Jamie Yachera

Draft Project Paper – DC-DAT-9

11/19

**Model Overview**

Rossman Stores, a European drug store chain, provided store and sale information for 1,115 of their German stores as part of a 2015 Kaggle competition. Rossman seeks to eliminate the variability in their manual sales forecasting process by requesting participants develop a model that can accurately predict the next six weeks of daily sales. Kaggle posted two csv files of anonymized sale and store data for use in this exercise:

* Historical sales data from 1/1/13-7/31/15 (1k+ rows), including the following features:
  + Store ID
  + Day of Week
  + Date
  + Sales ($)
  + Customers (#)
  + Store open flag (0,1)
  + Store promo flag (0,1)
  + State holiday flag (category)
  + School holiday flag
* Store-level metadata (1115 rows), including the following features:
  + Store ID
  + Store Type (categorical)
  + Assortment (categorical)
  + Competition distance
  + Competition open month
  + Competition open year
  + Flag as to whether the store participates in “Promotion 2”
  + Week Promotion 2 started
  + Year Promotion 2 started
  + Consecutive intervals in which the promotion starts new

**Exploratory Data Analysis**

Each file was first evaluated for completeness. The isnull() function indicated that out of the 1115 stores, 3 were missing the Competition Distance, 354 were missing the Competition Open Month/ Year, and for the 544 stores that did not participate in the special promotion, they were missing the Week/ Year that the promotion started.

Two data sets were created to accommodate for the missing information. Both data sets removed the three stores that were missing Competition Distance. “Data Set 1” replaced the null Competition Open Since Month/ Year and the Promotion Week/Year with 12/1900 and 52/1900. This was initially performed to allow for new features calculating date differences from the sale date to when the promotion started or the competition opened. A second “Data Set 2” was created and each row containing a missing Competition Open Since Month/ Year was dropped resulting in 761 stores. The Promotion Week/Year was again replaced with 52/1900 to allow for later date difference calculations. Both data sets were evaluated and used throughout the model build and showed only slight differences. The remainder of this paper focuses on Data Set 2.

The groupby().Sales.mean(), .decribe, plot(kind=’hist’) plot(kind=’scatter’) functions were used on the majority of features to determine whether or not these would be indicative of sale price for the final model. The majority of variables showed slight differences, which suggested they be tested for inclusion in the final model. For example: weekends had less sales than weekdays; less sales occurred on state holidays; store type ‘b’ had higher sales on average than a,c, or d; stores running promotions generated higher sales; and the four stores with the top sales all had low Competition Distances. Promotion Interval Type was one of the few features that did not appear to have an impact on sales.

**Feature Creation to Support a Classification Model**

The Rossman sales data is affected by seasonality and is not linear. Without significant fine-tuning and adding in a time series component, a linear regression model will poorly predict six weeks of future sales. A KNN classification model was instead chosen. KNN was chosen over Naïve Bayes as there may be highly correlated attributes in the data set that would be better handled by KNN. New features were created to support a classification model as well as enhance the model’s overall accuracy.

New features included:

* Response variable ‘Sales’ was created into bins at intervals of $500 and $1000 using the pd.cut function. These bin were chosen after reviewing descriptive sales statistics indicating the mean of $5780 and standard deviation of $3784
* Promo Week and Promo Year were used to create a promotion start date. By default the first day of the week was used in this calculation, based on the isoweek standard.
* Sale Month and Sale Week were extracted from Date using the .dt.month and .dt.weekofyear function
* Cumulative Sales and Cumulative Number of customers
* Sale Since Promo Started as calculated by the date difference between the current sale date and the date the promotion started. These date differences were then classified as integers. All negative numbers (i.e. the promotions started after the sale date) were reclassified as 0.
* School Holidays were augmented to factor in State Holidays. Rossman’s data dictionary indicated that all schools were closed on state holidays
* Weekend flag for when the Day of the Week equaled Saturday or Sunday
* Dummy features for StoreType, Store Assortment and State Holiday

**Feature Selection and Model Evaluation**

Over 20 combinations of features were evaluated to determine the best model prediction. Features were added and subtracted based on intuition from the data exploratory phase. The KNN model with the highest accuracy (.379 using the $500 sale bins, k=10, .506 using the $1000 sale bins, k=12) included the following features: Store, Day of Week, Store Type (via dummy variables), Store’s Assortment (via dummy variables), Sale Month, School Holiday, State Holiday, Promotion and Year Competition Opened. Each model was evaluated at different values of k {10,12,15,20,30,60} to determine the best accuracy. This model outperforms the null accuracy (.17). The RMSE is 3167.

**Next Steps**

Prior to the submission of the final paper I plan to continue on the model evaluation with other metrics/ cross-validation. I would also like to test the Naïve Bayes and Decision Tree models to determine if a more predictive model can be generated. Additionally, decision trees may help refine features for the KNN model.