Our 2015 Data Mining Cup Solution

Iowa State University Team 2

or

Creating Useful Features in A Wasteland

Our github page:
github.com/imouzon/dmc2015

Representing Iowa State University

and

My Incredible Team

Iowa State

The Stats Dept

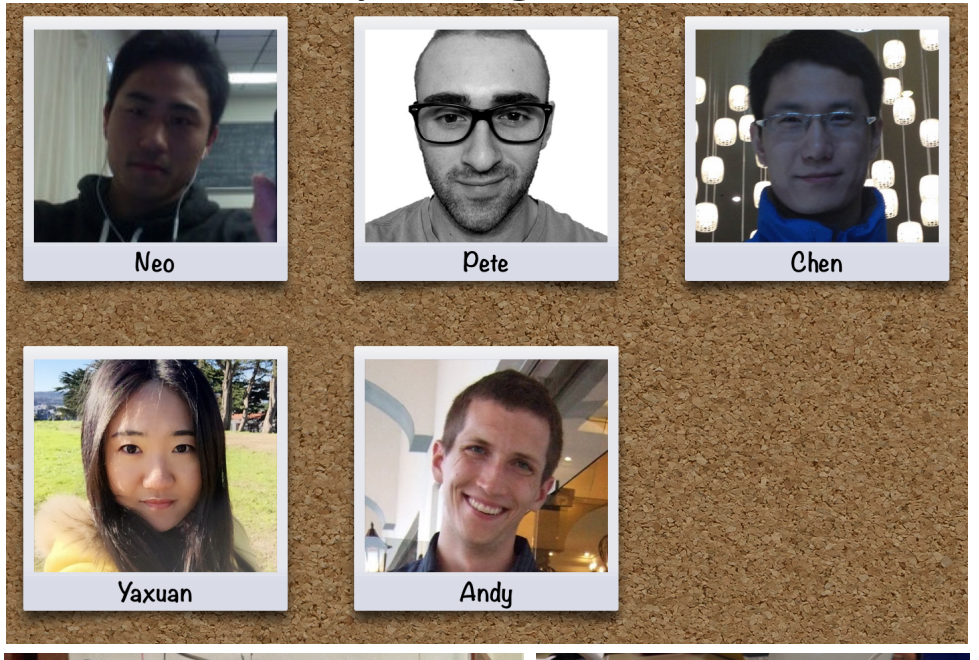
Statistics At Iowa State Univeristy

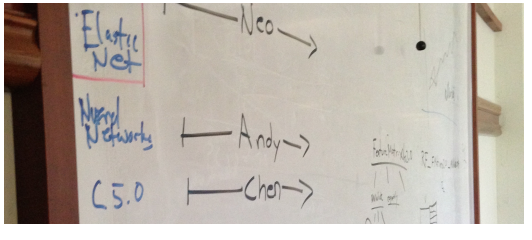
A Long History of Science with Practice

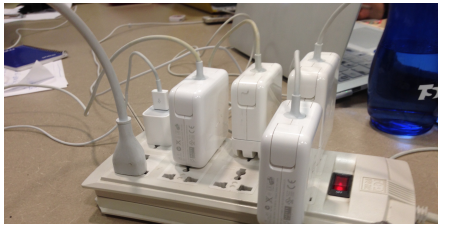
- First statistics course offered in 1915
- Department established in 1947 (the first statistics department in the US)
- Snedecor Hall built in 1953

Our Team

We were able to put together a terrific team







What We Did
The Loss and The Features

Dealing with Loss

$$E = \sum_{i=1}^{n} \left[\left(\frac{u_i - \widehat{u}_i}{\frac{1}{n}\sum_{i=1}^{n} u_i} \right)^2 + \left(\frac{v_i - \widehat{v}_i}{\frac{1}{n}\sum_{i=1}^{n} v_i} \right)^2 + \left(\frac{w_i - \widehat{w}_i}{\frac{1}{n}\sum_{i=1}^{n} w_i} \right)^2 + \left(\frac{b_i - \widehat{b}_i}{\frac{1}{n}\sum_{i=1}^{n} b_i} \right)^2 \right]$$

The Loss Function Is Weird

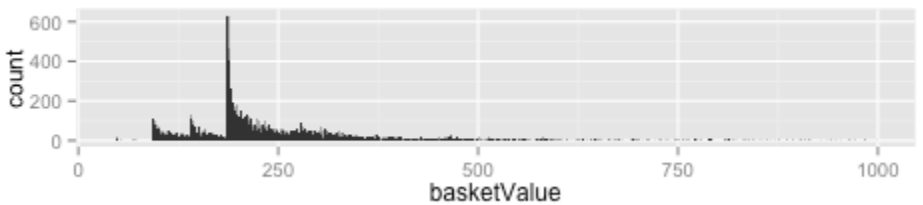
The basic piece of the loss function is

$$\sum_{i=1}^{n} \left(\frac{x_i - \hat{x}_i}{\bar{x}} \right)^2$$

- We get punished for for missing in either direction
- We get punished for more as the more we miss by
- we get punished more if the mean response is low

We have to predict the mean response too!

Consider `basketValue`:



There are some enormous values in there - how should that impact our prediction?

Error and `basketValue`

- **Possible Scenarios:** True Value Contributes to **LOSS** and **WEIGHT**

	True basketValue is huge	True basketValue is NOT huge
Prediction is huge	Small Error Huge Mean	**Huge Error** **Small Mean**
Predict is small	Huge Error Huge Mean	Small Error Small Mean

- Obviously, we are never better off making incorrect predictions
- *But* we could lose the entire competition just by making one mistake in basketValue

Decision

Use exponential tools that restrict prediction size

Cory Lanker's LLR Approach

Let k be some class of observations found in a historical data set $\mathbf{H}$

Let $u_{i,k}$ be the response-value related to ith observation from class k in $\mathbf{H}$.

We define the log-likelihood ratio statistics for class k :

$$\begin{aligned} \text{LLR(k)} &= \log\left(\frac{\sum u_i + \epsilon_1}{n - \sum u_i + \epsilon_2}\right) \\ &= \log\left(\frac{\text{adjusted number of times class } k \text{ is used in } \mathbf{H}}{\text{adjusted number of times class } k \text{ is NOT used in } \mathbf{H}}\right) \end{aligned}$$

(the greek terms are chosen in a clever way involving Bayesian estimation).

For each class $1, 2, \ldots, k$, we have a different response - the higher the response the more likely future outcomes from the class are to take the value 1.

Cory Lanker's LLR Approach

All the categorical values can be replaced with meaningful numeric values

Basic LLLRs

- select some observations from our data to create \(\mathbf{H}\)
- We can calculate LLRs for coupons (coupon ID = class)
- We can calculate LLRs for brands (brand = class)

LLR Interactions

- We can calculate LLRs for brand and product group interaction:
- Consider each pairing of brand and product group to be a class

In this way ALL categorical variables can be transformed into numeric values

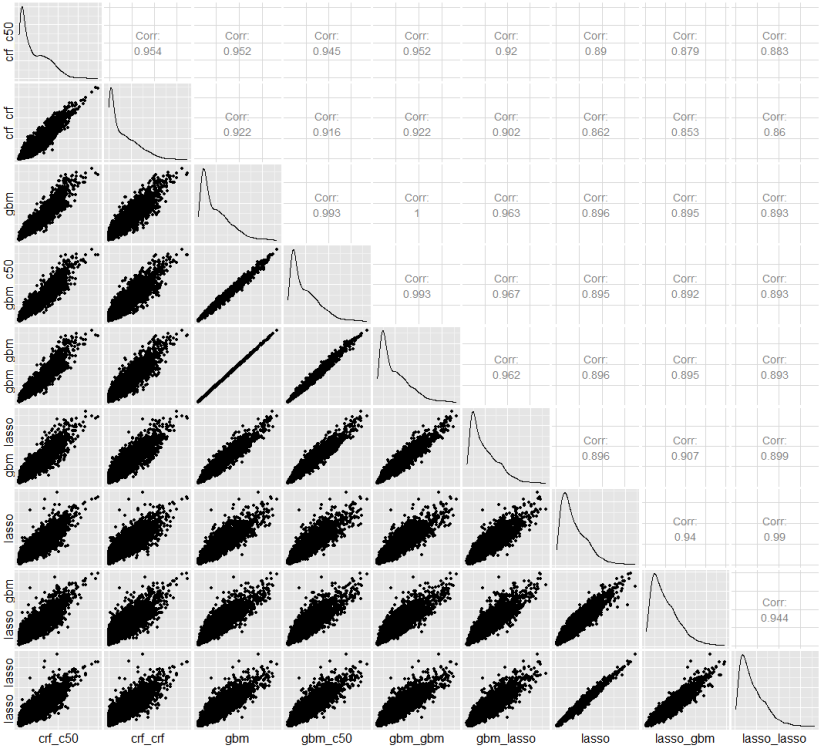
Final Predictions

Creating a Final Prediction

Multiple Methods

Aviod Correlation

Improve By Combining

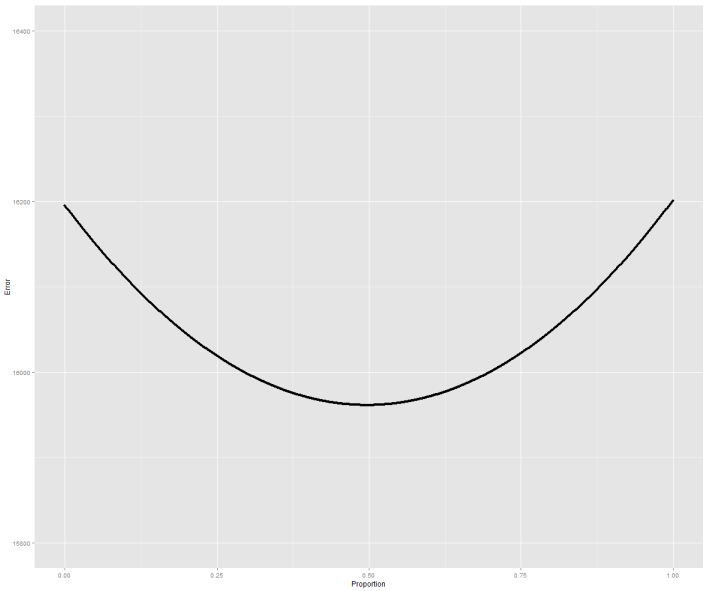


Putting Predictor Together

Combining predictors leads to the best final prediction

The best solution involved a combination of

1. Predictions fit by lasso based on variables selected by lasso
2. Predictions fit by Conditional Random Forest based on variables selected by Conditional Random Forest



We Are So Thankful To Have Been Invited

We Would Like To Thank prudsys For This Opportunity

and

ISU's Department of Statistics for Their Support