PREDICTING CUSTOMER BEHAVIOR ON RETAIL SALES

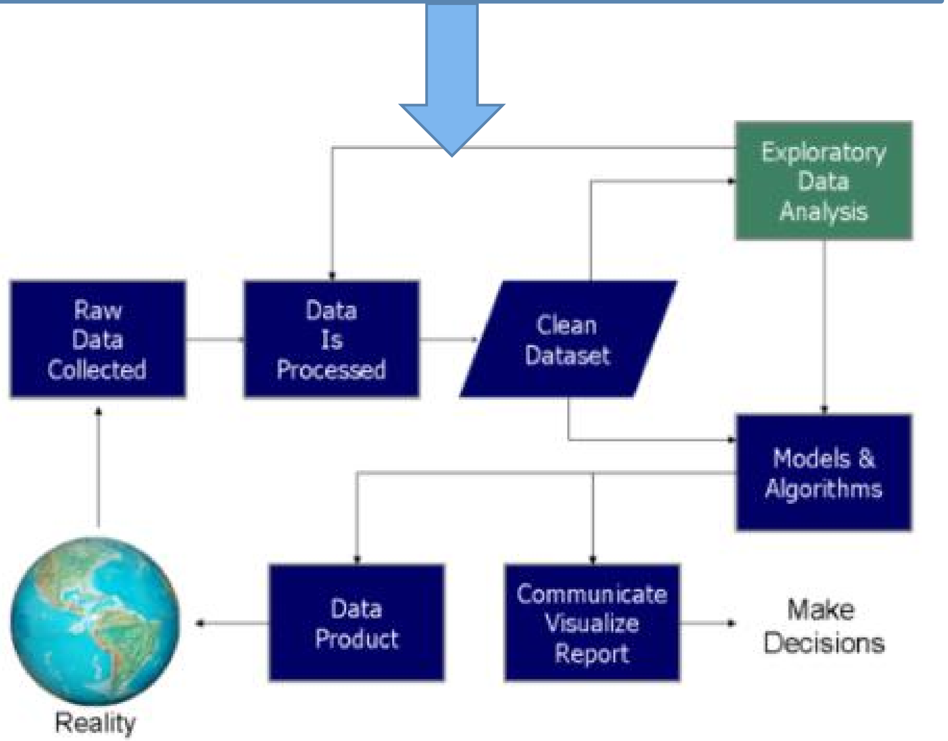
OPTIMIZING RETURNS BASED ON MARKDOWN SUCCESS

By

Angus Ogubuike

Objectives

- Evaluate Impact of
  - Holiday on sales
  - Major events that happen once a year (called markdowns) on sales
- Predict future performance based on these factors
- Optimize sales in different departments based on the features
- Answer Environmental Questions



Highlights

- The performance of the promos is dependent on the size of the stores - *The bigger the store, the more successful the promo (Markdown).*
- Success of the sales are more pronounced during routine holidays and weekends more than ordinary week days.
- While promo sales tends to behave independently from each other, sales during MarkDown 1 and MarkDown 4 have strong positive correlation
- The MarkDowns (promos) have more effect on sales of kids items and fashion items (for teens and adults) than other items.
- It is recommended that retailers pay more attention on these items (kids and fashion) during promos

Study Strategy

- Data Set Review
- Data Wrangling
- Data Exploration
- Hypothesis Testing
- Unsupervised Learning - Anomaly Detection
- Test of Multicollinearity - Variance Inflation Factor
- Clustering
- Dimensionality
- Regression
- Classification

DATASET - General

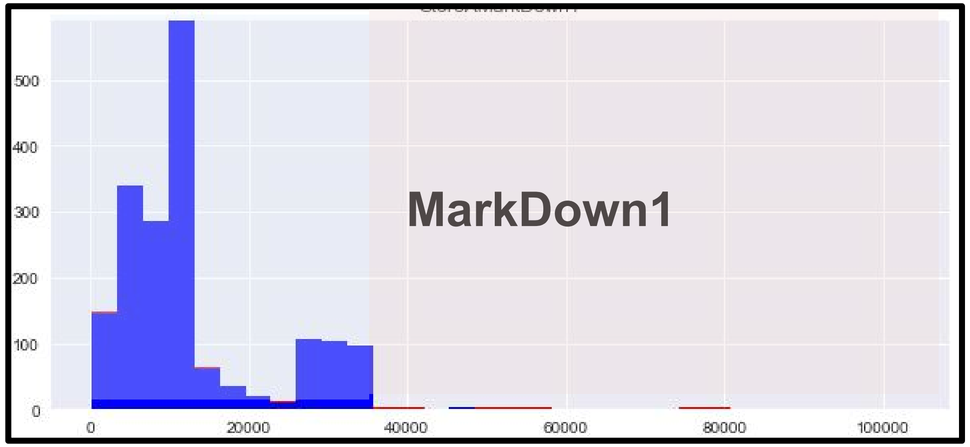
- Historical sales data for **45 stores** located in **different regions** in the United States
  - Features include
    - Departments
    - Promotional Markdowns: They precede prominent holidays. The five largest of which are the Super Bowl, Easter, Mother's Day, Thanksgiving, and Christmas
    - Environmental Variables: These are additional data related to the store, department, and regional activity for the given dates

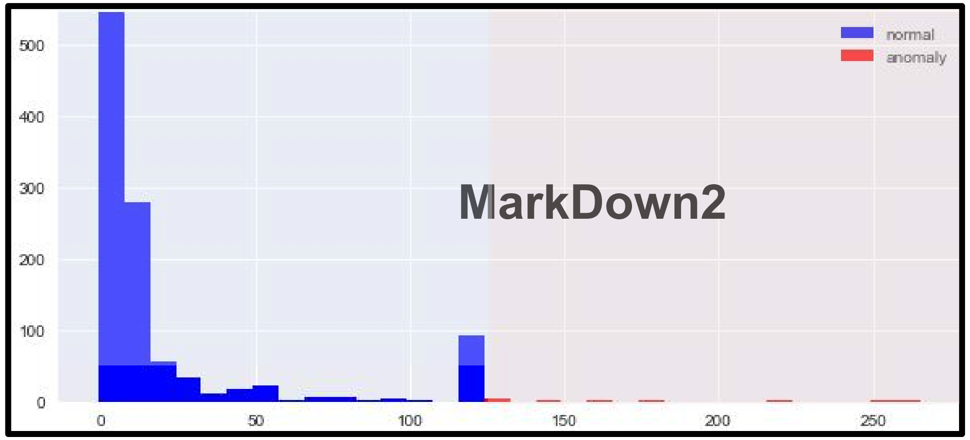
Statistical Overview

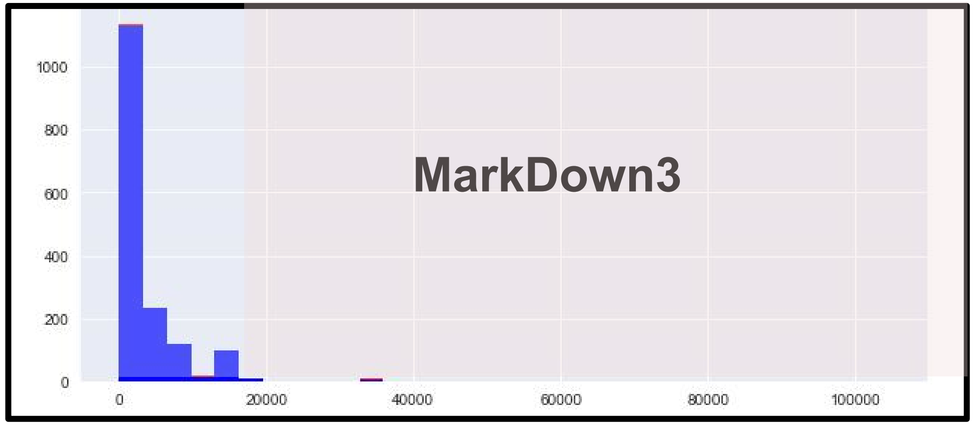
- The dataset has 8190 sales record and 95 features
- 7% of the total period of sales are holidays
- Missing values less than 2% of data set.
- Missing Values handled using interpolation method
- Promotions are generally more successful (more sales are recorded) during holidays than during non-holidays

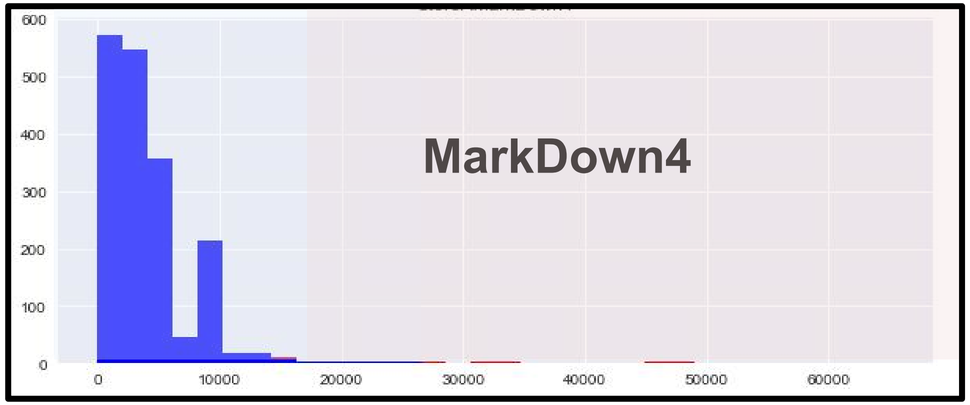
Anomaly Detection

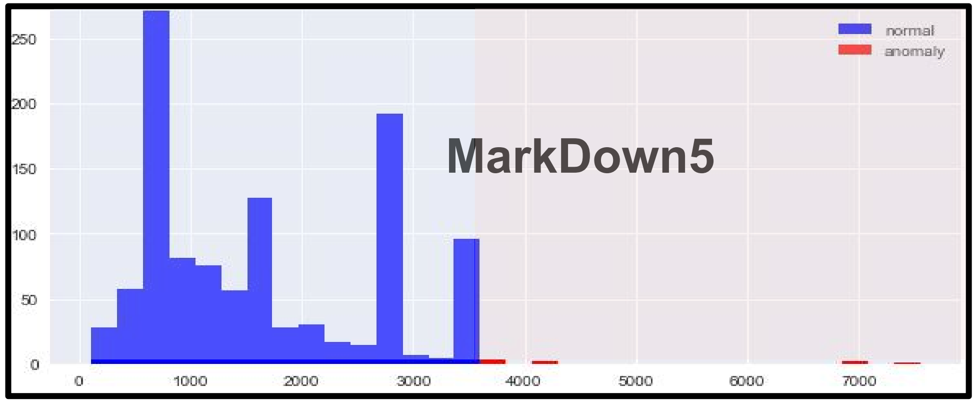
OBJECTIVE: To detect patterns in the data set that do not conform to an established normal behavior











* Anomalies < 2% of data set were removed from the data set prior to further analysis

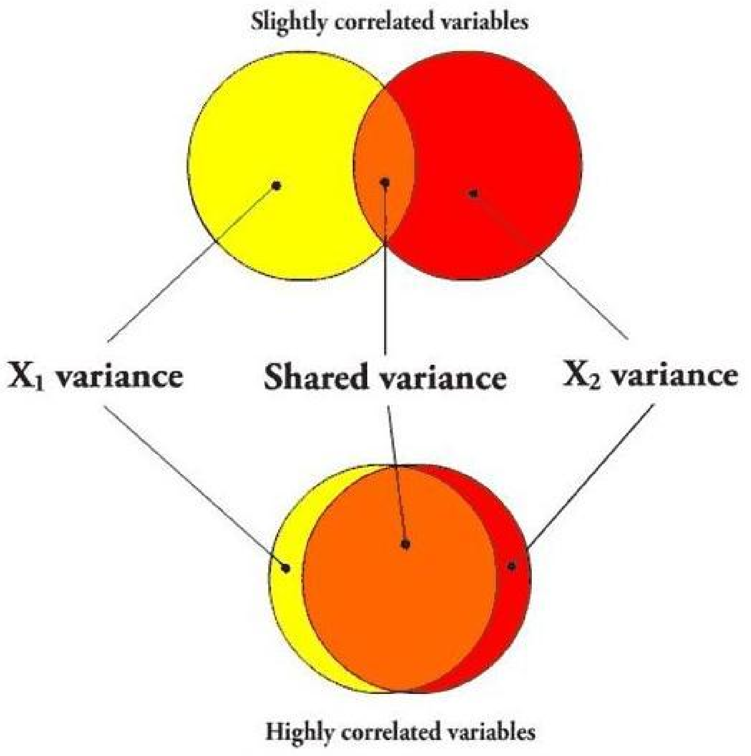
Multicollinearity - Variance Inflation Factor

- Out of 92 Features, 29 ~ 30% have VIF < 5.
- Group features with VIF > 5 into 8 Categories to create new target features

$$VIF_1 = \frac{1}{1 - R^2_{1.2...k}}$$

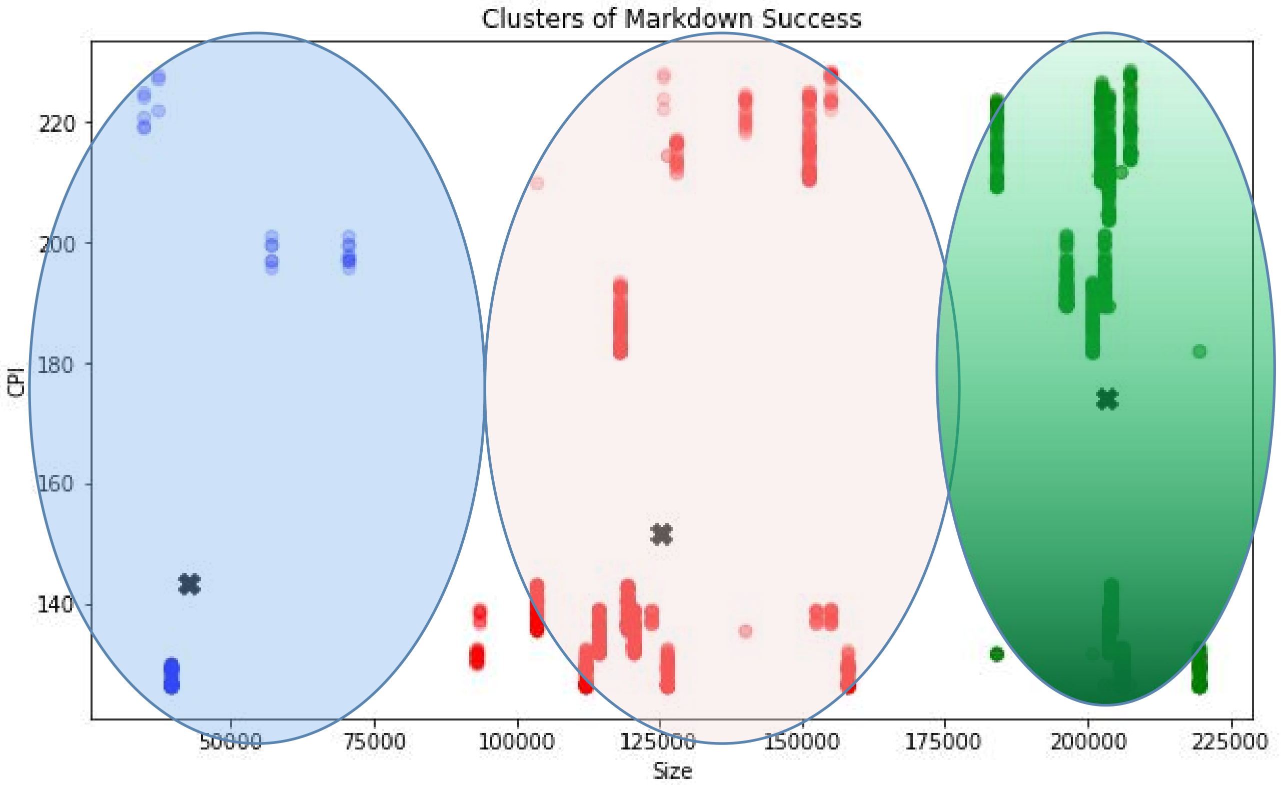
NEW TARGET FEATURES

- Home-items
- Electronic-items
- Health-items
- Kids-items
- Office-items
- Auto-items
- Wears-items
- Food-items



Clustering

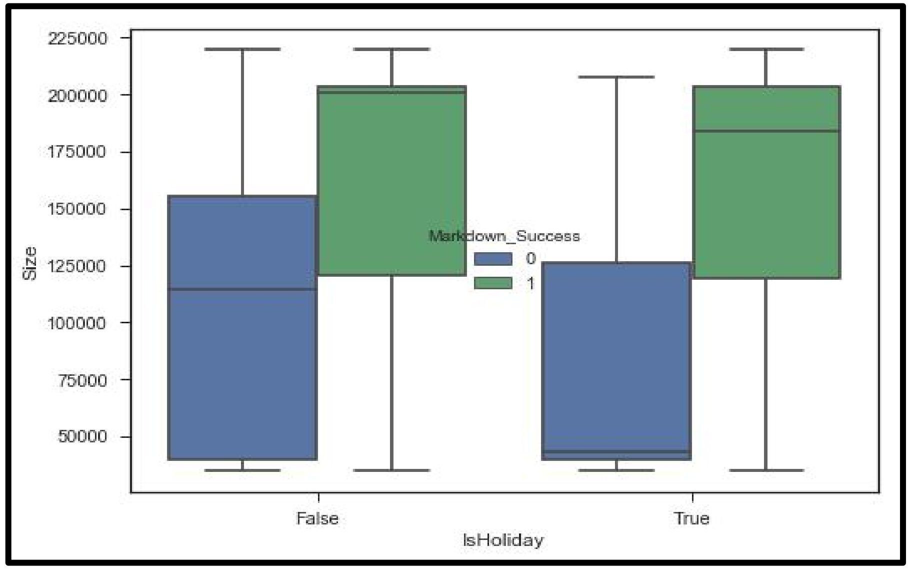
- MarkDown Success Used as Target Variable
- Used k-means clustering

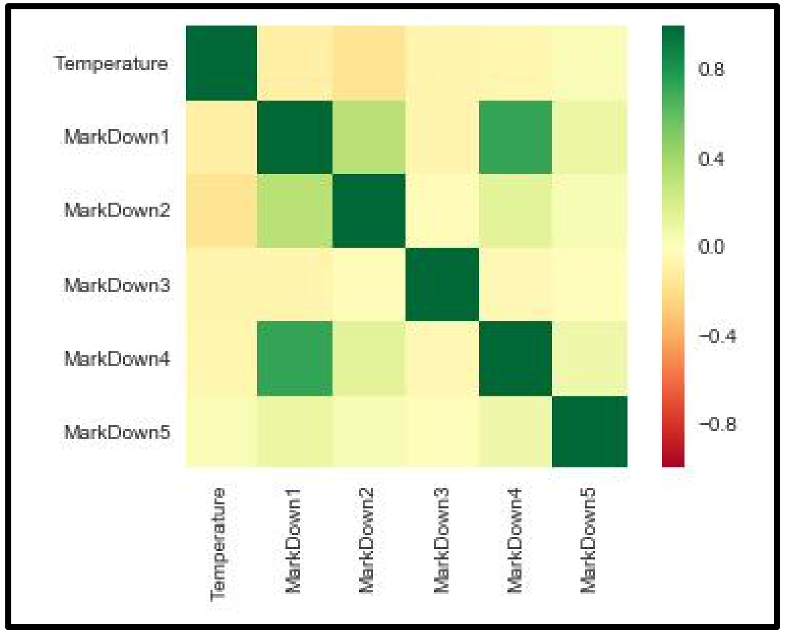


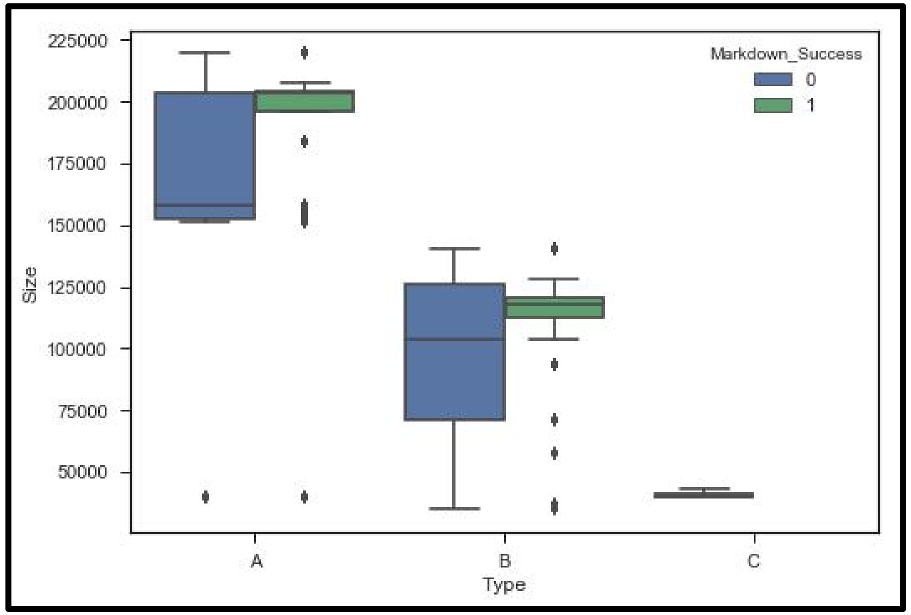
- Three Clusters distinguishable based on Store Size

Hypothesis Testing

- Markdown/ other sales are more successful during Holidays
- Irrespective of the the Store type, Store size does affect markdown success
- MarkDown 1 and MarkDown4 have strong positive correlation

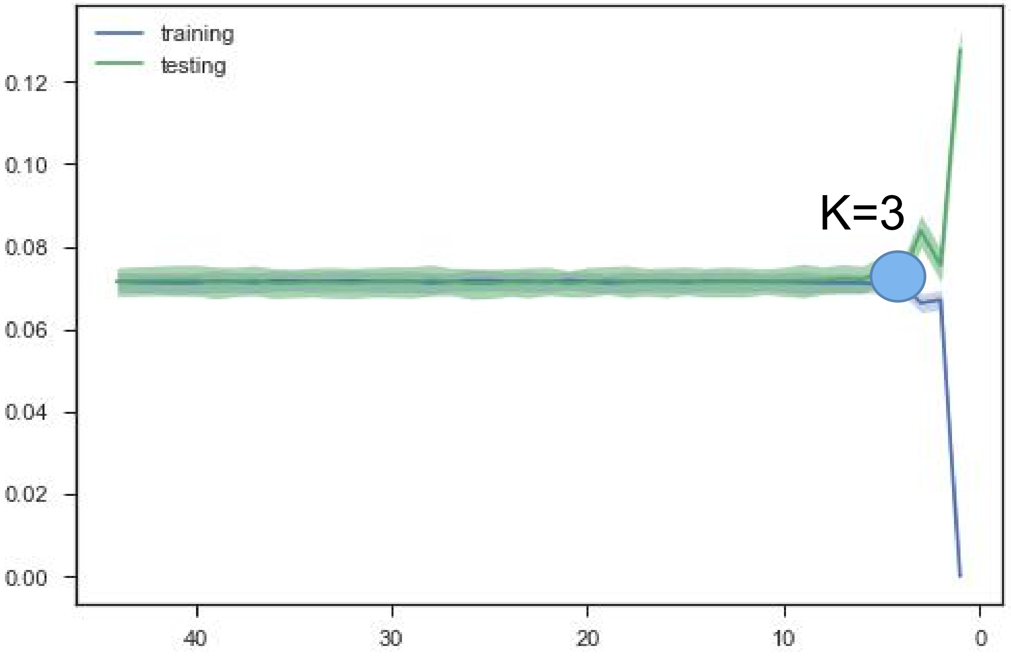




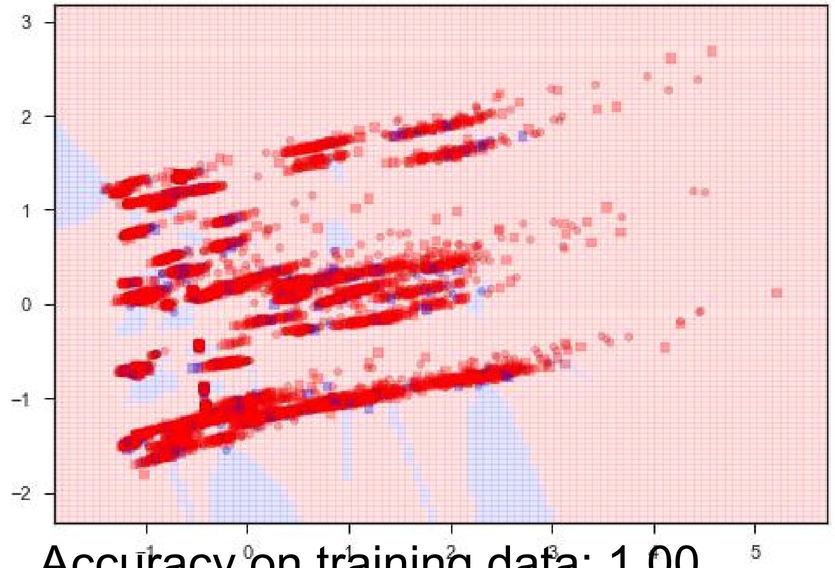


Dimensionality - PCA

- Using KNeighborsClassifier
  - A group of three features can explain much of the variations in the dataset

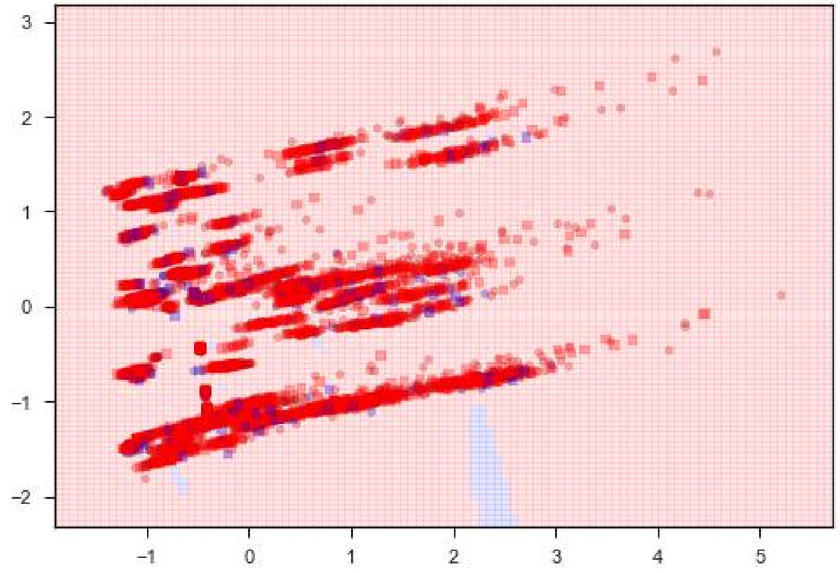


K=3



K=1

Accuracy on training data: 1.00
Accuracy on test data: 0.88



K=3

Accuracy on training data: 0.93
Accuracy on test data: 0.93

Regression

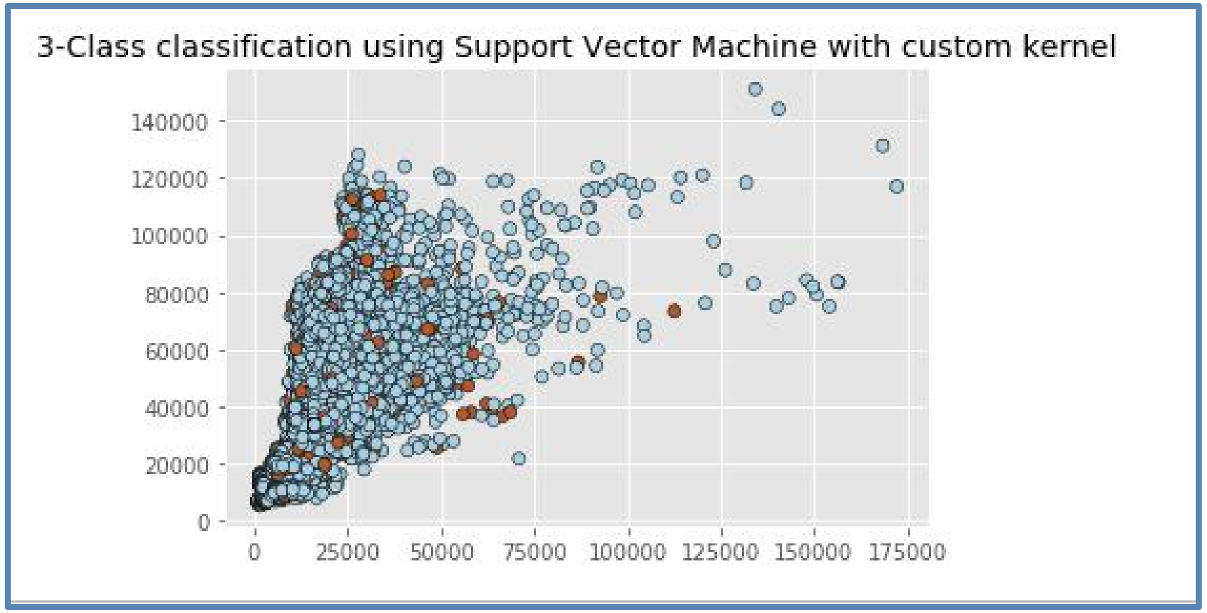
- Studied effect of MarkDown on sales of:
  - Health-items
  - Kids-Items
  - Office-Items
  - Transport (Auto)-Items
  - Fashion (Wears)-Items and
  - Food-Items
- These Target Features generated using the VIF results

Target Feature	Score
Health Care Items	75%
Kids-items	83%
Office Items	55%
Auto Care Items	66%
Fashion Items	78%
Food Items	55%

As Expected, 83% and 78% of the variation in the sales for Kids' and Fashion items are accounted for by the markdowns

Classification

Classification Method	Model Accuracy		
	Training Set	Test Set	
KNeighborsClassifier (K = 6)	0.93	0.93	
svm.LinearSVC		0.92	
DecisionTreeClassifier		0.93	
Gaussian Naive Bayes		0.61	
Neural Network		0.96	

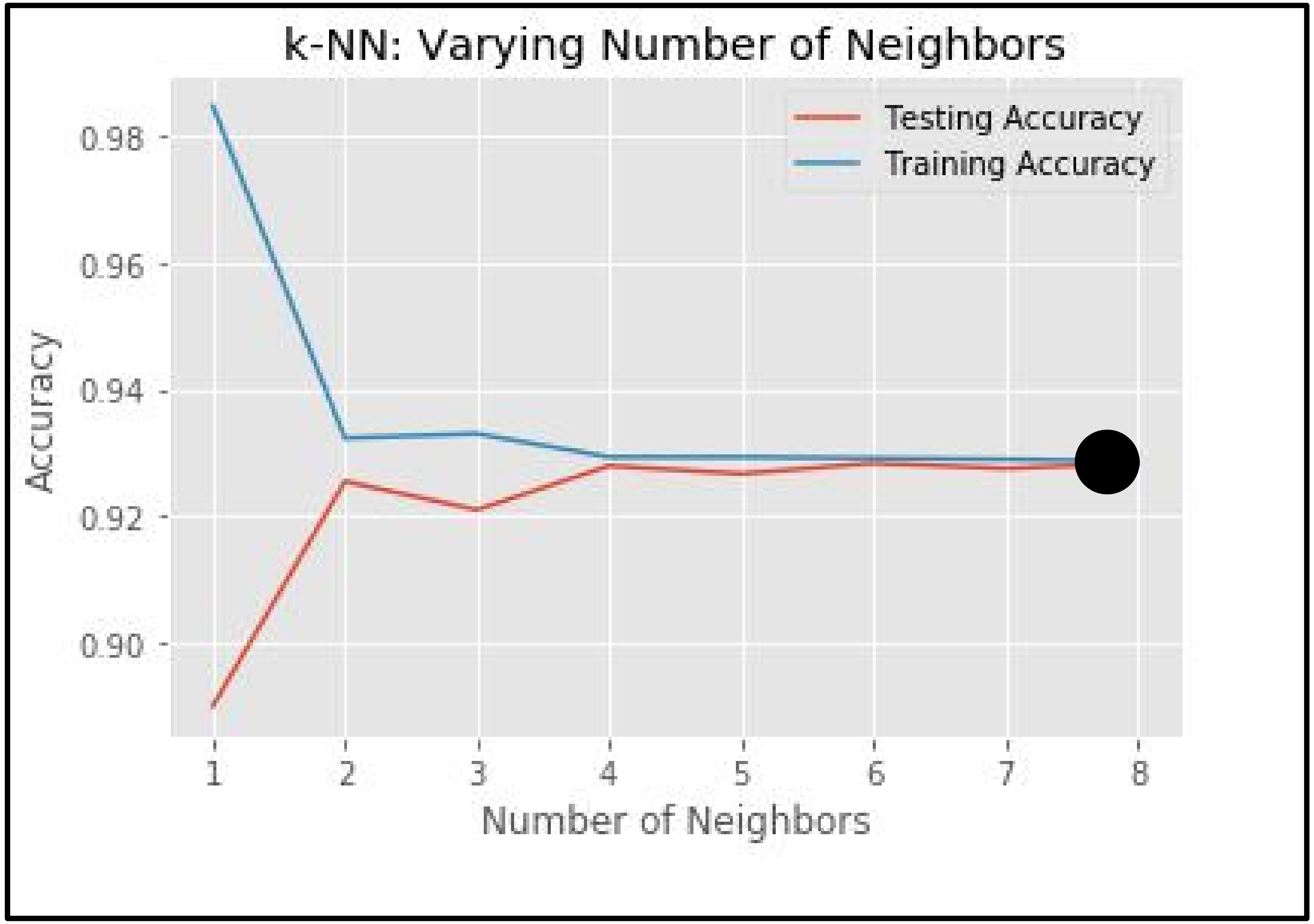


3-Class classification using Support Vector Machine with custom kernel

X - sales from 82 departments
y - Holiday Indicator (IsHoliday)

Overfitting/ Underfitting

K = 6 (Optimal K)


k-NN: Varying Number of Neighbors
Testing Accuracy
Training Accuracy
Accuracy
Number of Neighbors