on which the nearest competitor store was opened
- Whether the store is running a promotion
- Whether the store is participating a continuing and consecutive promotion
- The date on which the store started participating in promotion
- The consecutive intervals promotion is started every year

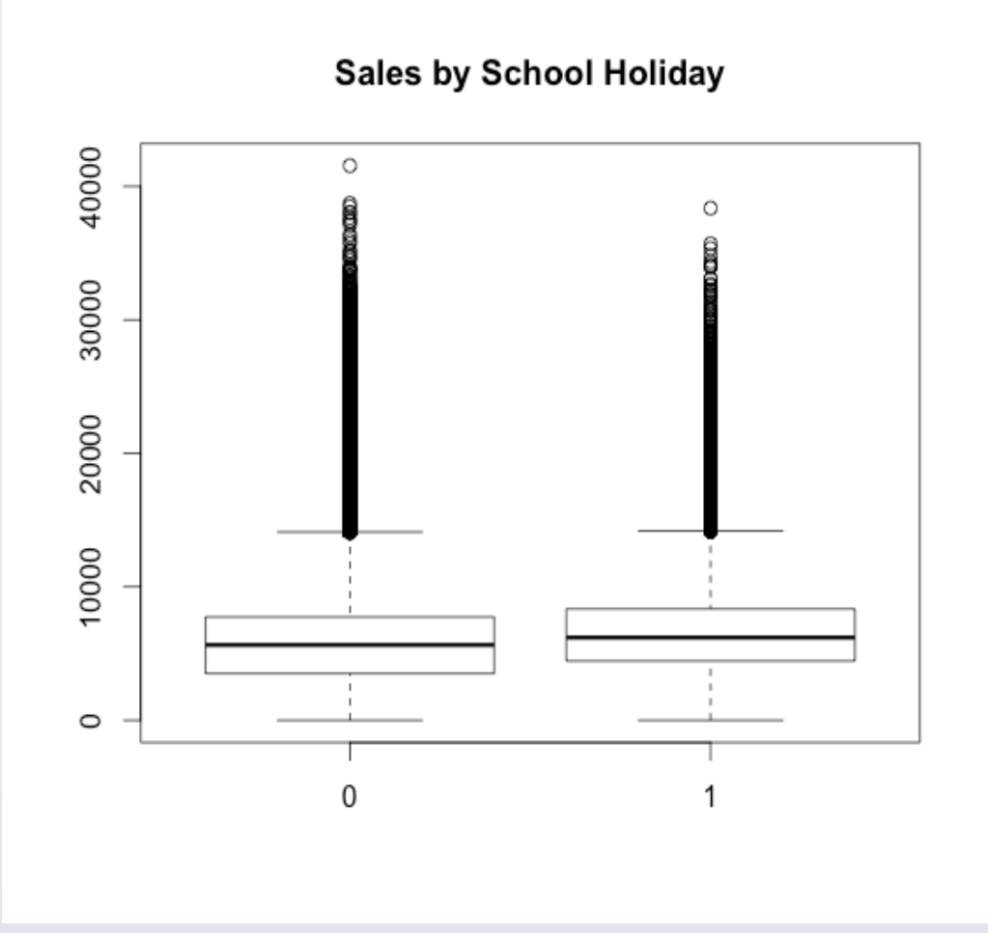
The main question:

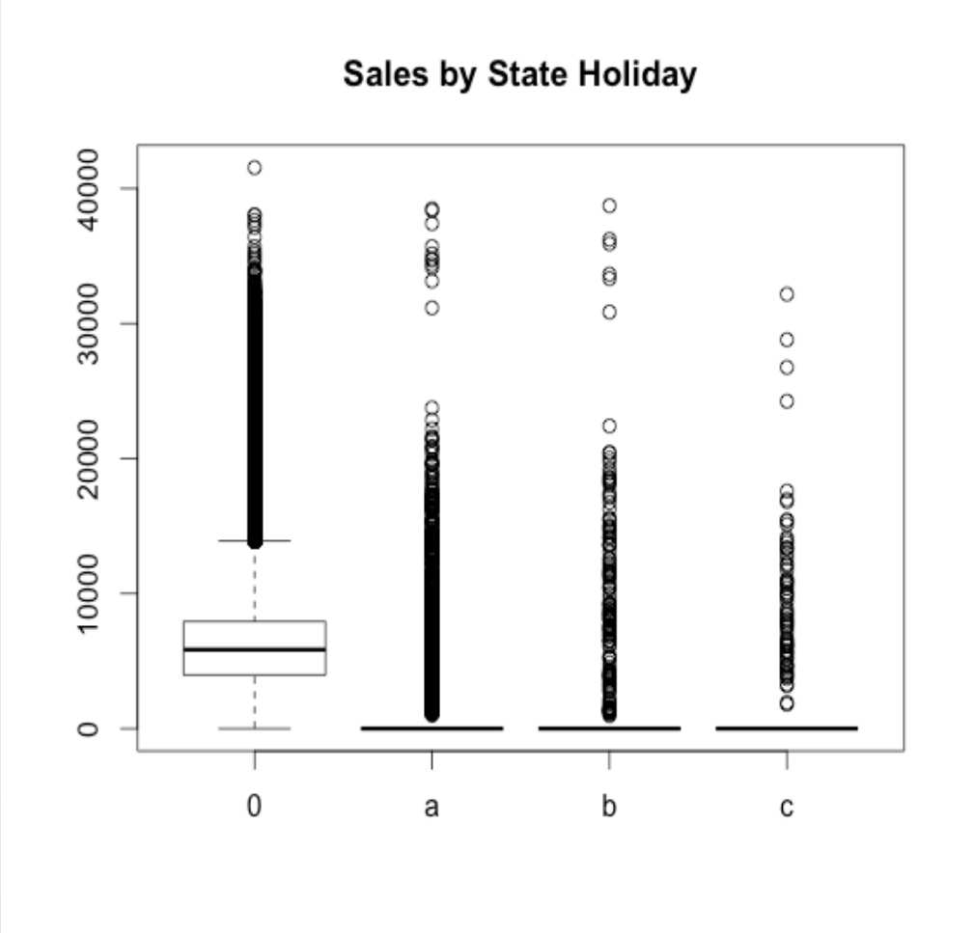
- Is there a way to categorize a store as a failing enterprise, an average one, or a successful one; *i.e.*, tell the company whether a given store should remain open?

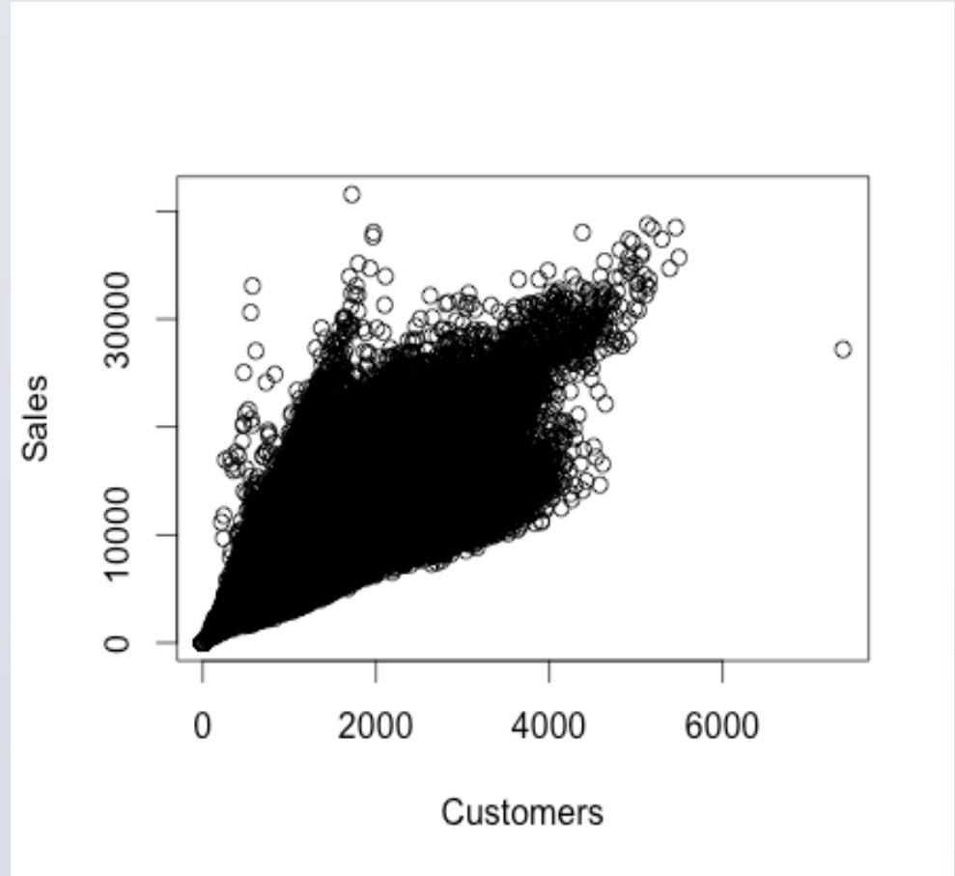
Exploratory Data Analysis

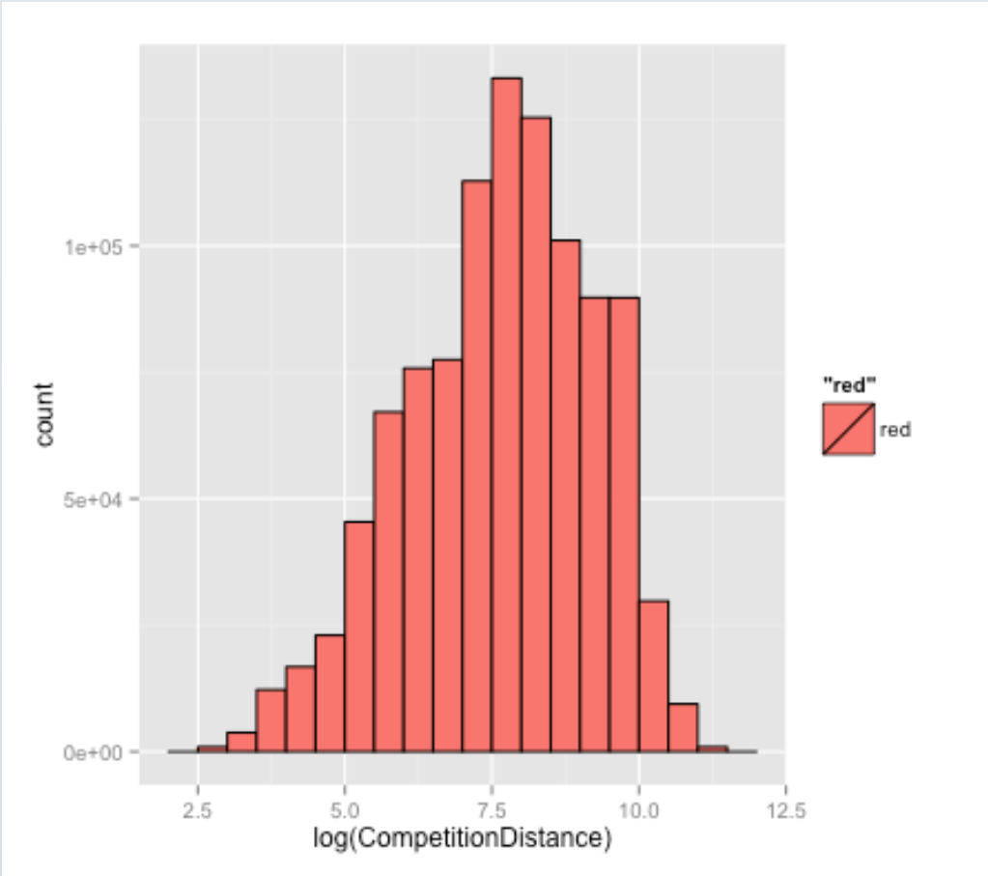
- Sales are highly correlated with Number of Customers
- Sales vary significantly by State Holiday and Assortment
- Sales does not seem to change much by School Holiday
- Log of competition distance and sales are roughly normally distributed
- Cross Tabulations show possible interactions among Promo, State Holiday, School Holiday, and Open

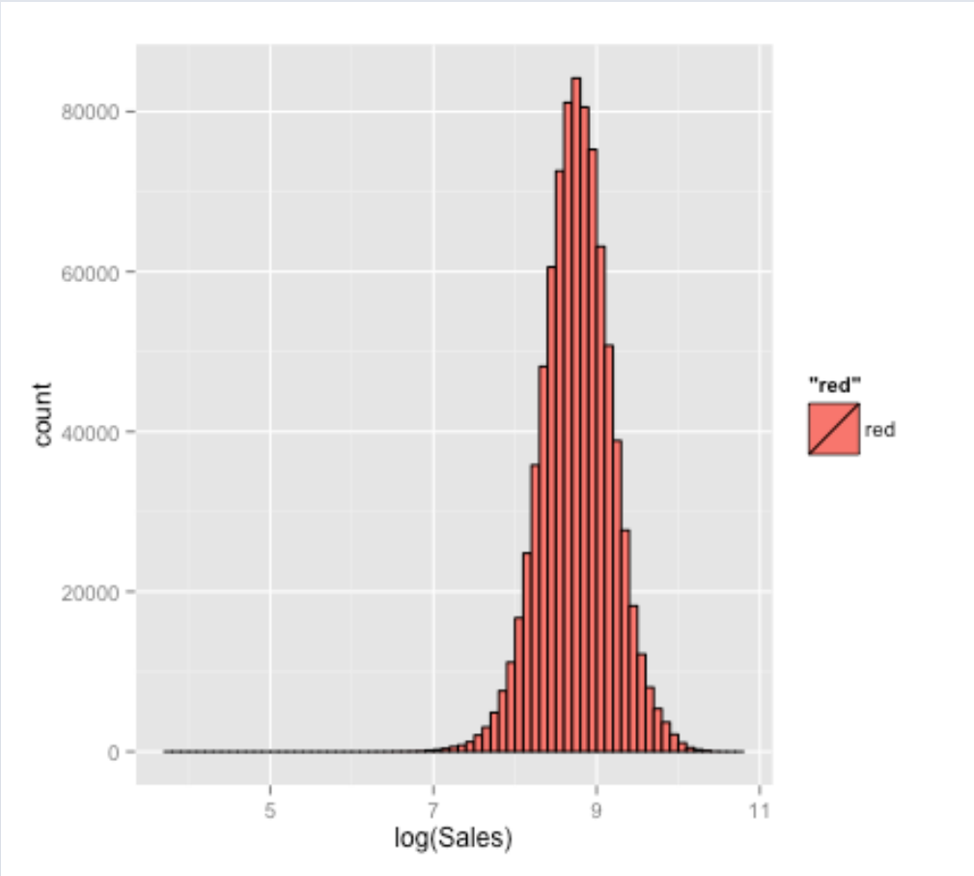












Motivations

- LASSO
  - Find the best predictors of sales
- Clustering
  - K-means was chosen due to the least computing time
  - Based on different types of stores
  - Useful for exploring mixture distributions
- Classification
  - Lack of numerical predictors
    - Only one relevant numerical predictor - distance to the nearest competitor
  - Poor cross validation MSE from LASSO
- Hidden Markov Model
  - Experiences tells us that stores have "good" and "bad" weeks.
  - An HMM allows sales to obey two distributions; *i.e.*, predict sales, given that it is a good or bad week.
  - A discrete-time Markov chain allows us to take into account the autocorrelation associated with temporal data.

Methods and Results

K-Means

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

- Given a large number of observations, k-means clustering provides the solution with the least amount of processing time required
- K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean
- There are 3 types of stores and 2 types of assortment

Log of Sales	Day of Week	Month	Promo	State Holiday	School Holiday	StoreType	Assortment	Competition Distance	Promo2
7.20	4.00	5.78	0.38	1.05	0.18	2.86	2.24	7284.79	1.00
7.24	4.00	5.83	0.38	1.05	0.18	2.34	1.89	17156.78	1.00
7.17	4.00	5.73	0.38	1.05	0.18	2.02	1.71	1476.51	1.00

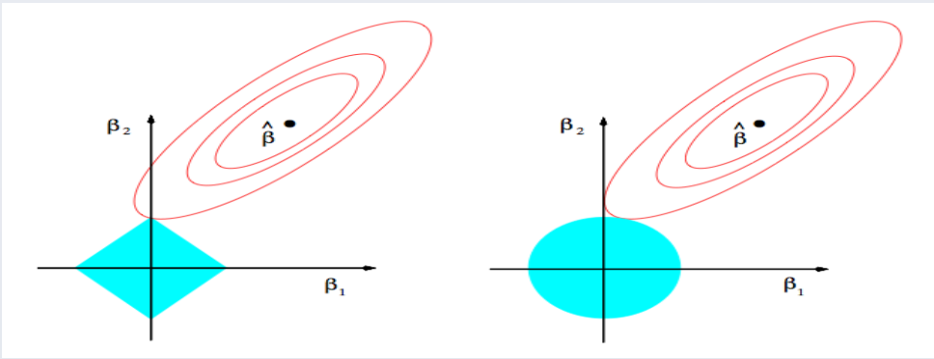
LASSO

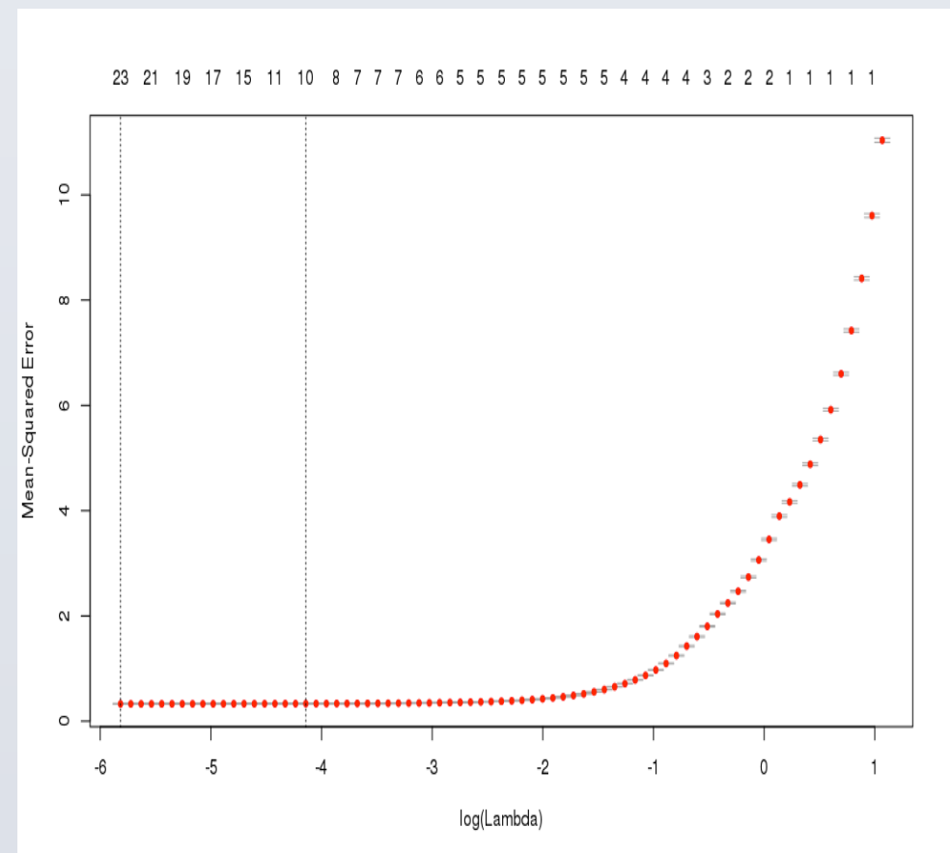
$$\hat{\beta} = argmin\frac{1}{n}\sum(y_i - x_i^T\beta)^2 + \lambda||\beta||_1$$

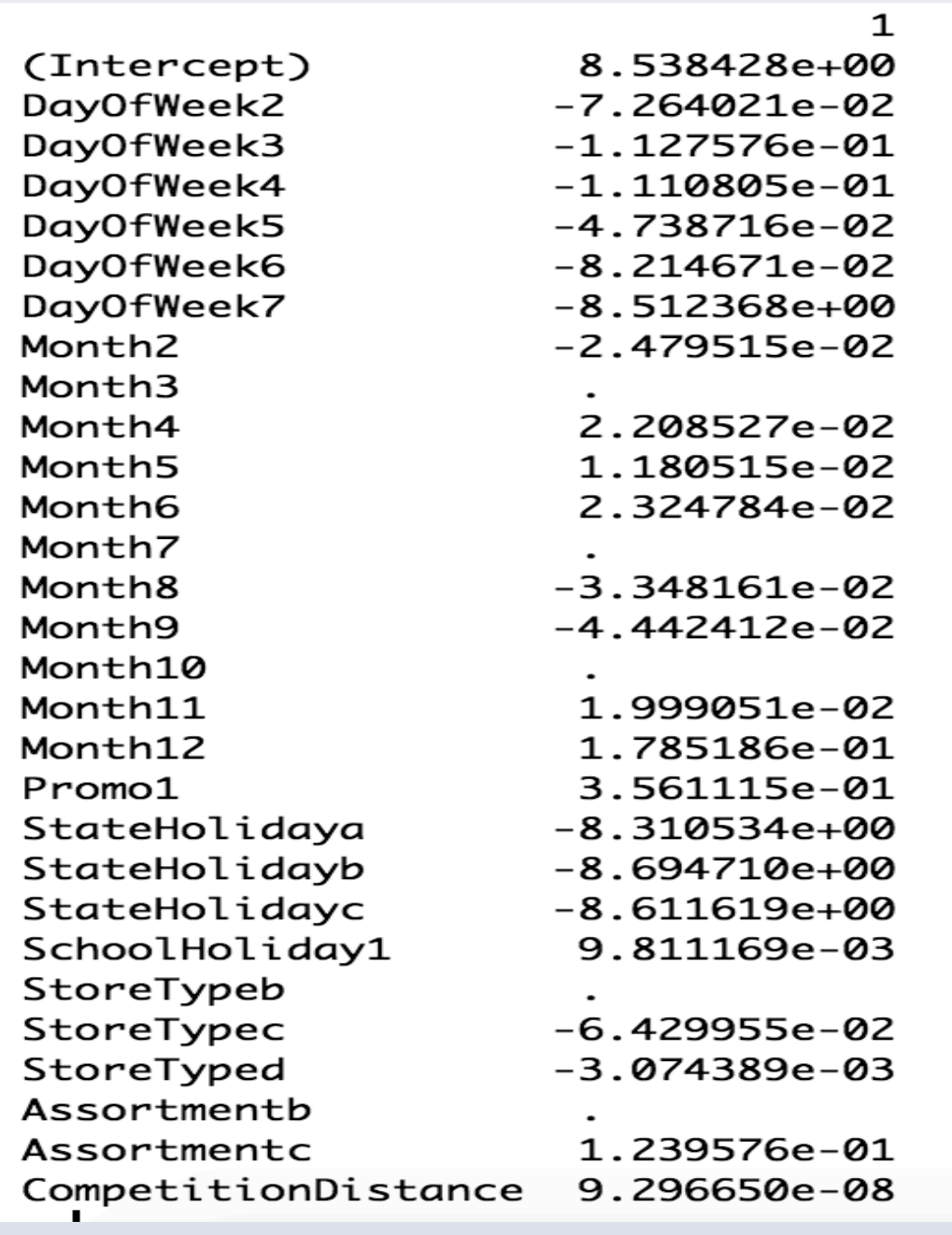
Sparse regression

- Regularization: sacrifice bias to reduce variance - more stable
- Interpretation: find a small set of strong predictors

Use 1-norm to approximate 0-norm







```
                                  1
(Intercept)            8.538428e+00
DayOfWeek2            -7.264021e-02
DayOfWeek3            -1.127576e-01
DayOfWeek4            -1.110805e-01
DayOfWeek5            -4.738716e-02
DayOfWeek6            -8.214671e-02
DayOfWeek7            -8.512368e+00
Month2                -2.479515e-02
Month3                 .
Month4                 2.208527e-02
Month5                 1.180515e-02
Month6                 2.324784e-02
Month7                 .
Month8                -3.348161e-02
Month9                -4.442412e-02
Month10                .
Month11                1.999051e-02
Month12                1.785186e-01
Promo1                 3.561115e-01
StateHolidaya         -8.310534e+00
StateHolidayb         -8.694710e+00
StateHolidayc         -8.611619e+00
SchoolHoliday1         9.811169e-03
StoreTypeb             .
StoreTypec            -6.429955e-02
StoreTyped            -3.074389e-03
Assortmentb            .
Assortmentc            1.239576e-01
CompetitionDistance    9.296650e-08
```

Support Vector Machine

There are a total of 3 categories:

- Failing stores whose sales are at the bottom 25%
- Average stores whose sales are between 25% and 75% quantile
- Star stores whose sales are at the top 25%

	Misclassification Rate
Linear SVM	0.276
Nonlinear SVM with Gaussian kernel	0.380

Hidden Markov Model

We can associate different Gaussian distributions with (unobserved) states in a Markov Chain. Here, we consider only two: the hidden states where the market is "good" and "bad." After all, it is conceivable that, on a "good" day (*i.e.*, when the store should have good sales), it still sells little merchandise.

Thus, we assume that the states obey the following transition distribution, which is homogeneous in time:in the good state, S ~ Normal(μ1, σ2); in the bad state, S ~ Normal(μ2, σ2).

From/To	Good	Bad
Good	π_1	$1-\pi_1$
Bad	$1-\pi_2$	π_2

Conclusions

- From LASSO regression, we can see that state holidays and sundays have the greatest impacts on sales.
- Using support vector machine, we can categorize a store very well with about 25% misclassification rate.