Kaggle Walmart sales prediction

Data Science General Assembly in DC

19 May 2014

Yifan Li

Contents

- The problem
- The data
- The model
- The result
- Conclusion

The problem

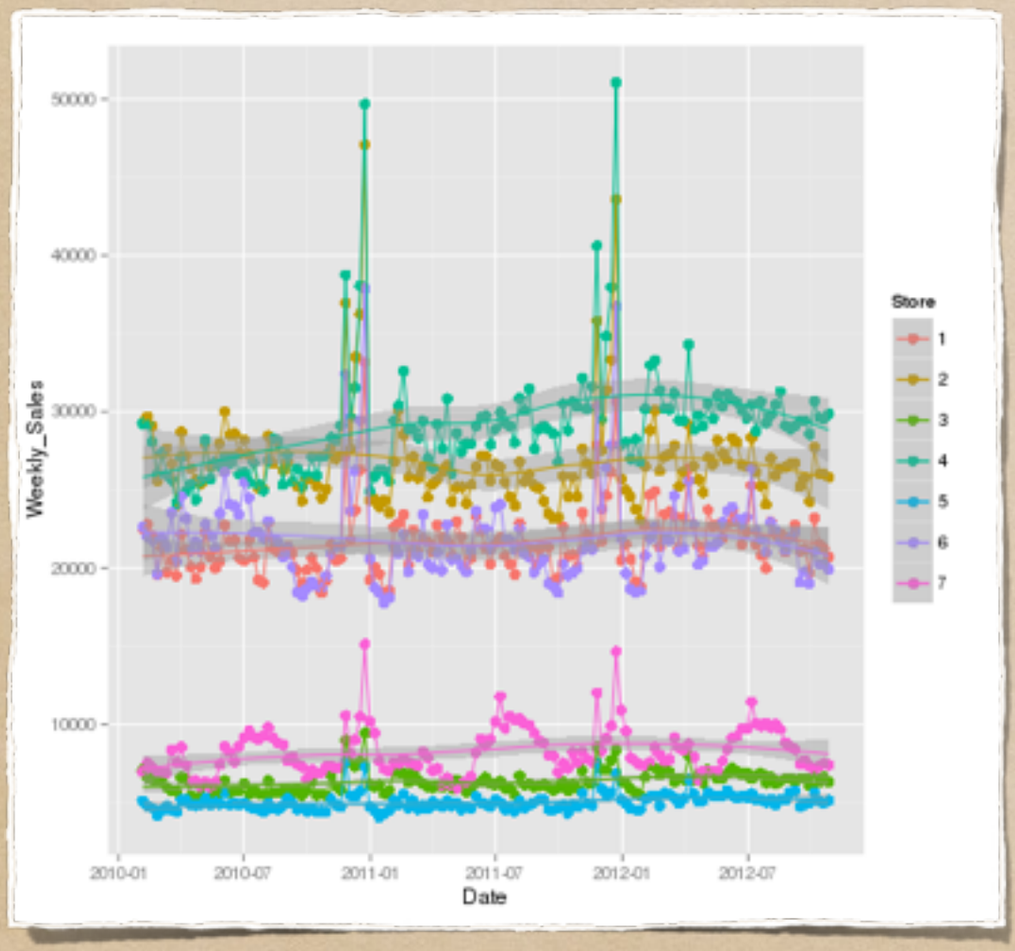
- To predict the Walmart sales number
- Regression problem

1. Given past sales, markdown(sales) events
2. Given associated CPI, temperature, unemployment, fuel_price, store type, store size

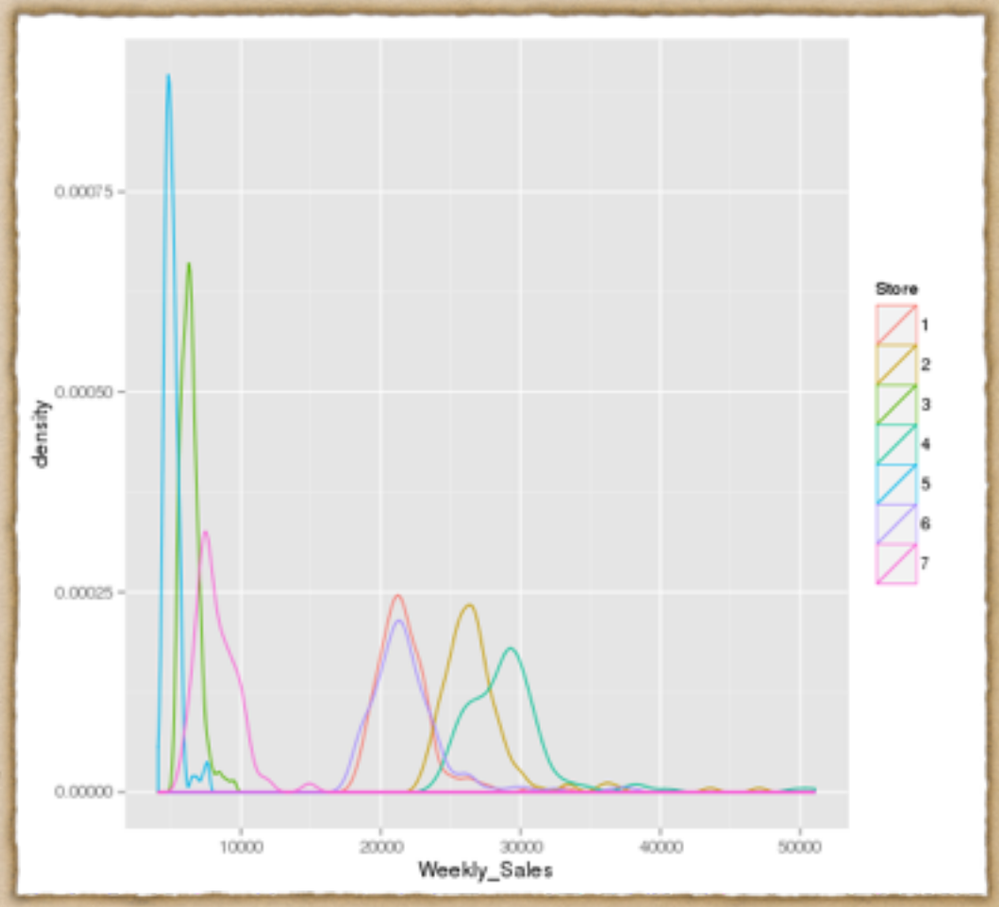
The data

- The weekly sales data corresponding to store

line plot



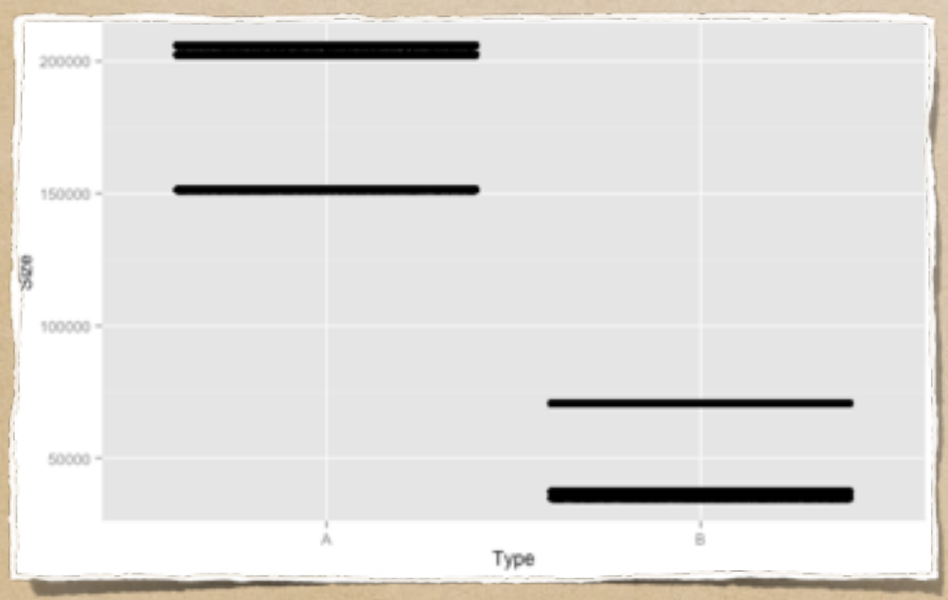
histogram



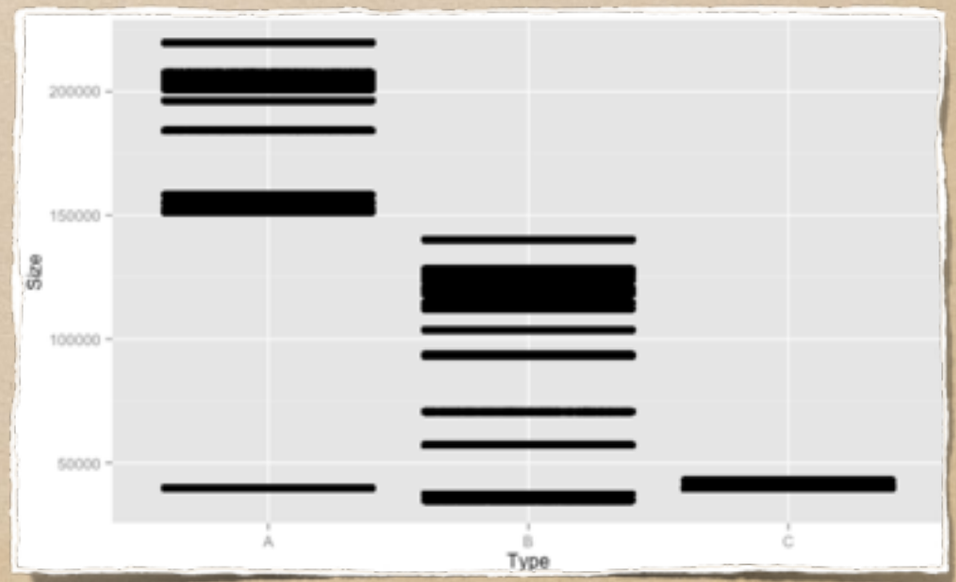
The data (cont.)

- Which feature to choose?
- type vs. size

Training data



Test data



The data (cont.)

- Handling missing Markdown data
    1. Fill it with zero
- Handling missing CPI, temperature data
    1. (CPI) Fill it with linear prediction of Temperature and Fuel Price
    2. (Unemployment) Fill it with linear prediction of CPI and Temperature

The model

- linear regression: lm{stats}
- regularization: glmnet{glmnet}
- least absolute deviation: rq{quantreg}

The result

- linear regression

656 new 一凡 李 22105.65108 28 Thu, 01 May 2014 22:05:22

Your Best Entry

You improved on your best score by 331.94296.

You just moved up 1 position on the leaderboard.

```
with model <- lm(Weekly_Sales ~poly(TotalDayFeature,3)+ Size+IsHoliday:Month  + Temperature:Fuel_Price
          +Dept * (MarkDown1+MarkDown2+MarkDown3+MarkDown4+MarkDown5)+CPI+Unemployment,data=tfs)
```

The result(cont.)

- regularization(result is not improved)

656 new 一凡 李 22105.65108 29 Thu, 01 May 2014 22:18:10 (-0.2h)

Your Best Entry

Your submission scored **22800.75939**, which is not an improvement of your best score. Keep trying!

lasso and ridge is not better……

The result(cont.)

- linear regression with weight(1 for normal week, 5 for holiday week)

657 new 一凡 李 21968.62304 32 Fri, 02 May 2014 00:48:15

Your Best Entry

You improved on your best score by 137.02805.

You just moved up 1 position on the leaderboard.

using the weight function to lm, moved up 1 position
But it's only weighted square sum, not weighted absolute sum.

```
weight <- rep(1,nrow(tfs))
weight[tfs$IsHoliday == T] <- 5
```

The result(cont.)

- least absolute deviation with weight

579 new 一凡 李 18079.20341 34 Sat, 03 May 2014 05:51:21

Your Best Entry

You improved on your best score by 3889.41962.

You just moved up 88 positions on the leaderboard.

```
fit1 <- rq(weights = weight,tau = 26/50,
Weekly_Sales ~poly(TotalDayFeature,3)+ Size+IsHoliday:Month  + Temperature:Fuel_Price
+Dept * (MarkDown1+MarkDown2+MarkDown3+MarkDown4+MarkDown5)+CPI+Unemployment
,data=tfs)
```

The result(cont.)

- first Kaggle experience
- 44 submissions(maximum 5 submissions per day)
- 592nd/694(18416.22852 points)
- beat the all zero benchmark(647th,22265.71813 points)

Conclusion

- Linear regression is a good starting point for feature selection (which takes a lot of time)
- Using model that corresponding to the evaluation method may improve score
- The best five on leaderboard uses Autoregression, Random Forest:

With limited data, more sophisticated algorithm would be beneficial