Rossmann Sales Kaggle Competition

Jonathan Dickerson, Madhav Kapoor, Matthew Osinski
Stevens Institute of Technology

Maintaining a profitable position in the food and drug store industry requires the ability to continually increase efficiencies, in order to maintain a competitive presence in a low margin environment. With over 3,000 stores in 7 European companies, Rossmann GmbhH is looking to improve the accuracy of its store sales forecasts. Currently, store managers in 1,115 stores in Germany judgmentally predict their respective store sales up to 6 weeks in advance. These metrics are then used to help with purchasing, inventory and staffing needs for each individual store.[2]

Business Understanding

Rossmann GmbhH is a German based Drug Store company with over 3,000 stores located across 7 European countries. Within Germany alone, Rossmann has 1,115 store locations. In its current standing the company is faced with the dilemma of properly forecasting sales for each store located in Germany. The current formula for forecasting sales has been through multiple sales forecast reports provided by store managers individually. From a corporate standpoint, receiving roughly 1,115 different forecast reports from each of the stores provides a cumbersome task for senior management to properly evaluate what drives certain stores to yield higher returns than others. Rossmann has tasked us with creating a streamlined predictive model that is able to process, transform and model data from multiple store locations in order to provide the corporation with enhanced sales forecasts.

With this task in mind there are some considerations that have to be made towards what affects sales over multiple locations. As the problem addresses, there are multiple factors that apply directly to sales of any given company. It is our job to consider these factors such as promotions, holidays, seasons, and in particular our competition to properly evaluate and predict sales performance. Out of the many factors that can be applied to projected sales of Rossmann, competition should be towards the top of this list. Since Rossmann has just under 50% of its total store locations situated in Germany, we can assume that there is a significant weight of competition within the Drug Store market. With this in mind, Rossmann also utilizes the factor of locality which helps identify what sort of distance from a store provides the most amount of frequent sales to occur. In order for Rossmann to properly evaluate forecasted sales properly we must also consider all significant effects.

Our approach focuses on not only conveying the prediction of Store Sales to Rossmann, but also the significant drivers of sales, the accuracy of the forecast, the most inaccurate predictions, and information on what could be causing the inaccuracies.

Data Understanding

Predicting same store sales is a widely researched topic. According to Grewal et al, some of the key factors that influence store sales include:

- Store factors: Accounting for the physical location of the store, the stores atmosphere and the store condition
- Service rates: Accounting for the customers' interaction: efficient payment transactions and good customer service for post-purchase problems
- Merchandise: Accounting for the management of SKUs
- Price: Accounting for price, promotions and competitors pricing These factors help determine the number of customers that will visit a particular store, and how much each of those customers will spend, which will ultimately have to be accounted for within the model.[4]

Training Data Set

Keeping these factors in mind, Rossmann has provided the following data in their training set:

- **Id** - an Id that represents a (Store, Date) duple within the test set
- **Store** - a unique Id for each store
- **Sales** - the turnover for any given day (this is what we are predicting)
- **Open** - an indicator for whether the store was open: 0 = closed, 1 = open
- **StateHoliday** - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools
- **StoreType** - differentiates between 4 different store models: a, b, c, d
- **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
- **CompetitionDistance** - distance in meters to the nearest competitor store
- **CompetitionOpenSince[Month/Year]** - gives the approximate year and month of the time the nearest competitor was opened
- **Promo** - indicates whether a store is running a promo on that day
- **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating

- **Promo2Since[Year/Week]** - describes the year and calendar week when the store started participating in Promo2
- **PromoInterval** - describes the consecutive intervals Promo2 is started, naming the months the promotion is restarted. E.g. “Feb,May,Aug,Nov” means each round starts in February, May, August, November of any given year for that store[2]

Sales Variable

The first course of action that was taken was to review the target variable, Sales. In the examples below, the distribution is skewed right, and according to prior work with supervised learners such as regression trees, transforming the data to produce a normal distribution helps to lower the error rate[3].

The sales data closely follows the log-normal distribution, as illustrated in Figure 1. This suggests a particular transformation, which we’ll discuss in the Data Preparation section. The normal probability plot for the untransformed data is given in Figure 2.

We then moved to the variable Open. The data description said: ”Note that some stores in the dataset were temporarily closed for refurbishment.”[2] These closures were confirmed plotting the number of open stores over time. This plot is given in Figure 3.

We also examined seasonality affects, to attempt to determine if a full time series model would be effective, or if we could get away with a simpler model, perhaps controlling for month or day. The average sales across all stores for each week number is given in Figure 4.

The weeknumber variable wasn't provided by Rossmann, but was calculated from the date field.

Data Preparation

As seen in Figures 1 and 2, the Sales variable is log-normal. This distribution is transformed into a normal distribution under the transformation $\log x$, as the name suggests. We re-plot the variables to inspect the results of the transformation. These plots are given in Figures 5 and 6.

To account for the closed stores, instead of imputing the mean or zeros, or any other technique, we removed the days the store were closed from the data set. This had an effect on sample size, but the final row count was still sufficient to train the model (roughly 850k rows).

One-Hot Encoding

The data set included several categorical variables, including store type, assortment, and state holiday. To be used as inputs into a machine learning algorithm, they needed to be transformed into integers. However, a simple $[a, b, c, d] = [0, 1, 2, 3]$ introduces both a ranking and a distance measure. Is d better than c? Is the difference between d and c the same as the difference between c and b? To get around these issues, we used a one-hot encoding scheme. This takes into account the full set of categorical variables, and all their possible values, then calculates the required number of bits to encode all the possibilities using only 1 and 0.

Modeling

An assessment of the Rossmann data yielded that the major drivers of store sales, as determined by Grewal, et.al [4], are not accounted for. However, the effect that those major drivers have may be seen more clearly by decomposing the variable Sales.

As a result, a time series decomposition was performed on the store sales. Breaking down Store 1s sales into a trend, seasonality and noise component demonstrates that there is a clear seasonal pattern. As a result, the algorithm will need to include variables that capture this information, such as a month and week variable. The decomposition is shown in Figure 7

A time series will be difficult to perform because there are large amounts of missing and non-continuous data within both the training and test data sets, which can be seen in Figure 3.

To best account for these effects, we viewed the issue as a supervised learning problem, with the input vector including the transformed data about the store, time and promotions, while the target vector would be the sales forecast that is to be projected.

Defining the input vector required additional feature engineering. As mentioned previosuly, the seasonality was encoded by splitting the DateTime variable into Week and WeekDay. The categorical variables, including Promotion, State Holiday, School Holiday, Store Type and Assortment were encoded into a matrix using the one-hot-encoding paradigm described above.

The original exploratory analysis was done using OLS regression and Decision Trees, and as they performed reasonably well, we kept them going forward with the reduced data set. However, we began to look into ensemble methods to potentially decrease the error.

After may attempts of utilizing Linear Support Vectors and Decision Trees, our next hypothesis was to test if using a boosting algorithm would improve the nature of the results. Because Adaboost trains weak learners with the aim of selecting a weak hypothesis with a low weighted error, or computing a weighted majority vote of the weak hypothesis.[1] In addition, AdaBoost works well when handling noise and outliers in real world data, which made it an attractive option as the training data and test data sets did experience several outliers within the store sales and store promotions fields.

After seeing the benefits provided by the AdaBoost algorithm, additional Boosting models were researched. Unlike the AdaBoost method, which improves the regression by varying the weights in the data points, the Extreme Gradient Boosting model (XGBoost) compensated by identifying the gradients and provides[5]. In addition, XGBoost allowed us to customize the loss function to fit Rossmanns method of evaluation (RMSPE)[6]. The performance improvements of the main iterations during the modeling phase on the training set, can be seen in Figure 8, and we selected to utilize the Extreme Gradient Boost model to evaluate the test set.

Evaluation

The performance of the model will be evaluated on the Root Mean Square Percentage Error of the sales defined as

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

As a simpler proxy in training and testing, we used root mean square error, since the two would yield the same ordinal ranking of algorithms. We tested a few algorithms in attempting to discover a good model, and the RMSE results suggested that the boosted decision trees provided the best results.

The current working model achieves a RMSPE of 0.12072 on the test training set. Because Rossmans managers use individual methods to compare store sales, our benchmark will simply be the prior quarter average store sales, which has an RMSPE of 0.20. Clearly, our model has substantial predict power in its current form. In terms of actual sales, our model predicts €275.50 M in total sales, while the benchmark predicts €250.68M in total sales.

Deployment

The deployment will focus around the key stakeholders involved in Rossmann Stores, which include the Store Managers (who are responsible for setting the forecasts), Area Managers and Vice Presidents.

The dashboard provided has the ability to select a particular store number or to view an aggregate mixture. The users will be able to view the historically predicted Sales values, Actual Sales values and Predicted Sales values, both within a spreadsheet and within a time series visualization. The visualization of the Historical and Projected Sales can be seen in Figure 9.

In addition to the projected Sales, we will also provide Rossmans management the key drivers of Sales, from the model perspective. In the variable importance chart in Figure 10, the key drivers are Customers, Day of Week, Store and Promotion. If over time a certain variable displays more importance, it can guide model refinement decisions.

With the development of our live time analysis dashboard we can first present Rossmann and the company stakeholders with a viable solution that is able to help predict and forecast results with a mean square error of 12% (currently). The dashboard allows us to properly evaluate sales from a high level perspective to help internal analysts determine what sales maybe in the coming days. Like any other good analysis dashboard or tool, all results must come with the understanding of noise. As weve seen based off the results of our variable importance analysis customers are the main driver of profits in this company. In order to properly create profit turnover for the number of customers to sales, the Company needs to invest more into the relationships built between the Store and Customer. In order to meet this requirement, the corporate office of Rossmann can begin to look into new promotional campaigns geared towards their most significant clientele. Seasonality and Holidays seem to play hand in hand with higher peaks of effect on over sales towards the end of the year. If the company plans on creating a new formula for holiday promotions to increase sales it may help increase the overall sales turnover for end of year profits.

Another plan of action that Rossmann can adhere to is enhancing community appeal around particular store locations. As we can see Store and Store Type have some general effect on sales, so by creating an initiative to become more involved in the community and general location of customers, sales should increase in turn. As mentioned previously, with just under 50% Stores located in Germany alone, Rosssmann has the ability to become a family name if the company finds a new marketing or promotional plan to help feel customers that utilizing their Drug Stores is better than competitors.

To promote full disclosure and transparency, Rossmann will be provided a report of the stores that are misclassified, along with the reasons why they are being misclassified. A report that shows the stores with the most inaccurate sales predictions with a Root Mean Square Percentage Error above Rossmanns current benchmark (20%) will be displayed to Rossmann leadership.

Same Store sales can trend up for numerous factors, including changes in: customers, average revenue per customers, competition, weather and new store managers. This area will need to be explored in depth with Rossmann to further identify what their personal experiences are in historical drug store sales trends.

Conclusion

The result of all of this effort if a full production-ready model which could be deployed on top of Rossmann's transactional database, to give a continuously updated forecast of the next 6 weeks. If we see that the performance extends beyond that, we could generalize the model to do longer projections. The change in sales from our model, as compared to the benchmark, on the test data set is about 25M Euros, or about a 9% difference. This becomes very significant, as Rossmann has to compete with other retailers in the low margin environment. Because our model is more accurate, it makes a positive effect within Rossmanns operations, and will solve Rossmanns problems of many different modelling and projection paradigms being used simultaneously by the various store managers. This will allow for better planning of promotional schedules and better ability to forecast the effects of new competitors in the area.

Some areas for future work would be introducing time series elements to the model, or expanding the current variables to further explain some of the variance. Overall for the relative simplicity of the model, achieving 12% error is a good result. Further refinement and algorithmic tuning could get the error down further.

References

[1] Schapire, R. (n.d.). Explaining AdaBoost. Empirical Inference, 37-52.

[2] Rossmann Store Sales. Retrieved October 15, 2015, from https://www.kaggle.com/c/rossmann-store-sales

[3] Jank, W. (2011). Business analytics for managers. New York: Springer.

[4] Grewal, D., Krishnan, R., Levy, M., & Munger, J. (n.d.). Retail Success and Key Drivers. Retailing in the 21st Century, 13-25.

[5] Schapire, R., & Freund, Y. (2012). Boosting foundations and algorithms. Cambridge, MA: MIT Press.

[6] XGBoost Documentation. (n.d.). Retrieved December 11, 2015, from https://xgboost.readthedocs.org/en/latest/

[7] Rao, T. (2012). Time series analysis: Methods and applications. Amsterdam: North Holland.

Appendix

Distribution of Work

The team was comprised of Jonathan Dickerson, Madhav Kapoor, and Matthew Osinski. All three team members contributed to the writing of this document. The code was also worked on in parallel, with the final model comprising components of Jon and Madhav's data munging with Matt's tuning and choosing of the algorithms. The source code used in all data analysis and report writing can be found at https://github.com/jaydik/bia656.git. The code was completed entirely in Python using pandas and scikit-learn, with this report being written in LaTeX.

Kaggle Results

As of this writing, the competition is not completed, however our current RMSPE is 0.12072, which is good for 1869th place of 3429 participants. Our team name is justwhateverbia656 and the results can be retrieved from https://www.kaggle.com/c/rossmann-store-sales/leaderboard.

Figure 1: *Sales Histogram*

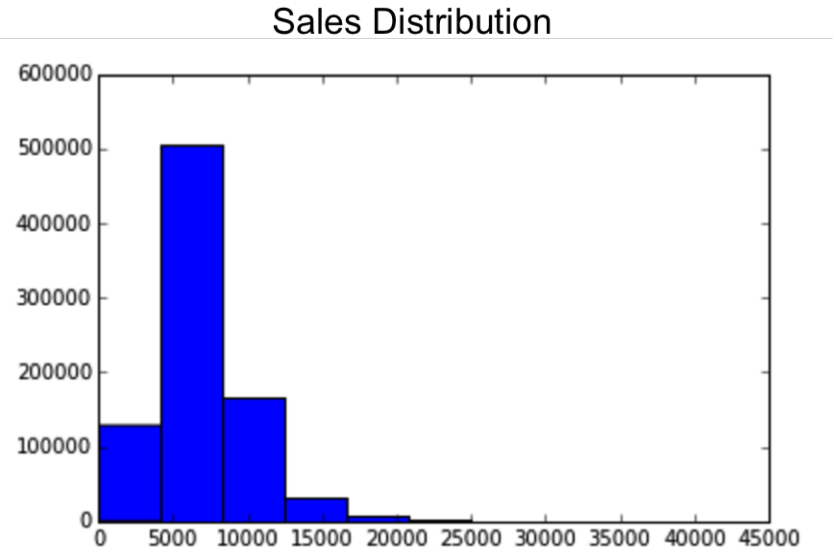


Figure 2: *Sales QQ Plot*

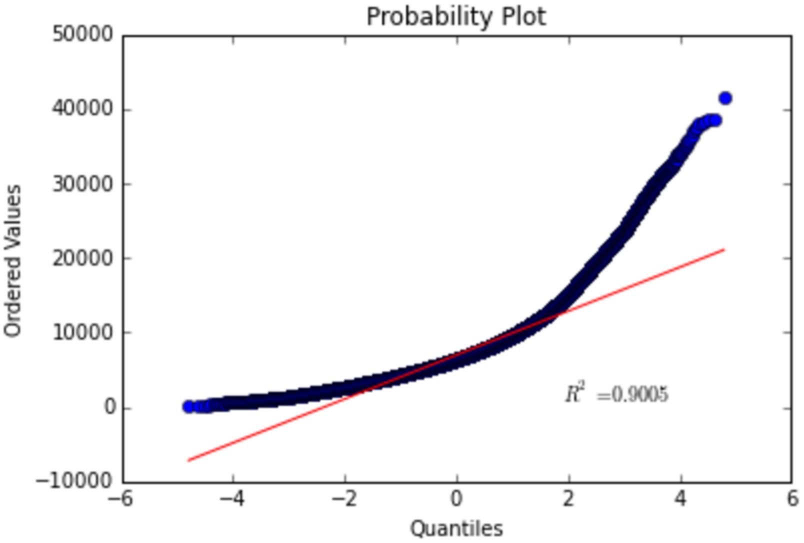


Figure 3: *Count of Stores*

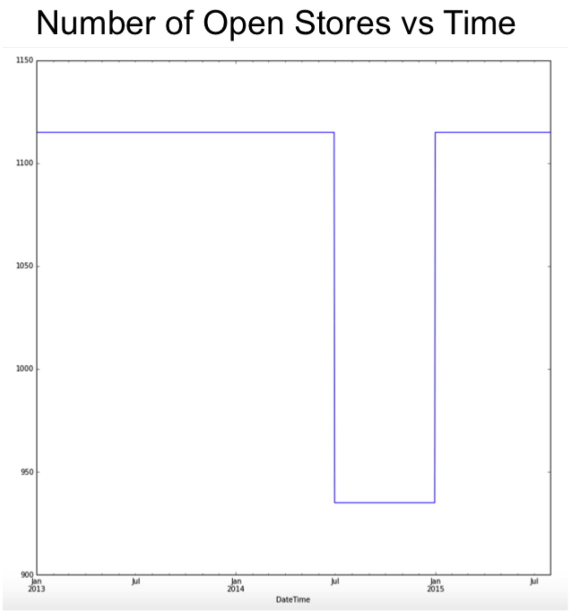


Figure 4: *Sales per Week*

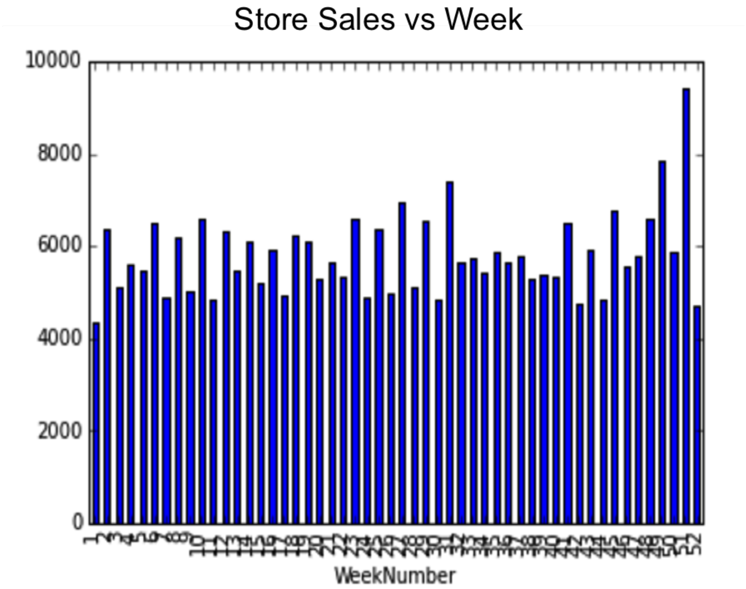


Figure 5: *Transformed Sales Histogram*

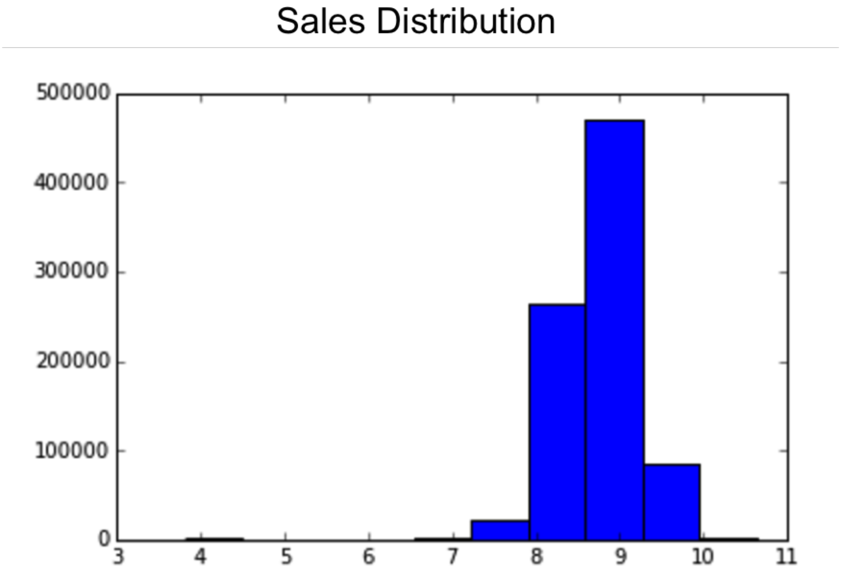


Figure 6: *Transformed Sales QQ Plot*

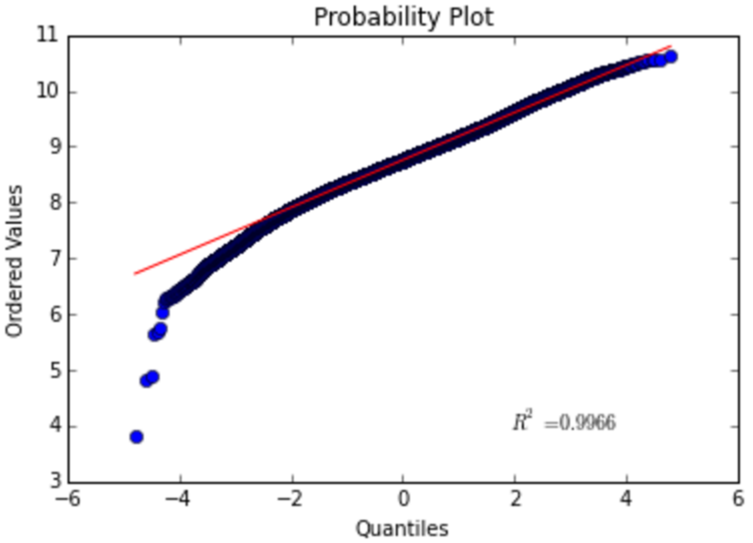


Figure 7: *Sales Decomposition*

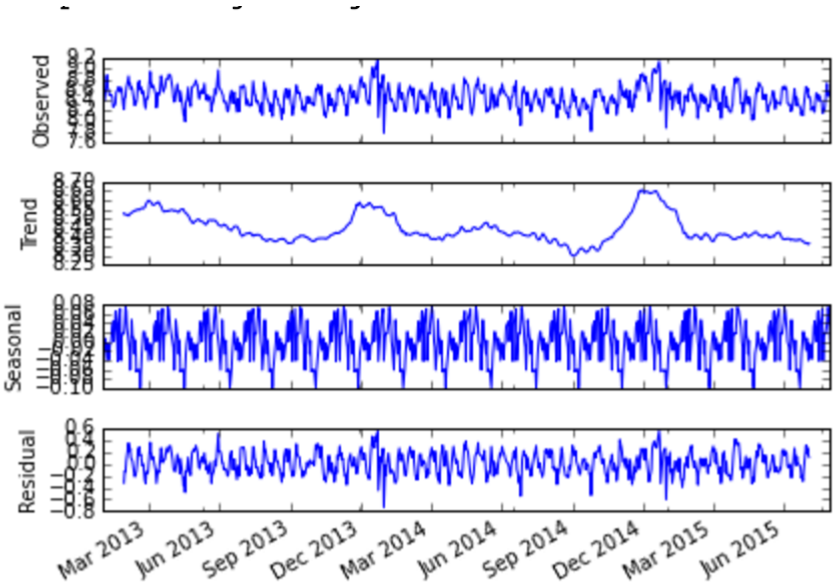


Figure 8: *Regressors and RMPSE*

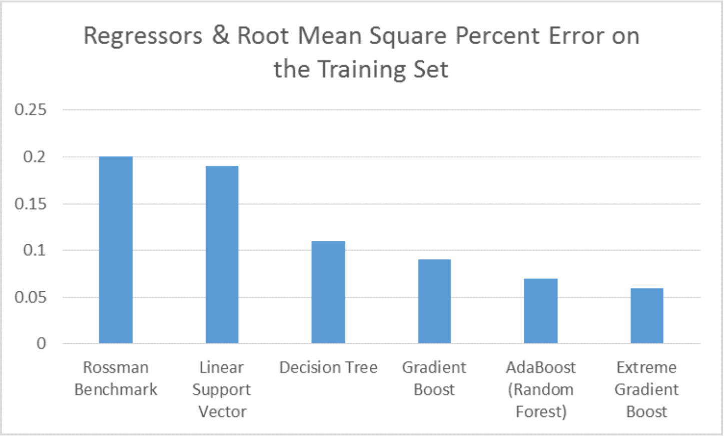


Figure 9: *Example Dashboard*

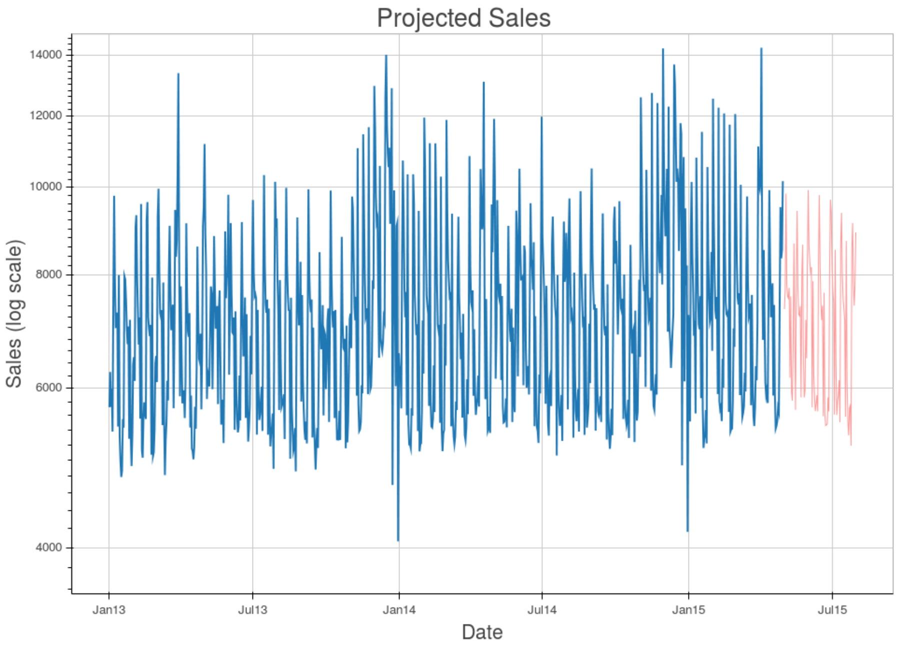
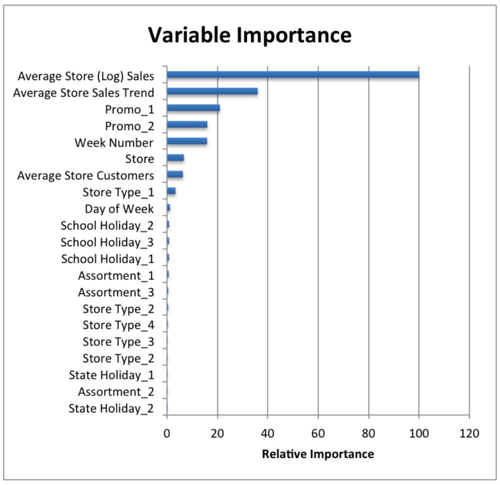


Figure 10: *Variable Importance*


Variable Importance
Average Store (Log) Sales
Average Store Sales Trend
Promo_1
Promo_2
Week Number
Store
Average Store Customers
Store Type_1
Day of Week
School Holiday_2
School Holiday_3
School Holiday_1
Assortment_1
Assortment_3
Store Type_2
Store Type_4
Store Type_3
Store Type_2
State Holiday_1
Assortment_2
State Holiday_2
Relative Importance