DSE220 final project

Sale Forecasting and Targeted Marketing

Team Sharknado

Alex Egg, Deepthi Mysore Nagaraj, Mai Huynh, Peyman Hesami

Introduction

Supply chain management is one of the most crucial part of managing any store. To survive in nowadays competitive market, stores need to leverage smart strategies to target customers.

Business Understanding

Demand Forecasting

Grocery retailers must strive to meet daily consumer demand for fresh food. We propose a model to accurately forecast inventory demand based on historical sales data. This model will also help us to understand the effectiveness of our promotions by seeing the result on sales.

Targeted Advertising

We propose to use customer demographic and sales information to build targeted advertising. Targeted advertising can increase conversions on marketing efforts.

Data Understanding

Transactional Data

We have a collection of 11 years of transactional data from grocery/drug stores across the country.

Panel Data

We have data conducted from panel focus groups where a customer's trip to the store is documented by: time, retailer, product purchased, dollar amount, quantity. This data is limited to 2 regions: Eau Claire, WI and Pittsville, WI.

Demographics

For each panelist we also have a collection of information about their household:

- hh_lang
- occupation_code_of_household_head
- occupation_code_of_female_hh
- occupation_code_of_male_hh
- education_level_reached_by_household_head
- female_working_hour_code
- education_level_reached_by_female_hh
- male_working_hour_code
- education_level_reached_by_male_hh
- age_group_applied_to_household_head
- household_head_race
- race3
- all_tvs
- marital_status
- combined_pre_tax_income_of_hh
- children_group_code
- device_type
- age_group_applied_to_male_hh
- cabl_tvs
- family_size
- age_group_applied_to_female_hh
- iri_geography_number
- number_of_dogs
- number_of_cats
- type_of_residential_possession
- fem_smoke

Data Preparation

Our primary interest is in the relationship between product sold and the demographics of the customer. In order to facilitate this analysis, we need prepare the data.

Postgresql Import

The original form of the IRI data set bodes well to a relational database. All the files are in the form of tables and have keys linking to other tables. We loaded all the tables into postgres using a set of import scripts. The data we want to work with is spread across 3 different tables. This degree of normalization is not necessary for our modeling task, so we merged the: stores, panels, and products tables into one: panels_stores_products table. In addition to denormalization, we also made the business decision to consider only data from years 8 - 11 due to data quality concerns. In the end we had a denormalized table w/ a reduced dataset with 4 years of data.

This preprocessing work allowed us to iterate quickly on our modeling b/c we were able query the database once, cache the result locally (csv, etc) and then run code experiments as much as we wanted. Other alternatives such as running our code on a spark cluster would have slowed down our feedback loop and lowered our iteration speed.

Cleaning and data exploration (Also part of Assignment 1 - Q3)

Files:

Assignment 1 - Q3 - Using transaction data.ipynb

Assignment 1 - Q3 - using panelist data and KNN (extra credit part).ipynb

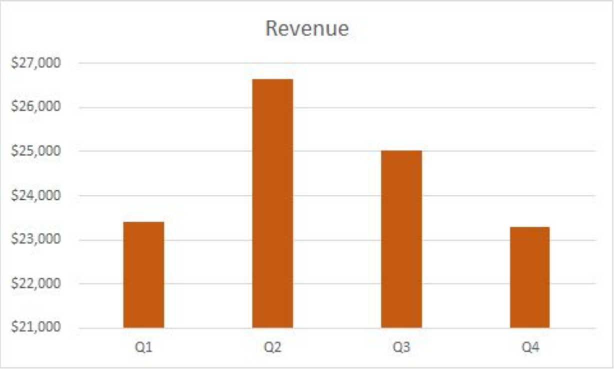
We looked at 8-11 years of data (both panelist and transaction). This exercise was done to ensure that panelist data is indeed a snapshot of transaction data.

During our initial exploration on transaction data, we found that POTATO CHIPS, TORTILLA/TOSTADA CHIPS and CHEESE SNACKS are the top 3 popular items during the period 2008 and 2011. Surprisingly, 90% of these items were from PEPSICO INC.

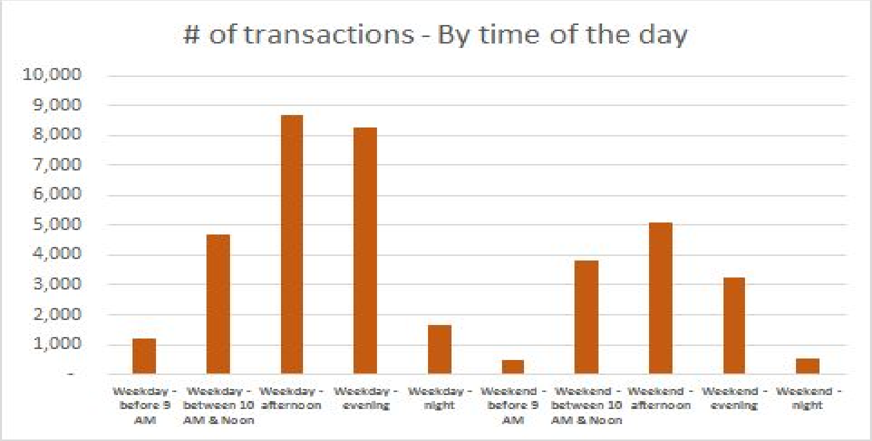
Similar insights were obtained on panelist data as well. POTATO CHIPS and TORTILLA/TOSTADA CHIPS were the top selling products year after year.

Year	Product Category	Units	Dollars	Rank
8	POTATO CHIPS	3,937	$8084	1
8	TORTILLA/TOSTADA CHIPS	2,723	$6898	2
9	POTATO CHIPS	4,038	$9842	1
9	TORTILLA/TOSTADA CHIPS	2,656	$6564	2
10	POTATO CHIPS	4,357	$9857	1
10	TORTILLA/TOSTADA CHIPS	2,365	$5951	2
11	POTATO CHIPS	4,169	$10187	1
11	TORTILLA/TOSTADA CHIPS	2,406	$6061	2

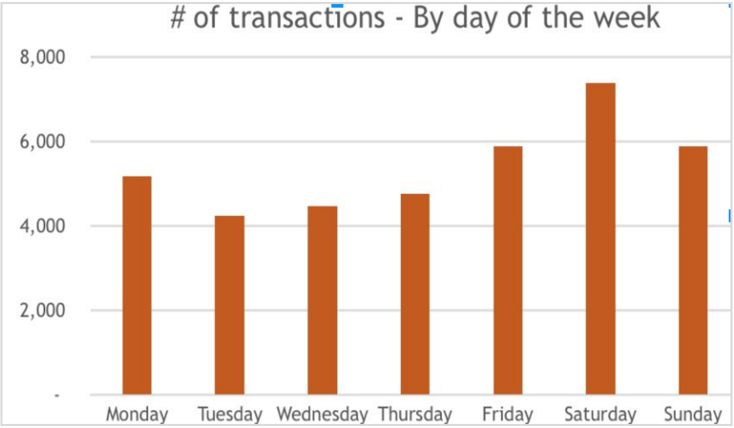
Salty snacks were most popular in Q2.



Salty snacks are mostly purchased on a weekday afternoon. We hypothesize that this is mostly by stay-at-home moms.



Saturdays are heavy on salty snack sales as well.



K-Nearest Neighbors Model (KNN)

File: End of the file hyperlinked below:

Assignment 1 - Q3 - using panelist data and KNN (extra credit part).ipynb

We applied K-nearest neighbor algorithm to the salty snacks data to estimate the units.

Data Cleaning:

1. Replaced NaNs with 0.
2. Categorized 99, 98 and 7 under the NA category (0)
3. Converted non number categories to numbers

Feature used:

age_group_applied_to_female_hh, age_group_applied_to_male_hh, education_level_reached_by_female_hh, education_level_reached_by_female_hh, occupation_code_of_male_hh, occupation_code_of_female_hh, est_acv, events, colupc

Feature Engineering:

Didn't apply any specific feature engineering method here.

Target:

Units

Models used:

K-Nearest Neighbor

Evaluation:

Obtained an MSE of 0.39

Modeling

In this section, we have described different models that we tried and their corresponding results. We have tried to stick to CRISP-DM format for each model.

Original Model

File: 01_Original_Model.ipynb

Data Cleaning:

See cleaning steps above.

Feature used:

Week, minute, outlet, est_acv, marketname, open

Feature Engineering:

Created the features like season, month, hour, time of the day (morning, noon,...) using minute and week flags in the transaction data.

Target:

We used product UPC codes for our targets.

Models used:

We used a random forest classifier

Evaluation:

We trained our initial model against a product UPC code and all demographic and store features. This is like saying: If a person from a household of demographic D walks into store S at time T what exact product will they buy. Using a random forest classifier w/ all the features scored %20 accuracy.

Model 2

File: 02_model.ipynb

Data Cleaning:

We used the same data as original model. So all the data cleaning steps are applicable here.

Feature used:

We used the same features as the original model. Namely:

Week, minute, outlet, est_acv, marketname, open

Feature Engineering:

Created the features like season, month, hour, time of the day (morning, noon,...) using minute and week flags in the transaction data.

Target:

Changed the target from colupc (sku) to L2 (product category). This reduced the total number of classes from 3616 to only 8. Following are the 8 classes:

- TORTILLA/TOSTADA CHIPS
- PRETZELS

- PORK RINDS
- READY-TO-EAT POPCORN/CARAMEL CORN
- POTATO CHIPS
- CHEESE SNACKS
- OTHER SALTED SNACKS (NO NUTS)
- CORN SNACKS (NO TORTILLA CHIPS)

Models used:

As with original model, we used Tuned Random Forest Classifier.

Evaluation

Using 10 fold cross validation, we were able to achieve a score of 32% which was a good improvement from 20% that we saw in our original model.

Model 3

File: 03_Model.ipynb

In this model, we introduced demographic data.

Data Cleaning:

Data cleaning involved the following steps:

1. Converting the numbered categories in text format to numeric
2. Replaced the values 99, 98 and 7 and grouped them under N/A category

Feature used:

minute, week, combined_pre_tax_income_of_hh , family_size, age_group_applied_to_male_hh, age_group_applied_to_female_hh, education_level_reached_by_female_hh, education_level_reached_by_male_hh, occupation_code_of_female_hh, occupation_code_of_male_hh, all_tvs, cabl_tvs

Feature Engineering:

1. Created new features using minute and week flags to indicate the transaction time:
    a. Time of the day (using minute)
    b. Day of the week (using minute)
    c. Season (Quarters or Spring to Winter) (using week)

2. Using Demographic data (By combining M&F HH values):
    a. Income per person
    b. Age group
    c. Education

d. Occupation
e. # of TVs (using # of TVs and # TVs hooked to cable)

Target:

Used L2 (product category) with 8 classes:

- TORTILLA/TOSTADA CHIPS
- PRETZELS
- PORK RINDS
- READY-TO-EAT POPCORN/CARAMEL CORN
- POTATO CHIPS
- CHEESE SNACKS
- OTHER SALTED SNACKS (NO NUTS)
- CORN SNACKS (NO TORTILLA CHIPS)

Models used:

Decision Tree Classifier

Evaluation:

Using 10 Fold cross validation we obtained a score of about 42%.

Model 4

File: 04_Model.ipynb

Data Cleaning:

Same as model 3

Feature used:

In efforts to boost our accuracy we attempted to reduce the amount of features by using the SelectKBest algorithm. We were training with up to 62 features, some of which did not contribute much information for the model. In order to help prune down our feature set we employed the SelectKBest algorithm with ANOVA F-value scoring between label/features. Using this routine we were able to score a 10-fold crosvalided score of 45% using only 16 features.

Feature Engineering:

None

Target:

We used the same target as the previous model: l2 in the product dataset

Models used:

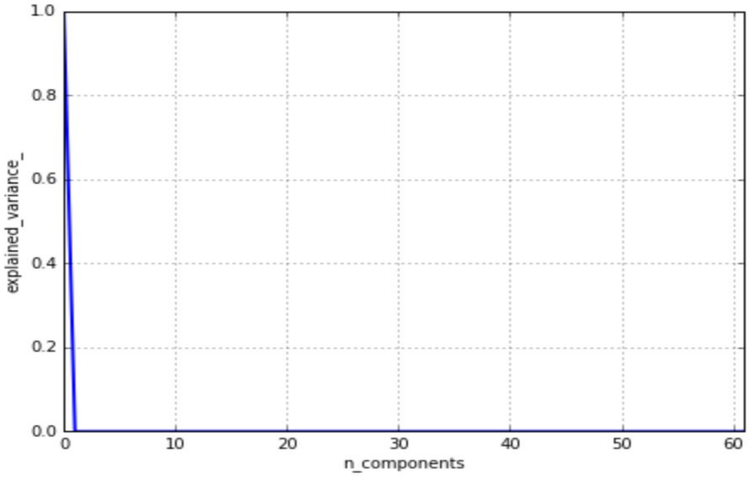
We used the grid search routine to find the best parameters for our Random Forest classifier: 16 features and max depth in the classifier.

Evaluation:

Using 10 fold cross-validation, we were able to achieve a score of 45%

Aha Moment

We were dazzled by the low performance of our classifiers (low score on both training and test set) despite all the complex models and feature engineering that we did. So next we did PCA analysis to understand the nature of the dataset. We realized that 99.99% of the variance of the data is focused in the first dimension/eigenvector. And we compared that one dimension correlation with our target output and realized that the correlation is very low and that's the main reason of our low performer models.



Final Model

File: 05_Final_model.ipynb

After our Aha moment where we realized that the data is one dimensional and limited in nature, we started looking at the transaction tables for drug/grocery stores which had more data and slightly modified our goals for our business objective (as our new dataset was slightly different). Here is our new modified business goal:

- Forecast the sale (quantity or Dollar value) of different products based on store characteristics, the marketing strategies (display size, coupon, ...), and time of the day/month and study the impact of promo on sale!

Data Cleaning and Feature Engineering:

We used transaction tables for both drug and grocery stores on salty snacks, coffee, and sugar substitute. We aggregate the final table at year, season, month level and created the following new features and Targets:

- Feature: # of Promotions given out on each product category by the store
- Feature: Display Size Sum which signifies the sum of display size used on each product category
- Feature: Ad importance
- Target: The most popular item within each product category

Ad importance was one of the key engineered feature which boosted the performance of our models by 10%-15% and was created by assigning a relative weight to each Feature (F) based on their importance and our intuition. For example we assigned higher weights to A+ (QR code and coupons) compared to C (small size texts). We also used the product and IRI Week Translation Table and aggregate all the three tables above and created new features for season and month from week.

Outlier Detection and Removal

We assume that all features have Gaussian distribution and detect outliers based on the following rule:Detect a column with extreme outlier if:

$$mean(col) - 10 * \sigma < value < mean(col) + 10 * \sigma|$$

Models used:

We created the following models (can be used for yearly, seasonal, and monthly data):

- Classifying the most popular item within each category (L2)
- Forecasting the sale of each product (DOLLARS)
- Predicting the number of units sold for each category (UNITS)

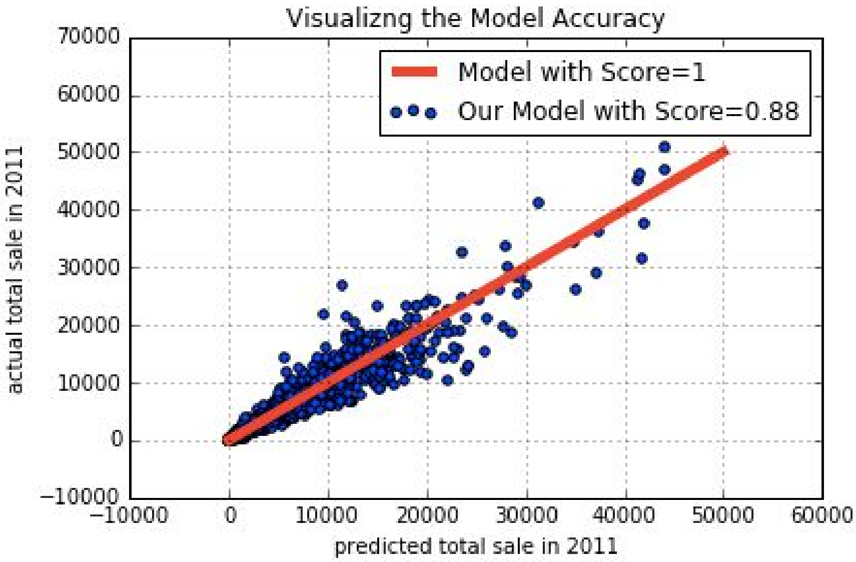
We used RandomForestRegressor and RandomForestClassifier and used GridSearchCV to tune the parameters. We also tried a 2-level stacking with 3 base classifiers/regressors and 1 meta classifier/regressor to see if it can improve the accuracy of our models. We tried all combinations of several classifiers/regressors and picked the one with the best score but realized that stacking didn't help to increase the performance of our models whenever the accuracy of our model was already high (for example score of 0.88 for yearly model). However, for monthly model where the performance of our model was slightly lower (score of 0.7) it boosted the score of our model by 2%-5%.

Evaluation:

While we used 10 fold CV for evaluation our models 0,1,2, and 3, for evaluating the performance of our final model, we hold out the 2011 data and used it as a test set while training our model on the data for 2008-2010. Please note that if we use 10 fold CV, we get even higher scores (2%-3%) but we believe for time varying target, using the future data set as test set makes more sense. Here are the score results for our models on 3 different targets and 3 different time level of forecast.

Target Model Type	Total Sale (Dollar)	Number of Units	Most Popular Item
Annual	88%	86%	98%
Seasonal	81%	80%	97%
Monthly	72%	71%	95%

And here is a visualization of our model accuracy by plotting the predicted values of the sale (DOLLARS) by the model versus the actual values of sale. The closer the blue dots are to the red line in the below plot, the more accurate the model.



Deployment

How to use these analysis/results to get business insights?

- Stores can use the sale forecast model for yearly/seasonal/monthly sale forecast of any product. This can help them to manage their supply chain efficiently
- Analysing the forecasted sales in combination with the marketing strategy can help boost sale of specific products through targeted marketing