

Sale Forecasting and Targeted Marketing

using IRI Marketing Dataset

Team Sharknado:
Alex-Deepthi-Mai- Peyman

DSE 220 Final Project



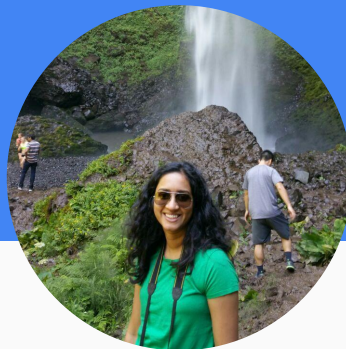
The Team



Mai



Alex



Deepthi

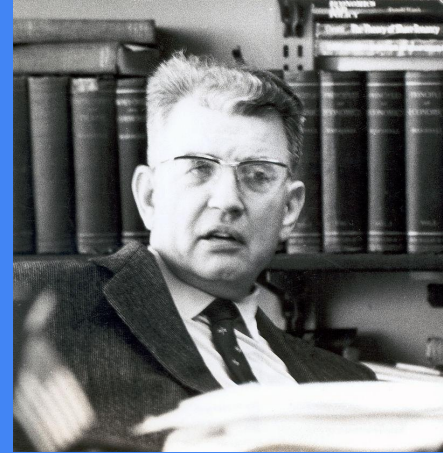


Peyman

Lesson Learned

“ If you torture the data long enough, it will confess”

Ronald H. Coase



ENOUGH

“ If you torture ~~data~~ long enough, it will confess”

Ronald H. Coase+Team Sharknado

Initial Business Objectives

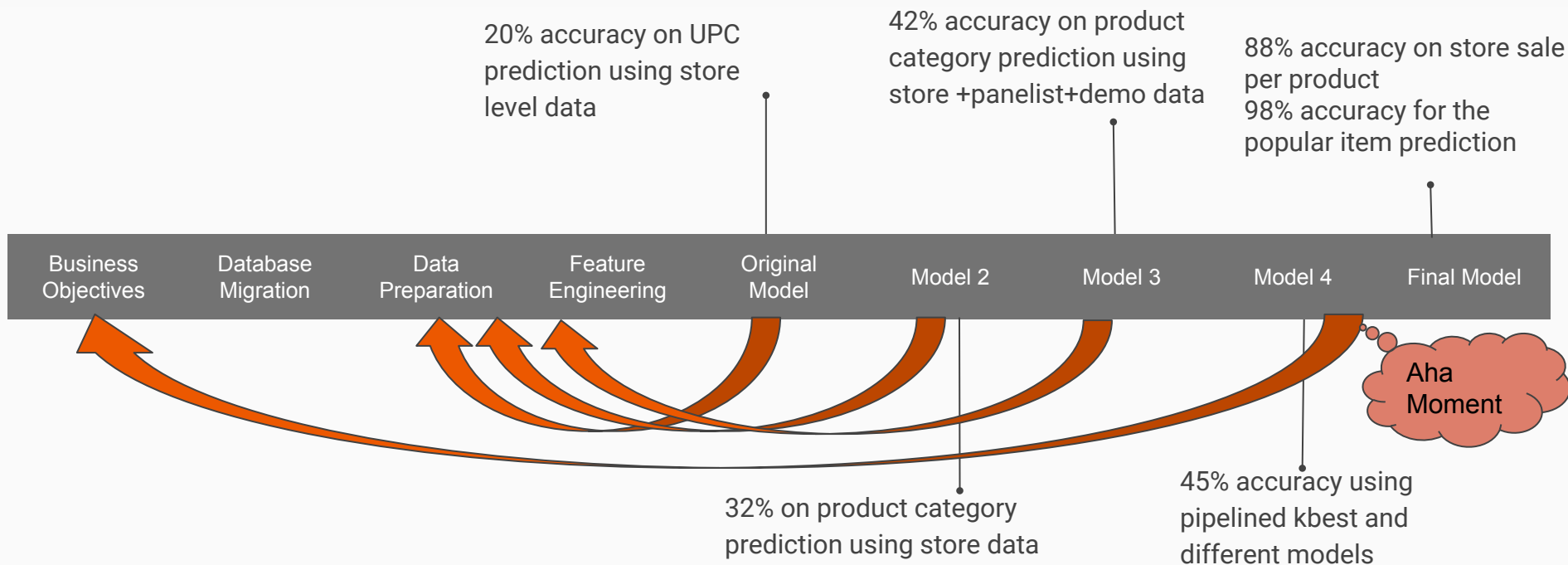
Supply chain management is one of the most crucial part of managing any store.


To survive in competitive market, stores need to leverage smart strategies to target customers.

- Creating a *model of consumer demand* to *forecast future sales (product type and sales value)* for a subset of stores and for a *specific time of the month/day*, and for a *specific type of demographics*.



The journey





Data Preparation

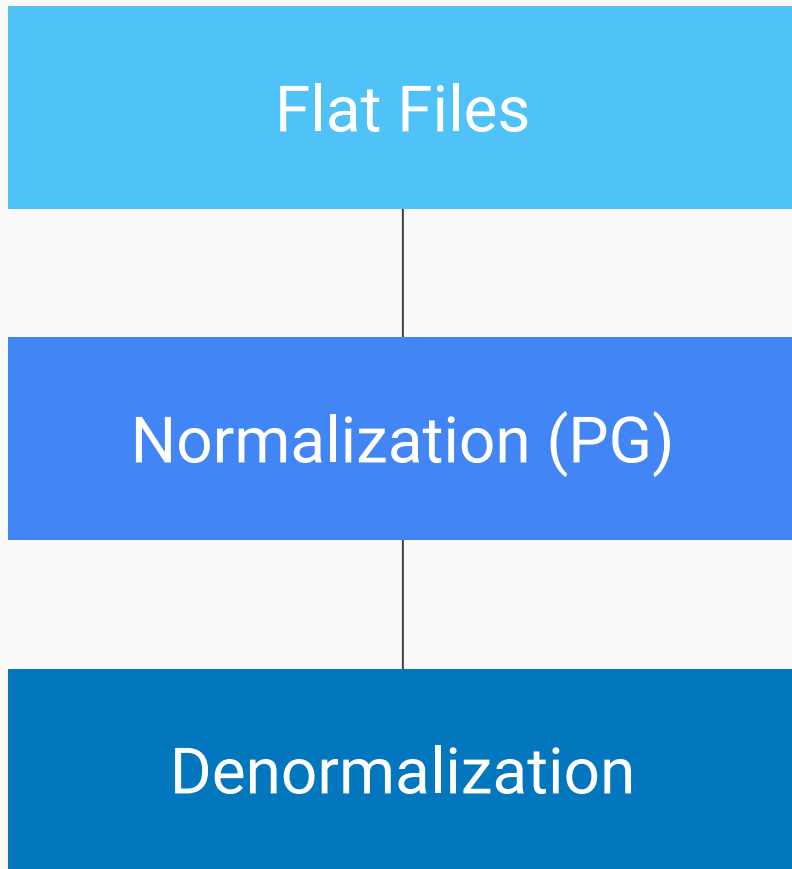
Migration to AWS

Data Preparation: Database Migration

We leveraged a PostgreSQL database hosted on Amazon AWS.

The normalized database tables allowed quick experimentation

We denormalized the database tables to remove bootstrap code from notebooks





Data Exploration

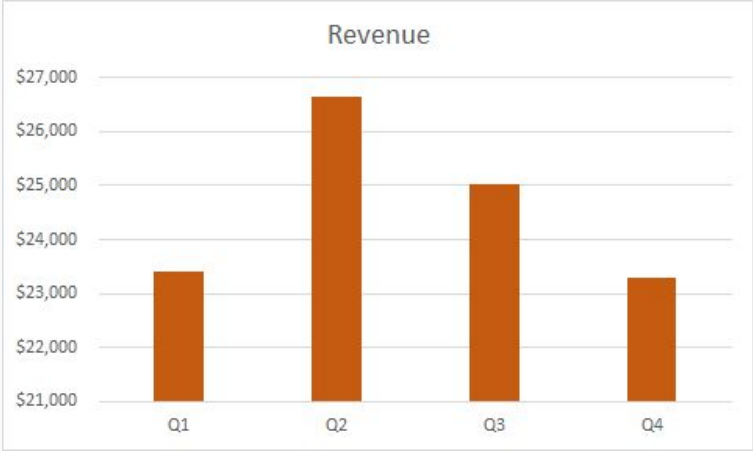
Exploring the products and
demographic data

(Weka, ipython Notebook, and
Excel)

Potato chips is the most popular product across years. Followed by Tortilla/Tostada Chips. Salty snacks are most popular in Q2

Note: We found similar trends on the transaction data as well

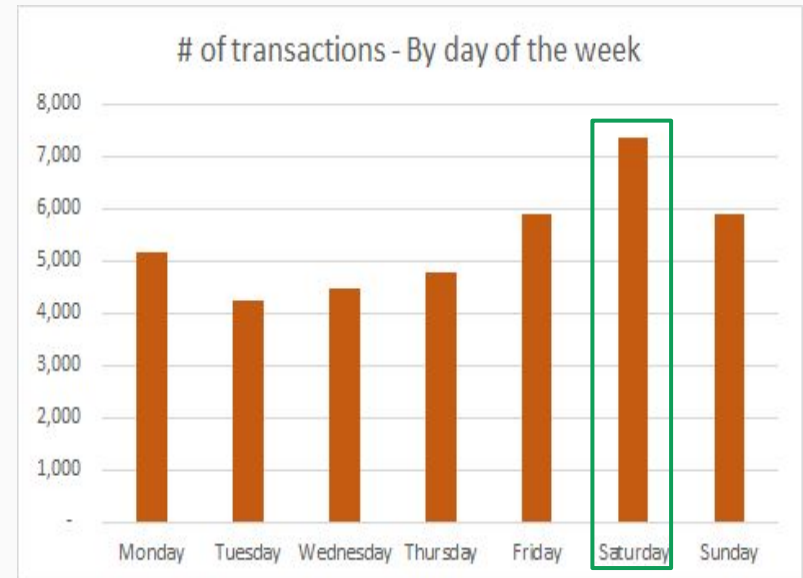
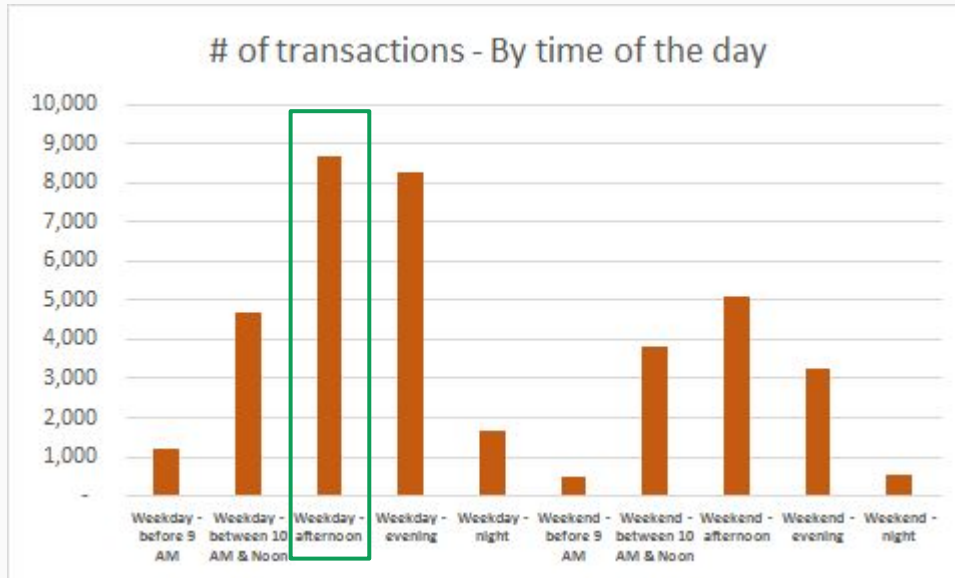
Year	Product Category	Units	Dollars	Rank
8	POTATO CHIPS	3,937	\$8084	1
8	TORTILLA/TOSTADA CHIPS	2,723	\$6898	2
9	POTATO CHIPS	4,038	\$9842	1
9	TORTILLA/TOSTADA CHIPS	2,656	\$6564	2
10	POTATO CHIPS	4,357	\$9857	1
10	TORTILLA/TOSTADA CHIPS	2,365	\$5951	2
11	POTATO CHIPS	4,169	\$10187	1
11	TORTILLA/TOSTADA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



Data exploration (Panelist data)

Salty Snacks are mostly purchased on a weekday afternoon.
Mostly by stay-at-home moms? Also, Saturdays are heavy on salty snack sales

Note: We found similar trends on the transaction data as well





Modeling and Evaluation

Feature Engineering

Target Selection

Model Selection/Tuning

Evaluation Criteria

panelist transaction+delivery store data

Features: week, minute, outlet, est_acv, marketname, open

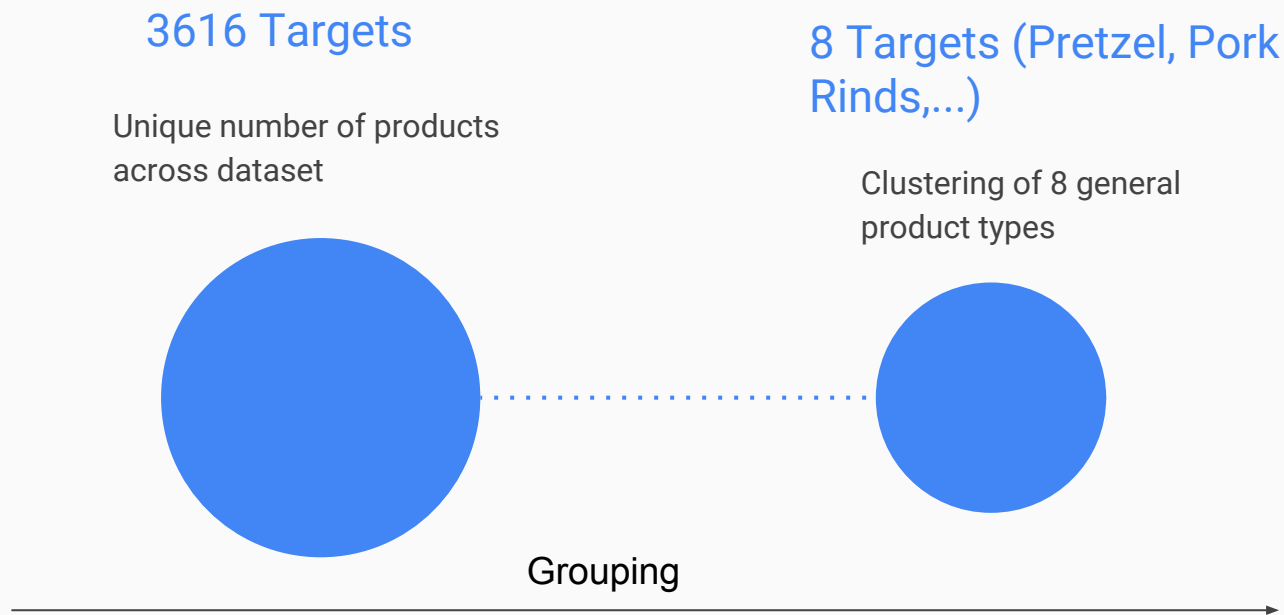
Features engineered: season, month, hour, time of the day (morning, noon,...)

Target: colupc (sku)

Model: Tuned Random Forest Classifier

Evaluation: 10-fold CV score of 22%

Caveat of the original model: Too many classes



panelist transaction+Product+delivery store data

Features: week, minute, outlet, est_acv, marketname, open

Features engineered: season, month, hour, time of the day (morning, noon,...)

Target: Changed target from colupc to product category

Model: Tuned Random Forest Classifier

Evaluation: 10-fold CV score of 32%

Caveat of the model 1: not enough data/features

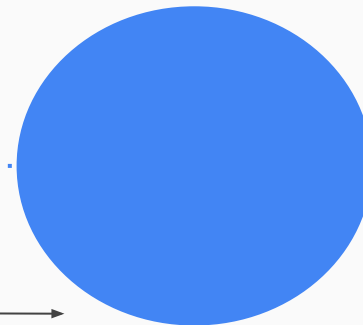
6 Features

From transaction+store table



11 Features

From transaction+store+demo table



We switched gears a little and decided to see if we can predict the product for an individual (person) given the demographic details and added the demographic features.

panelist transaction+delivery store+demo data

Data Preparation:

1. Used panelist data for the years 8-11 as transaction time was found to be inaccurate for years 1-7
2. Used panelist demo, transaction and week translation data
3. Converted the numbered categories in text format to numeric
4. Replaced values 99, 98 and 7 and grouped them under N/A category

Modeling and Evaluation:

1. Used a Decision Tree Classifier and got accuracy of 42% on prediction the product category
2. Tried a few stacking methods on this model but didn't see much improvement

Features Engineered:

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 # TVs and # TVs hooked to cable)

panelist transaction+delivery store+demo data

Features/Features engineered/Target: Same as Model 2

Model: Tried SelectKBest ANOVA F-value scoring and Random Classifier

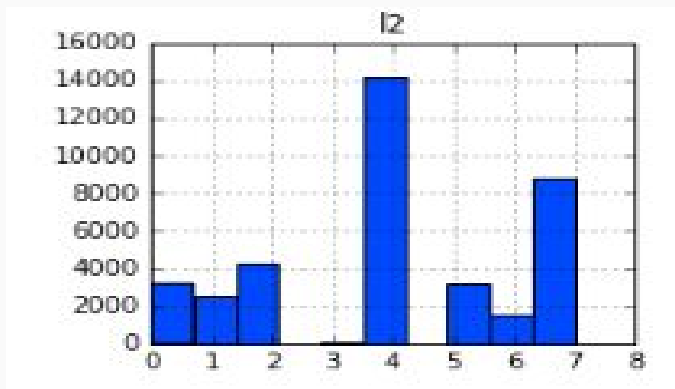
Selected Features: Family_size, panelist_type, fipscode, iri_k, zipcode, age_group_applied_to_female_hh, age_group_applied_to_male_hh, iri_geography_number, combined_pre_tax_income_of_hh, ou_GK

Evaluation: 10-fold cross validation score of 45%

Done?!

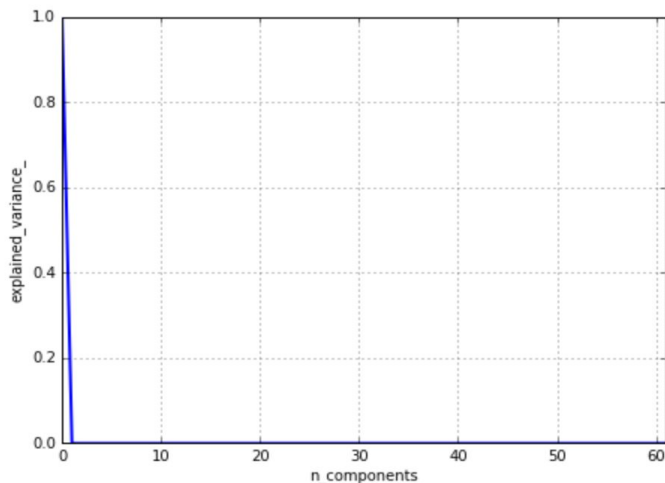
Not really!

A Not-So-Smart classifier that always outputs the most popular item (potato chips) has 42% accuracy!



Aha Moment!

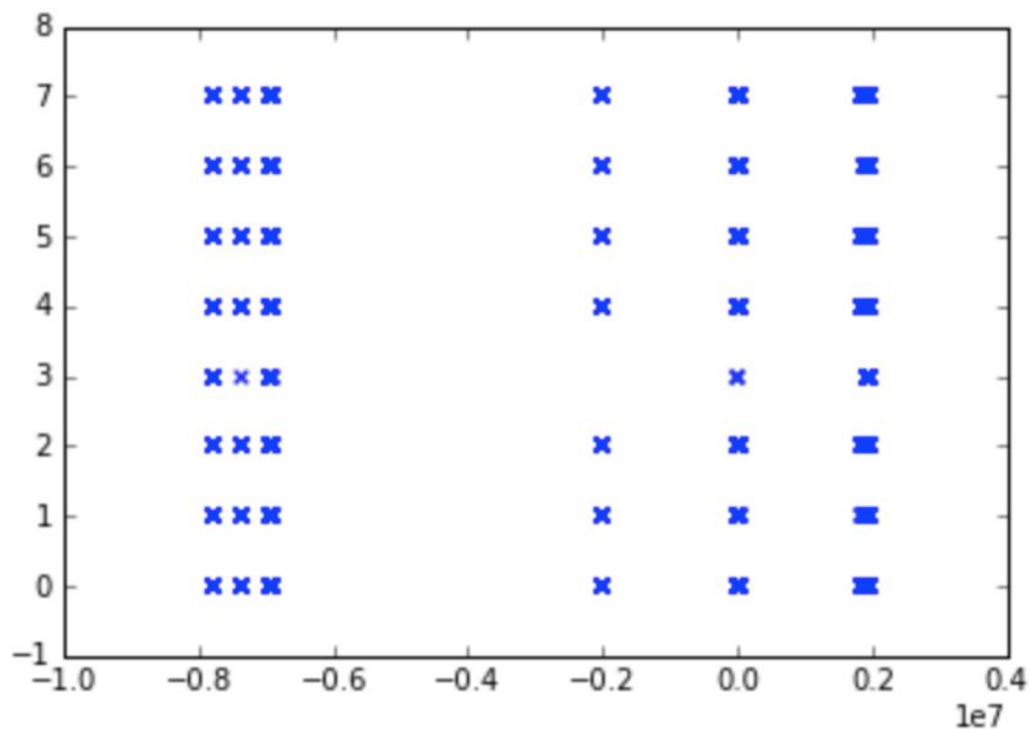
We did PCA on our data set and realized *99.8% of the variance is focused on the first dimension/eigenvector*



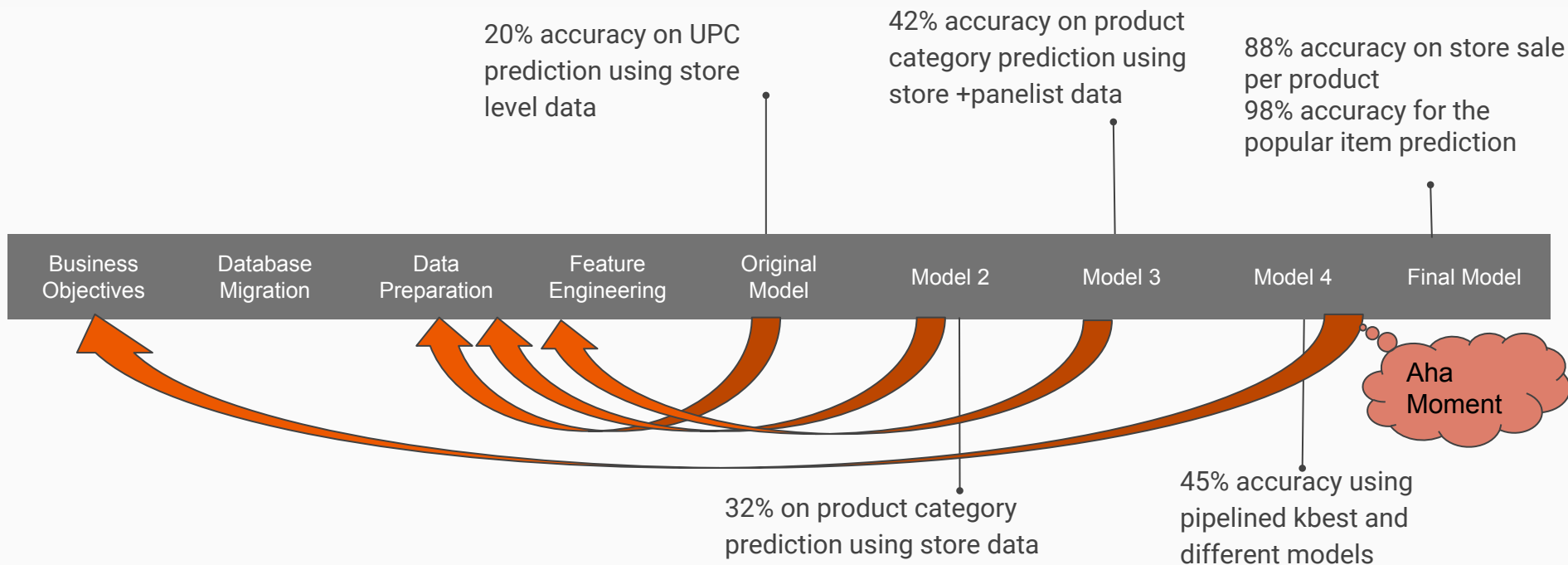
and that one dimension doesn't have much correlation with the target!!!



```
In [79]: _=plt.scatter(X_reduced, target, marker="x")
```



The journey



New Objectives + New Data

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!

Use the main transaction table instead of limited panelist transaction table

General transaction+product data

Features: week, Feature (type of ad), D (Display size), PR (promotion), L1 (product category)

Features engineered: season, month, total number of promo per category, Average Display Size, ad importance (assign a weight to each type of ad)

Outlier Detection: Assuming Gaussian distribution for the futures (based on their statistics), we removed any extreme outlier value which does not satisfy the following property:

$$\text{mean}(\text{col}) - 10 * \sigma < \text{value} < \text{mean}(\text{col}) + 10 * \sigma$$

General transaction+product data

Target:

1. sales per product category
2. number of unit sold per product category
3. most popular product in each product category

Model: Random Forest Regressor/Classifier

- Yearly
- Seasonal
- Monthly

Stacking:

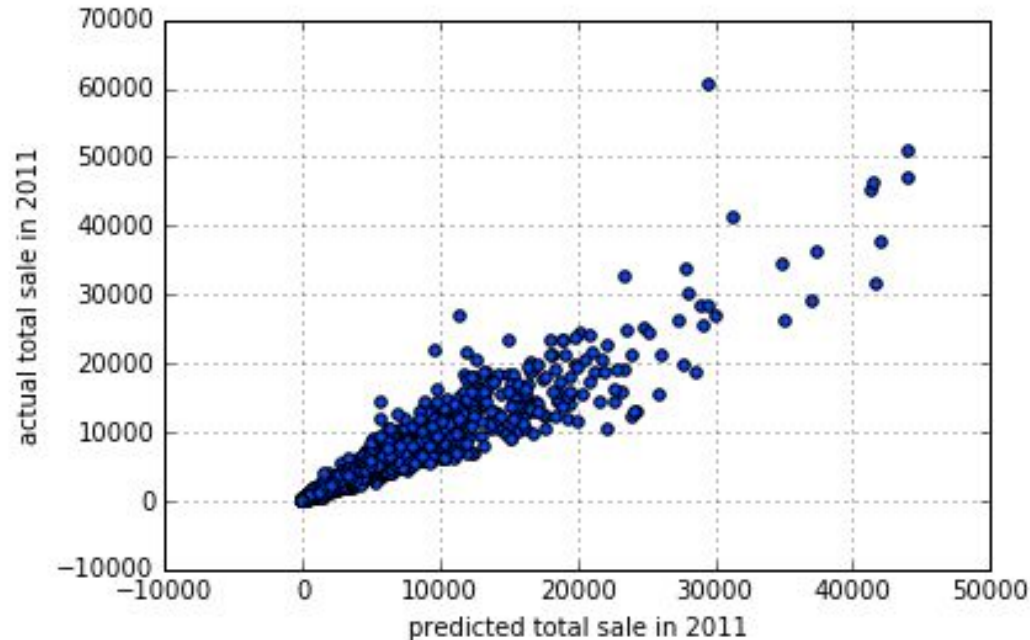
- base (Logistic Regression+Decision Tree+Ridge) + meta (KNeighbor)-->Classifier
- base (Lasso+Random Forest+BayesianRidge) + meta (Extra Tree)-->Regressor

General transaction+product data

Evaluation: We trained on data from 2008-2010 (for Salty snack-Coffee-Sugar subs) and used the entire 2011 dataset as our test set

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

Final model prediction versus actual sale



Final Model: Prescriptive Analysis

Total Number of Promotions	Average Display Size	Ad Importance	OUTLET	Product Category	predicted sale	actual sale
224	5	1122	0	0	10102.708	9244.18
380	200	1838	0	1	7160.160	8208.54
60	0	211	0	2	1175.044	983.82
322	109	1196	0	0	12564.724	14778.83
172	249	1174	0	1	4805.750	5824.10
26	0	136	0	2	970.156	1074.03
35	37	331	0	0	2569.789	3083.77
167	443	1810	0	1	12472.628	15847.39
5	0	71	0	2	531.743	564.05
322	18	1321	0	0	10671.119	12077.77

2. Impact of promotions

Average \$ during promo period is 1.8x times that of \$ during no-promo period

1. Most products have a positive impact on sales due to promotions
2. Potato chips, Cheese snacks and Corn Snacks are the most popular products during sale. They generate more than 2.2x \$ during the promo period

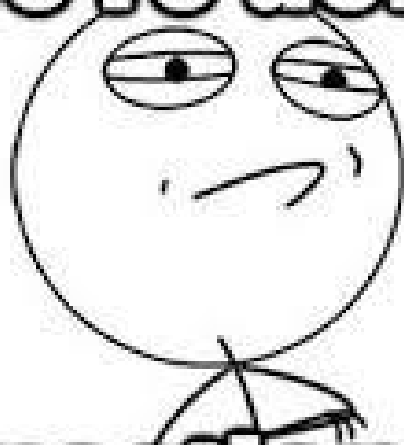
Product category	Average Revenue - With Promo	Average Revenue - Without Promo	Promo vs No Promo
POTATO CHIPS	\$49.63	\$23.03	2.2x
CHEESE SNACKS	\$44.50	\$20.97	2.1x
CORN SNACKS (NO TORTILLA CHIPS)	\$50.77	\$25.39	2.0x
TORTILLA/TOSTADA CHIPS	\$51.97	\$29.64	1.8x
OTHER SALTED SNACKS (NO NUTS)	\$31.38	\$20.42	1.5x
PRETZELS	\$23.26	\$16.60	1.4x
PORK RINDS	\$13.37	\$10.36	1.3x
READY-TO-EAT POPCORN/CARAMEL COR	\$20.29	\$17.05	1.2x

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

We're done.



Questions?

Appendix

The transaction time was found to be inaccurate for the years 1 through 7, so we are using data only for years 8 through 11

With an ***assumption that panelist data is a good sample for the analysis***, we have used the following datasets for the analysis:

1. Panelist demo data: For demography details (*demos.csv*)
2. Panelist transaction data: Transactions on a weekly basis
(*Saltsnck_PANEL_DR_strtwk_endwk.dat* and *Saltsnck_PANEL_GR_strtwk_endwk.dat*)
3. Manual store entry data: To adjust for the stores that do not report the store details (*manual store entry external 8_11.csv*)
4. Week Translation: To create seasonality flags (*IRI week translation_2008_2017.xls*)

Created new features using the available data to capture transaction time, seasonality and demographic details (Used one hot encoding where relevant)

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 (Spring/Summer/Autumn/Winter) (*Using week*)
2. Used demographic data and created the following features:
 - a. Income per person (*using combined income and family size*)
 - b. Age group (*by averaging the male and female HH age*)
 - c. Education (*by averaging male and female HH education level*)
 - d. Occupation (*by averaging male and female HH occupation level*)
 - e. # of TVs (*using total # TVs and # TVs hooked to cable*)

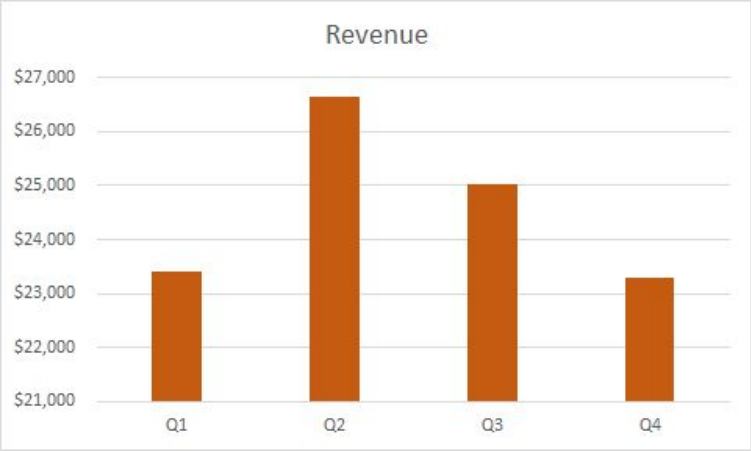
Created several models to..

1. Understand purchase propensity of a product at a particular store at a certain time of the day
2. Understand the impact of promotion on the sales
3. Understand purchase propensity of a product by a particular customer knowing the demographic details

Potato chips is the most popular product across years. Followed by Tortilla/Tostada Chips. Salty snacks are most popular in Q2.

Note: We found similar trends on the transaction data as well

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



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.
- Stores can use the sale forecast model based on the marketing strategies in prescriptive analysis to boost the sale of specific products through targeted marketing

