Capstone Project Proposal –

Models for Sales Prediction

**Foundations of Data Science**

Christopher Himmel

April 28, 2016

**Motivation**

In retail, predicting sales volume can be used in maximizing in-stock positions and maximize revenue. The more accurate sales can be projected, the better replenishment models can anticipate out of stocks, and keep inventory on the shelves by reordering product in a timely manner.

Replenishment is a core part of JDA’s retail ERP, Merchandise Management System (MMS), which is the reordering of inventory based on current levels of inventory, seasonality, and sales history. The algorithm uses sales history by week, and could be more accurate if it had an idea of projected sales. Integrating this kind of data into the model would translate into decreasing lost sales due to stock outs.

This project will be developed with my current company’s data, but is meant to be generic to any JDA MMS implementation. The relevance of the SKU attributes’ may vary from company to company, and the correlation will have to be measured and which attributes make it to the final model with vary.

**Methodology**

The source data to be used for this comes directly out of SRG’s ERP, Merchandise Management System. This system resides on a legacy Series-I (AS/400) platform, in a DB2/400 database structure. The tables used are as follows:

INVMST – SKU/Item Master

This table contains all of the attributes of the selling items entered by the user or loaded in directly from the vendor.

INVBAL – Inventory Balance by Store

This table contains summary information of the selling items, down to the store level. This information includes On Hand Units, Sales by Week for the last eight weeks, and much more status fields for the SKU at that store.

INVCBL – Inventory Balance at Chain

This table contains summary information of the selling items combined for all stores. Information like On Hand Units and Sales by Week (and other statuses of the items) are at chain level, in contrast to broken down by store in INVBAL.

A subset of stores is used for this analysis, which limits the INVBAL to just the three stores that make up the Niquea’D concept (1001, 1002 and 1003). SQL is used to filter these store numbers out of the much larger table of 400+ stores. Once we have a limited number of INVBAL records we can then limit the number of INVMST Item Mater records to only include those sold through the Niquea’D concept. After the filters, these tables are transferred from the Series-I and stored on a Windows file system in comma delimited format. They are then ready to be loaded into R for analysis.

**Approach**

Once these tables loaded into R, the first task will be to get it into a format that can be fed into a learning algorithm. This entails combining the tables into one, with one row consisting of one observation. Each observation will be one SKU, with every attribute as a different column. As well, the units sold for the previous eight weeks, with the current units on hand will be attached as additional columns. Finally, the description field will be broken down into word vectors, with the top fifty words considered for the model. This translates into fifty fields on each observation (SKU), each having a 1 if the word is in that SKU’s description.

Once the observations are built, each attribute column is analyzed as to whether it has any significant data in it. If the column is the same for all SKU’s, or only a few SKU’s contain any differing information, then the SKU will be removed from the model. This is done by graphing the distribution of each column, and visually determining whether any data is present. Any missing data is filled in with defaults at this point.

Now correlation between each column is calculated and reviewed as to whether the information that will one column contributes to the model is the already covered by another column. This can only be done for numeric fields, and does show some field correlations.

Finally, to get the data ready for building models, a Linear Regression model is built over the complete data set. This exposes which attributes of the SKU have no correlation to the last complete week of sales, which is being predicted.

Now the minimal columns of data will be split between Training and Test sets, and models will be built. Linear Regression, Decision Trees, Pruned Decision Trees, and Random Forest models will be used to develop models on the Training set. Then each tree will be evaluated on the Test set.

Finally, a comparison of each model will be made and a determination numerically and visually will be made as to the best fit.

**Deliverables**

The final results of this project will be a Knitted Word document, containing the story of analysis from beginning to end, with lots of explanation.