Topic: Walmart Sales Forecasting

Team Members: Xinya Zhao, Doug Meyers, Matthew Aaron, Jessica Mondon

Business Understanding

For major retailers such as Walmart, predicting shopping levels can be a challenge. Without this information they are unable to know how much inventory they need, which could result in buying products that won't be sold or lost revenue from being unable to appropriately predict demand. If individual Walmart stores are able to better predict whether there will be an increase in sales, they will be able to estimate the inventory levels needed to maximize their revenues.

One factor that is important to look at is the weather. According to a study by the National Oceanic and Atmospheric Administration (NOAA), thirty percent of consumer patterns are influenced by the weather (Forbes). Due to the holiday season there is a spike in retail sales during November and December, however studies have shown that the colder temperatures lead people to shop less. Colder weather and rain may lead consumers to want to stay home, while they may seek out indoor activities, such as shopping, during hotter temperatures (Starr-McCluer). In particular, snow storms tend to result in losses in sales for retailers (Money). Each day that retail stores close due to the weather, they are reported to lose about $10 million. Even online sales suffer during snow storms as the majority of online shopping occurs while people are at work, and the fact that they are home due to the weather combined with power outages has a negative impact on online sales (Money). Bad weather can also deter people from driving to the store, and can cause delays in products bought online to be delivered. In

order to account for the seasonality of sales, it will be useful to predict sales per store per season.

Another factor that has been shown to play a role in retail sales is fuel prices. It's been shown that higher gas prices result in fewer retail sales (BBC). Decreases in gas prices of $0.30 would result in an additional 1% of potential retail spending (Seeking Alpha). As the gas prices go down people have more money in their pockets, which leads to them drive more and to increased retail spending. This also allows people to travel further to make purchases that they otherwise might not make (USA Today).

The Consumer Price Index (CPI) is the measure of inflation in the economy. CPI looks at the prices of commonly consumed products and compares changes in the price over time (Investopedia). Inflation would decrease a person's likelihood to shop. If someone earns the same amount of money, but goods are more expensive then they won't be able to afford to buy as much, therefore sales will be down (Small Business).

Unemployment rates have also been shown to play a role in spending. Higher unemployment rates weaken consumer spending as people have less money to spend on purchases (Bolden-Barrett). Additionally, those who are unemployed tend to exhaust the money they receive for unemployment and then are unable to spend after that point (Ganong).

Since we were able to establish connections between these features and the sales levels of retailers such as Walmart, we will use these to create a model to predict whether Walmart's weekly sales are over or under the average seasonal sales. Being

able to estimate if they will be at average sales will help them to better understand how these factors play a role in the sales and inventory levels they will need when those conditions are met. Our target variable is whether the sales for a particular department of a particular Walmart store is above or below that store's average for the season.

Data Understanding

Data Understanding is critical in solving business problems because the historical data that is often used is rarely collected with the intention of helping to solve a specific problem and was most likely collected without a specific purpose. It is typically one of the most time-consuming aspects of analysis and requires detecting and addressing missing data and outliers, determining if any data features need to be excluded, and also addressing missing data elements (Data Science for Business). The data we used for our models was provided by Walmart, however, it would be erroneous on our part to think that all of the data is accurate and reliable. Our features csv file contains additional information about each store, the department, and regional activity for the mentioned dates with details like the store number, the average temperature in the region, the cost of fuel in that region, the unemployment rate, the consumer pricing index, whether the given date/week is a special holiday week or not, and data related to promotional markdowns that Walmart was running.

Kaggle, the website that provided us with the data, gives us no information regarding the sources of the other feature variables such as average temperature, fuel prices, etc. We can't verify that their sources are accurate or that the values could differ

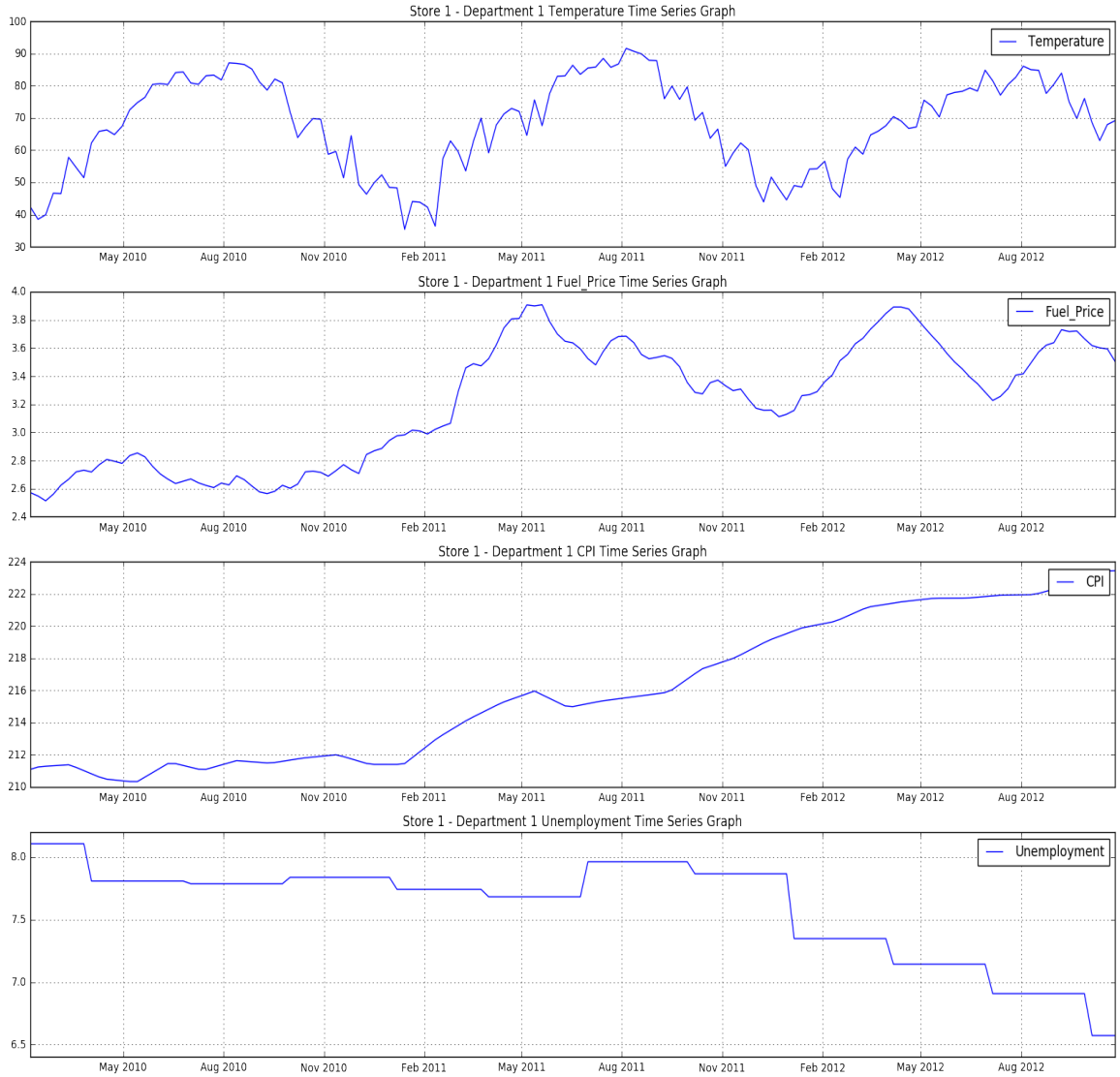
if the variables were more explicitly defined. Our team researched what these feature variables represent in the real world and how much they can differ throughout the year and by location.

As shown with our data, the fuel costs from 2010 to 2012 dramatically increased but it's hard to know the true effect that would have on the average Walmart consumer. We need to obtain additional data that would provide us the percentage of consumers for a given Walmart store that drive to and from the store, how many miles the trip is, what the average gas price is in their area, and even the average MPG for cars in that area. We could also normalize it by saying that gas prices at that given time were up or down X% from the previous week. The dollar values that are currently listed gives us little context as to whether those prices are expensive or cheap relative to factors such as the metropolitan area or the historical prices in that area.

We also researched the variations in consumer price index values from the mean over time to help put it in perspective. The average CPI index in 2010 was up 1.6% from the previous year and in 2011 it was up 3.2% from the previous year. Lastly, in 2012, it increased 2.1% from the previous year (CPI). These percentages allow us to better understand the effects of inflation and their subsequent effect on the cost of commonly bought goods by the average US consumer.

Quantifying the unemployment rate is also equally important. The unemployment rate can vary drastically from state to state and additional data should be gathered to more accurately represent the unemployment rate in a given geographical location. To provide even further clarity, the number of unemployed people should be given in actual

number of persons rather than a rate. Through additional research, we found out that the number of unemployed people in the US in 2010 decreased from approximately fifteen million people to twelve million people in 2012. The implications on inventory levels for a certain store given that there are now three million additional wage earners in the US will have immense impacts on the number of items sold in their stores. The graph shown below illustrates how our four main feature variables vary over the time span that our data set contains.



Part of the data understanding phase is estimating the costs and benefits of each

data source and deciding whether additional investment is merited. The additional investment could be cleaning noisy and variable data or verifying that the data is accurate. For example, there are many different sources that provide accurate weather data and the minor discrepancies between the sources wouldn't have a material effect on our target variable. It wouldn't be worth additional investment into obtaining other sources of data. However, small fluctuations in gas price data could dramatically affect our target variable.

Data understanding requires digging beneath the surface to examine the data. We need to dive deeper and figure out whether this data represents the average consumer in the area and also figure out a way to better represent this (Data Science for Business). One idea our group came up with should we want a higher level model was to normalize gas prices so readers, such as high level executives at Walmart, could more easily intercept our results and also understand the data in real terms. Our first step would be to more precisely label our feature variable and also to provide the sources of the data. It is common for the costs of data to vary and for additional undertakings to be conducted in order to obtain additional data. Had our team had access to additional sources of data and additional time, our data understanding could be vastly improved through the aggregation of additional sources of data.

Part of data understanding is not only deciding whether to fund the costs of obtaining additional data but also cleaning the data and removing unnecessary feature variables and aggregating the data from different sources. As discussed in our next section, we dropped certain features that we believed would promote leakage and also

created our own feature variable called ‘Season’ to account for the temperature fluctuations throughout the year.

Data Preparation

The data for this project was sourced from a Kaggle competition, where Walmart asked participants to solve a similar problem, except using linear regression, instead of a supervised classification model, which we plan on using to solve this problem based on the scope of this class. Walmart provided 2 datasets. The first was a Weekly Sales csv, where each row pertained to the weekly sales figure by store and department.

Store	Dept	Date	Weekly_Sales
1	1	2010-02-05	$24,924.5
1	1	2010-02-12	$46,039.49

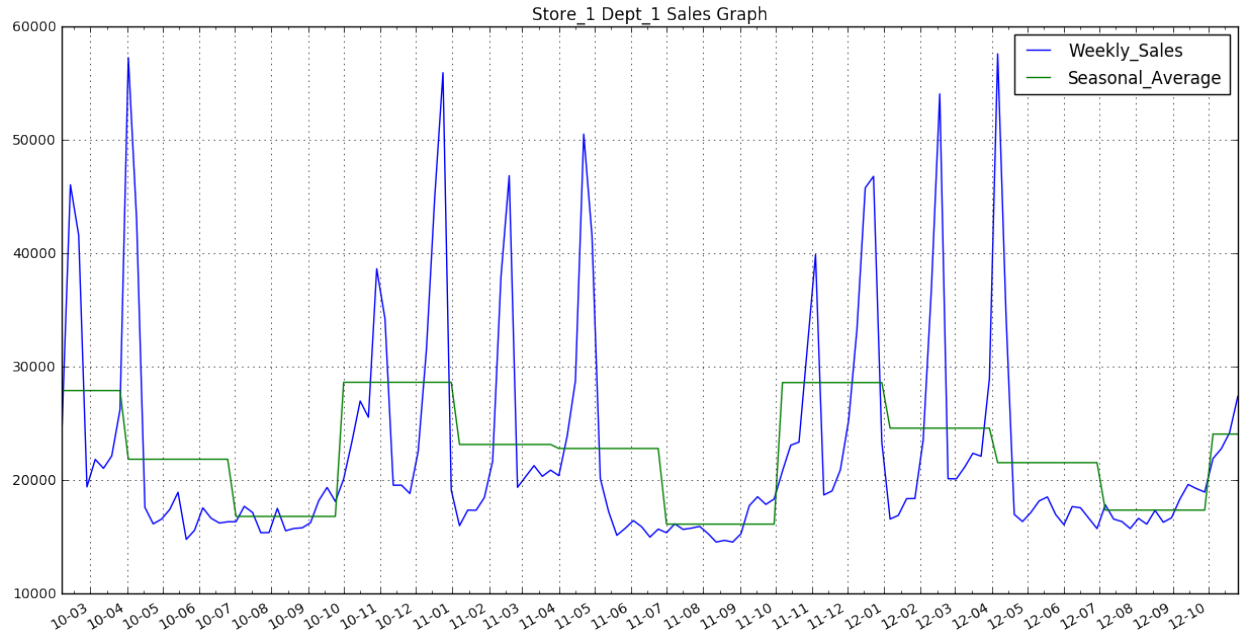
From the date column, we derived a ‘Season’ feature with values ordered from 1 to 4, and then calculated the average sales by Store, Dept and Season. We incorporated the Season attribute in the average sales figure, in an attempt to account for some level of seasonality for each store and relative department. For each row or vector, we compared the Weekly Sales figure of each store and department and labeled the data as 1 (above weekly sales) or 0 (below weekly sales).

The second data set (Features csv), described the exogenous economic and weather related factors for each store by date, including the data mentioned above

(Temperature, Fuel Price, CPI, Unemployment). The Features.csv was joined to the weekly sales figures to produce our initial merged complete set of training and test data. At this point in time we dropped features that would result in leakage, such as sales related data, and discarded other attributes which we perceive to be extraneous, at least for our model experiment.

Season	Store	Dept	Date	Temperature	Fuel Price	CPI	Unemployment	IsHoliday	Target Variable
1	1	1	2010-02-05	42.31	2.572	211.0963582	8.106	FALSE	0

Since our data was clean and there was no missing data, our next tasks for data preprocessing are to create dummies for categorical features and add in relevant time-series features. We first created dummies for two features that can identify the ‘Stores’ and ‘Department’. There may not be uniform patterns that would generalize across all stores or departments, and variance in sales for each individual stores / departments is expected. Second, we plotted the weekly sales in a time series to assess if there are any additional trends or seasonality observations. The chart below shows the actual weekly sales and average seasonal sales of store_1 department_1 over 3 years, from 2010 to 2012.



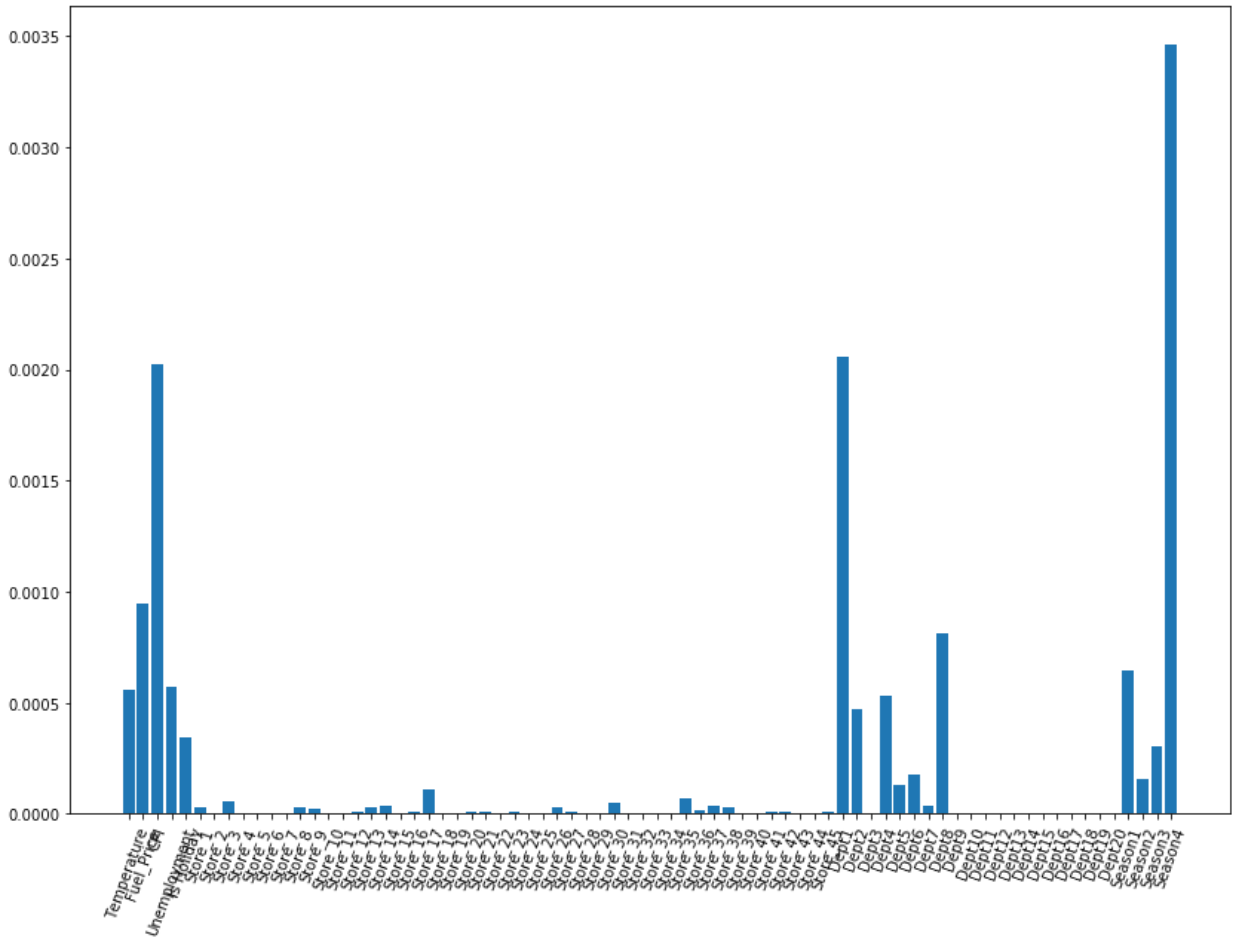
When we reviewed the time series data, there were 3 main factors we used to assess possible data patterns: seasonality, trend and random variation. Because random variation is an inherent hidden pattern existing in all data sets, and can hardly be depicted by simple observation, we mainly included seasonality trends when preprocessing the data and including time-series features.

The chart clearly depicts that sales for this department have some seasonality over periods. Sales will reach its regular peaks in same 4 months each year, which are in October, December, February and April. Also, this trend can be observed from the seasonal average sales. These seasonality factors are reflected in our model by creating dummies for different levels of the date attribute. The time unit that we used was month and year. Preprocessing these new dummy variables optimized the tree model by 9%.

The most recent historical data is generally more informative when predicting

market sales, thus we need to account for trends in the model. There are two kinds of trends that we considered. One is the trend where the sales continuously climb or fall for each period. We incorporated this trend by introducing a feature that gives the last-3-month (last 12 weeks) average sales of an observation. The other type of trend is the opposite, where sales ebb and flow over one period. This is observable in the chart when peaks and troughs occur alternately. For this trend, we added a feature representing sales moving average over a relatively shorter period - one month (four weeks). This feature is more granular than the 3 month moving average, and will make it easier to judge if weekly sales is likely to be higher (or lower) than the average seasonal one.

Are the features and attributes relevant to the target value? We ran the information gain metric, which validated that temperature, CPI, fuel price and unemployment (first 4 bars) all have relevance to the target attribute, as demonstrated by the bar chart below. The winter season, as expected, and the first department had relevant high information gain values as well.



The final clean dataset can be viewed from this link.

Modeling

Based on our data understanding, and evidenced by the information gain metrics on the selected features, our team selected to use a supervised classification model, as labeled data is available. A regression or value estimation model, would unequivocally be more useful to Walmart. However, proving that these selected features affect sales, and the probability of class occurrence would be simple, and help Walmart determine if they should invest in a more expensive complicated regression based model.

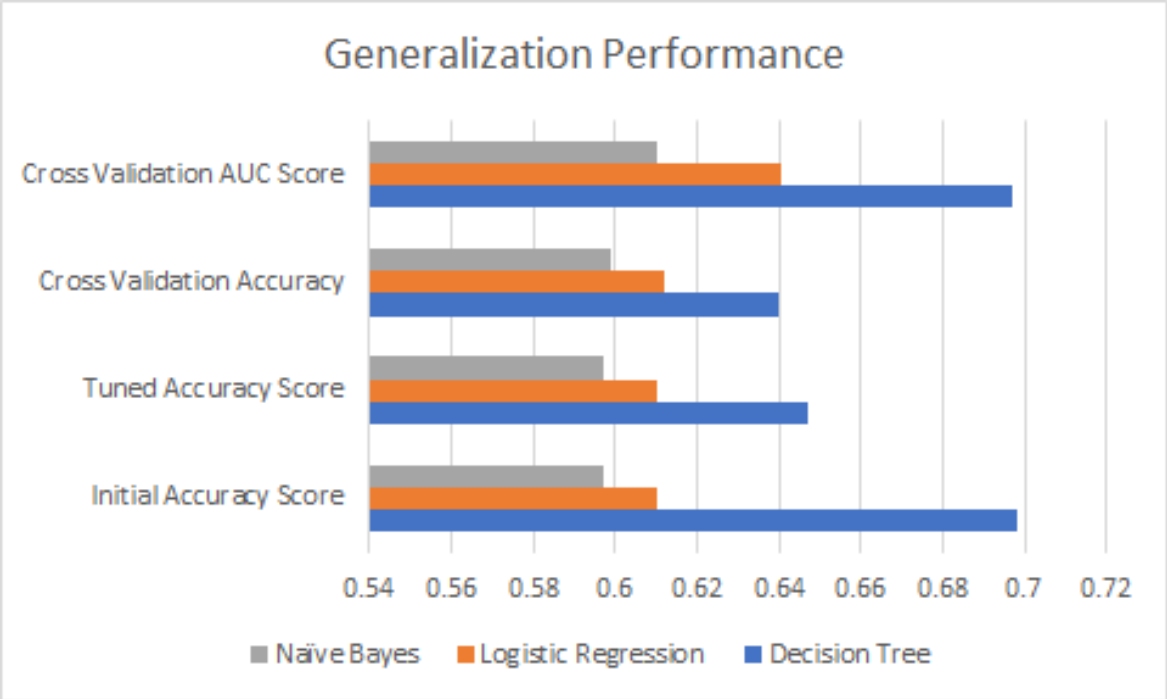
The 3 most popular data mining classification models (Decision Trees, Logistic

Regression and Naive Bayes) were examined and executed in our exercise. We did not select SVM, since we thought that Walmart would want to know the probabilities, which would be required to generate intuitive business friendly accepted metrics like Lift. If the Naive Bayes algorithm resulted in the best generalization results, we would need to be cautious before using any probability based metrics, since they would be skewed due to the independent feature assumption. Subsequently, the team ruled out that k-Nearest Neighbor (k-NN) wouldn't be intuitive to stakeholders, since our objective was not to find records that are close to each other.

Model	**Pro(s)**	**Con(s)**
Decision Tree	• Easily Interpretable to Business Stakeholders • Fast • Works well with a small feature set • Non-Linear Model (Low Bias)	• Complex • Prone to Overfitting • May not fit curved surfaces
Logistic Regression	• Difficult to Overfit • Fast in Use Phase • Would work well with many features	• Non-Intuitive to Business Stakeholders • Linear Model (High Bias) • Slower than Decision Tree to Model
Naive Bayes	• Would work well with many features that are high dimensional. • Very fast and simple	• Probabilities are biased/skewed • Treats features as independent which we know many not be true

Based on the criteria above, and the desire to use a data mining algorithm that was inherently business friendly and easy to visualize/understand, we started out with a

Decision Tree model. Our biggest concern was overfitting and that it would need to be managed using complexity control practices and hold-out/evaluation techniques. We also tested with Logistic Regression and Naive Bayes to understand and evaluate if those models would improve model generalization performance. All models were trained with 75% of the dataset and 25% hold out test data. Move over, the sklearn GridSearch function was used to optimize hyper-parameters to maximize accuracy or ROC scores and control for complexity with cross validation (folds = 5-6). The results of the model evaluation are below:



- <u>Decision Tree</u>
    - Baseline Model Accuracy Score: 0.698
    - Tuned Accuracy Score: 0.647

        - Grid Search Tuned Parameters: Max Depth = 10, Min Sample Leaf = 5, Min Samples Split = 64

    - Cross Validation Accuracy (with 5 folds): 0.640

    - Cross Validation AUC Score (with 5 folds): 0.697

- Logistic Regression

    - Baseline Model Accuracy Score: 0.6100

    - Tuned Accuracy Score: 0.6104

        - Grid Search Tuned Parameters: Penalty = L1, C=0.1

        - Grid Search with StandardScaler in Pipeline: Penalty = L2, C = 1.0

        - Grid Search with Polynomial Features in Pipeline: Penalty = L2, C = 1.0, Degree = 1

        - Grid Search with Feature Selection in Pipeline: Penalty = L2, C = 1.0

    - Cross Validation Accuracy on the whole data (with 5 folds): 0.612

    - Cross Validation AUC Score on the whole data (with 5 folds): 0.6405

- Naive Bayes
    - Baseline Model Accuracy Score: .597
    - Tuned Accuracy Score: .597
        - Grid Search Tuned Parameters: Alpha = 11
    - Cross Validation Accuracy (with 5 folds): 0.599
    - Cross Validation AUC Score (with 5 folds): 0.61

When comparing all three models, the Decision Tree data mining algorithm consistently resulted in higher accuracy and ROC generalization performance. The Decision Tree algorithm has a ~64 % accuracy and ~70 % ROC score with cross validation, which is a material improvement from random or 50%. If Walmart can use this model to successfully predict sales with ~65-70 % accuracy to help manage inventory, it will definitely affect their bottom line. Considering that Walmart holds ~ $ 43 billion of inventory at any given point in time, if this model helps the firm effectively manage inventory levels by even 1%, they can save ~ $430 million.

Evaluation

To evaluate the generalization performance of a model, understanding which performance metrics fit the business situation well is critical. The business problem we want to solve is to determine the weekly sales level of a Walmart department – whether it is higher than the seasonal average. If it is higher, Walmart will enlarge its order for inventory to meet the increasing demand; if lower, Walmart will maintain the inventory at

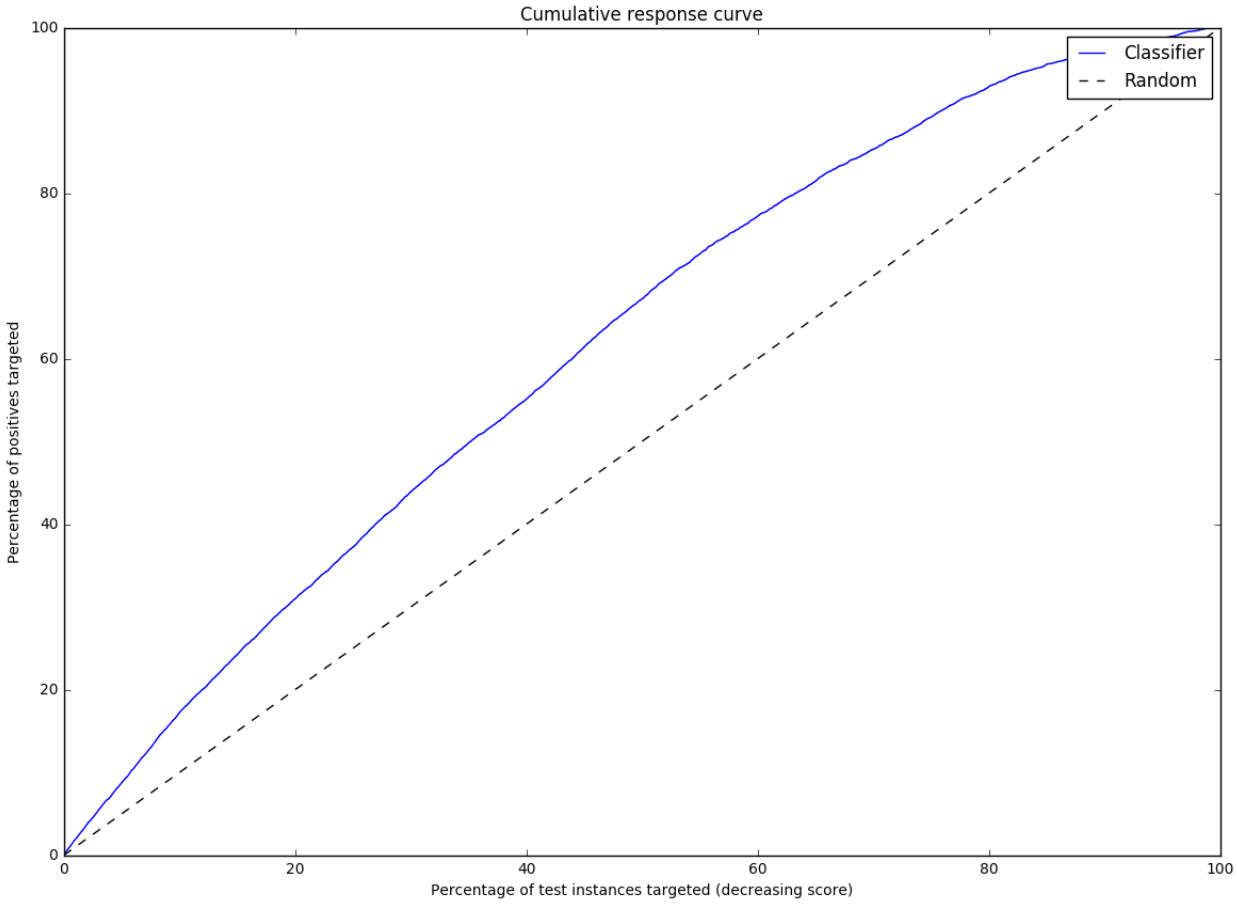
a relatively lower level for that period. Thus, the effect of a false prediction of our model will either be a surplus of inventory or an unsatisfied demand. The cost matrix is showed below. In retail terms, a true positive would imply that the company expected to be at average sales and they planned their inventory that way so it would result in a net gain of 0. A false positive would be when the model predicts the sales to be average, but the demand is higher than expected. This will result in unmet demand and they will not realize all of the potential sales that they could have if they had stocked more inventory. The benefit for a true negative would be the profit that they made, or revenue from increased inventory minus the cost of the extra inventory. The false negative in the confusion matrix would correspond to the surplus of inventory as they predicted above average demand, but did not get it. This would result in a loss to the company as they had to buy extra inventory, and the loss would be in the amount of the cost of goods sold of the extra inventory that they purchased. Different from traditional churn cases, which always tend to value true positive over true negative to gain the highest expected value, our model determines the level of penalty for the two types of wrong prediction by comparing their costs. If there is a preference to lower one of the costs more, then it will be more proper to use the metrics like ROC, AUC, and Lift, which are proficient in ranking entities according to their likelihood to be positive (if the surplus inventory, false positive, is costlier, this judgement will still be true if converting the target variable from higher than seasonal average to lower than seasonal average and retraining the model). But if these two false predictions are equally costly, accuracy will be an ideal metric for the performance evaluation. In our practice, we use the two metrics, ROC and accuracy, to do the evaluation. And considering the general low cost of holding extra

inventory for such large retailers, we suppose Walmart will ‘hate’ unsatisfied demand (false negative) more and so that target variable conversion is not needed.

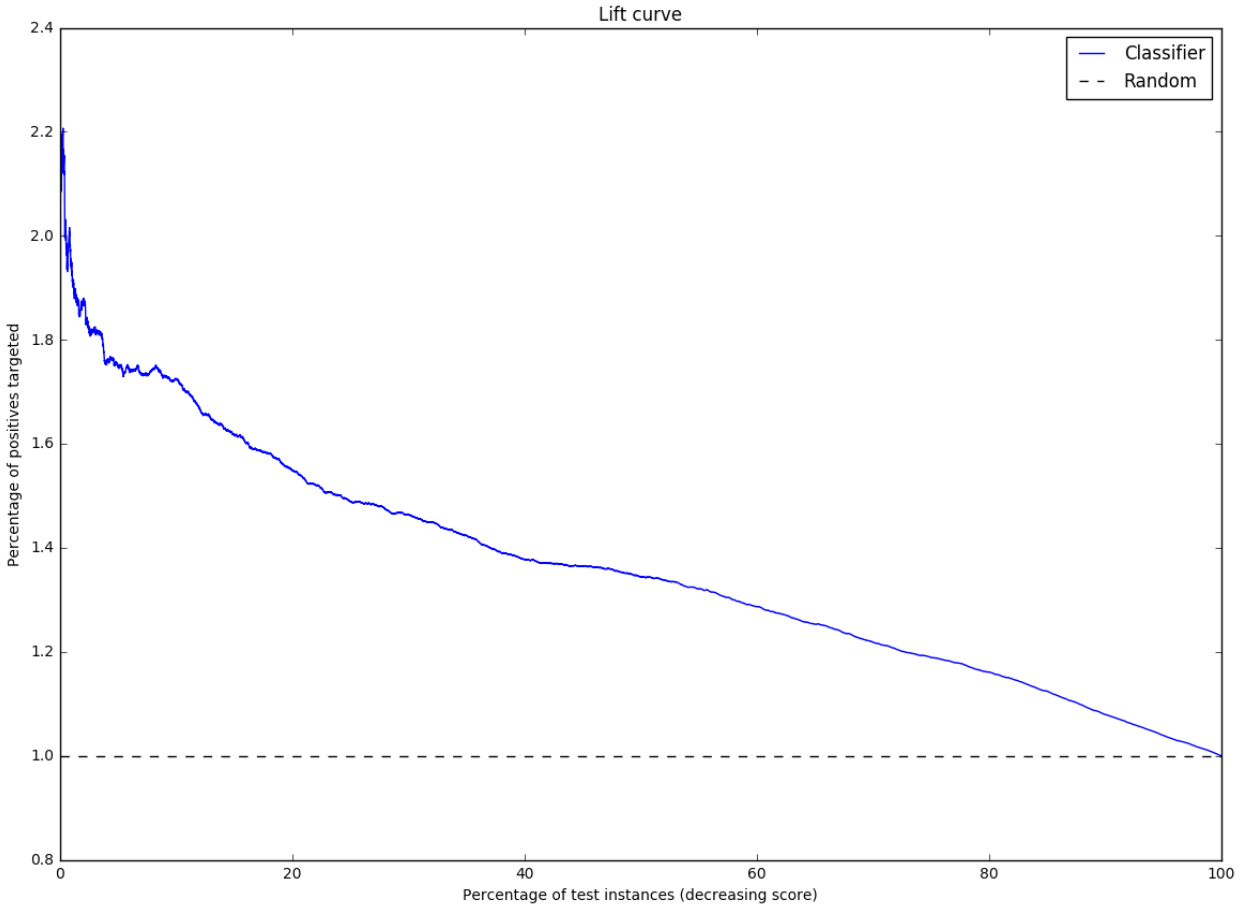
Actual Result	Confusion Matrix		
		Positive	Negative
	Positive	TP-3300	FN - 2201
	Negative	FP-3030	TN - 6153

	Cost / Benefit Matrix		
		Positive	Negative
	Positive	0	-COGS
	Negative	Lost sales	Revenue

As we do not have the cost / benefit matrix data to build a profit curve, this model clearly has benefits which can be built into a business case when assessing the cumulative response curve. The cumulative response curve clearly demonstrates that the decision tree model performs better than random, and would almost always create some benefit to Walmart.



Furthermore, if we only wanted to deploy the model where there are substantial benefits and more certainty, we could use the model when it has the highest probability, targeting the 20% of the instances, which has a lift of 1.5- 2.2 times the random base rate. The model never performs worse than random, so there would be no penalty for always using it.



Notwithstanding that we can't directly produce a ROI figure or profit curve, we believe that the combination of the proposed earned value framework and the model's demonstrated positive lift, would help build a strong business case to optimize Walmart's inventory management system, which has the potential to save them firm millions of dollars per year.

Deployment

The deployment stage is when the results of data mining are put into real use in order to realize some return on investment. A given model can be deployed immediately after data mining; however, it is not advisable to do since the advantages of testing a model in a laboratory setting will allow for a more efficient testing environment. A model

that successfully passes rigorous testing should not be taken at face value (Data Science for Business). Our model accounts for location (Store) and product (Dept), and Walmart can run the model weekly to forecast if sales will be above or below the average, contrast the model predictions against current inventory for each Store/Dept, and automatically choose to either replenish that department's inventory to a minimum 'base' level, or add inventory as sales increase. External considerations should be factored that could render the model less useful in a real-world setting. Through the deployment of the system rather than the models, Walmart can more easily adapt to a rapidly changing world and also not worry about creating a different model specifically catered to each line of their business. Our model should be taken with a grain of salt and the firm should be aware that every model is wrong, but some are useful. For example, our model can't take into effect external events such as weather related phenomena like an impending snowstorm or hurricane. Nor can it anticipate the level of demand for a hot new item that will sell out immediately.

We should advise Walmart of the old adage that, "our model is not what the data scientists design; it's what the engineers build." This model will be useful to help determine if the variables used in our model play a role in whether or not inventory is at or above average. Doing so will allow them to appropriately plan inventory levels to maximize the store's profit under normal conditions. Poor management, strikes, or theft are not taken into account in our model and yet, they can severely impact inventory levels unexpectedly. Our team would suggest that Walmart implements our model alongside their engineering team so that they can provide the critical insight needed to make necessary adjustments down the road.

Works Cited

1. Provost, Foster and Fawcett, Tom. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. Sebastopol, CA. O'Reilly Media. 2017
2. http://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/
3. Barnes, Ryan. "Economic Indicators: Consumer Price Index (CPI)." *Investopedia*, www.investopedia.com/university/releases/cpi.asp.
4. Bolden-Barrett, Valerie. "How Does Unemployment Affect Businesses?" *AZ Central*, yourbusiness.azcentral.com/unemployment-affect-businesses-15445.html.
5. Davidson, Paul. "Falling gas prices fuel consumer spending." *USA Today*, www.usatoday.com/story/money/business/2014/11/09/gas-prices-boost-consumer-spending/18663989/.
6. Ganong, Peter, and Pascal Noel. "Consumer Spending During Unemployment: Positive and Normative Implications." Harvard University Department of Economics, scholar.harvard.edu/files/ganong/files/ganong_noel_ui_2017.pdf. Manuscript.
7. "Higher fuel prices take toll on retail sales." *BBC*, www.bbc.com/news/business-39365048.
8. "The Impact Of Gas Prices On Retail And Restaurant Sales." *Seeking Alpha*, seekingalpha.com/article/2714935-the-impact-of-gas-prices-on-retail-and-restaurant-sales?page=2.

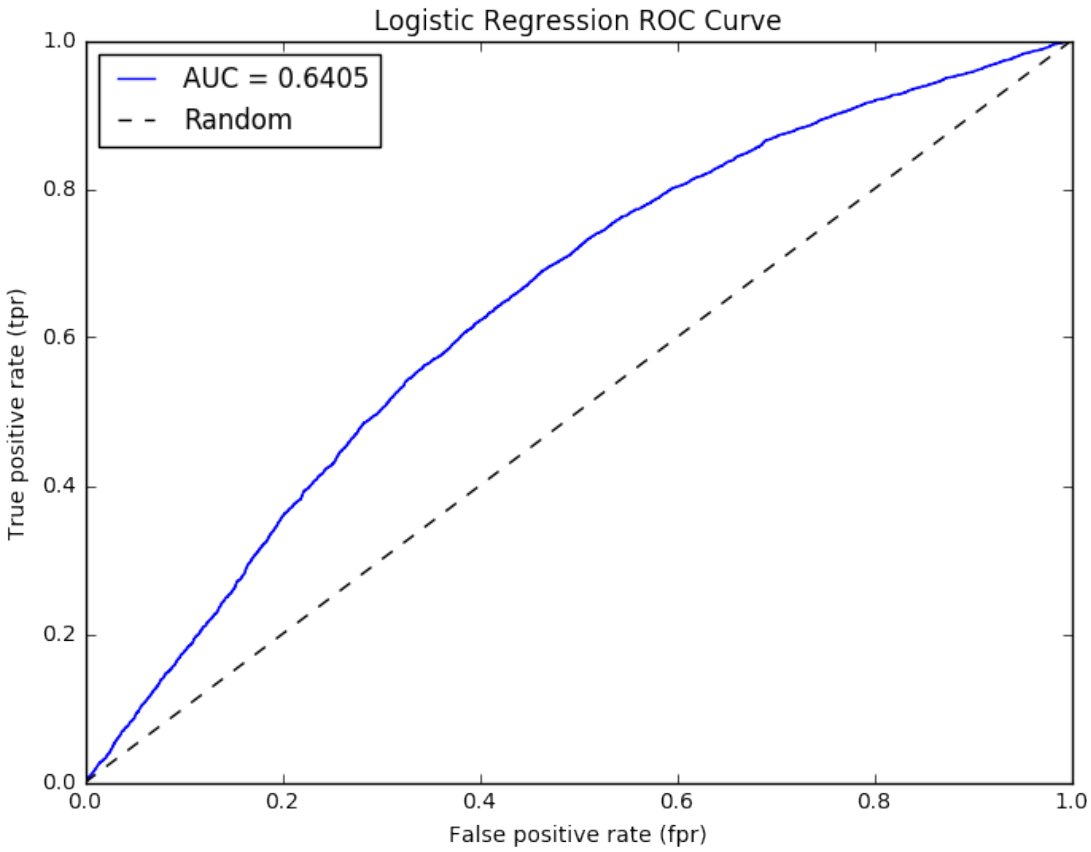
9. McBride, Carter. "How Does Inflation Effect The Purchasing Power of Money?" *Chron*, smallbusiness.chron.com/inflation-effect-purchasing-power-money-696.html.
10. Starr-McCluer, Martha. "The Effects of Weather on Retail Sales." *Federal Reserve Board of Governors*, www.federalreserve.gov/pubs/feds/2000/200008/200008pap.pdf.
11. Tuttle, Brad. "The Curious Ways Brutal Snowstorms Affect How We Shop." *Money*, time.com/money/3703336/snow-blizzard-impact-shopping-retail/.
12. "Walmart 2017 Annual Report." *Walmart*, s2.q4cdn.com/056532643/files/doc_financials/2017/Annual/WMT_2017_AR-(1).pdf.
13. Walsh, Paul. "This Holiday Season, Colder Weather Will Boost Retail Sales And Shoppers' Spirits." *Forbes*, www.forbes.com/sites/ibm/2016/12/19/this-holiday-season-colder-weather-will-boost-retail-sales-and-shoppers-spirits/#3ee947214c1c.

Appendix

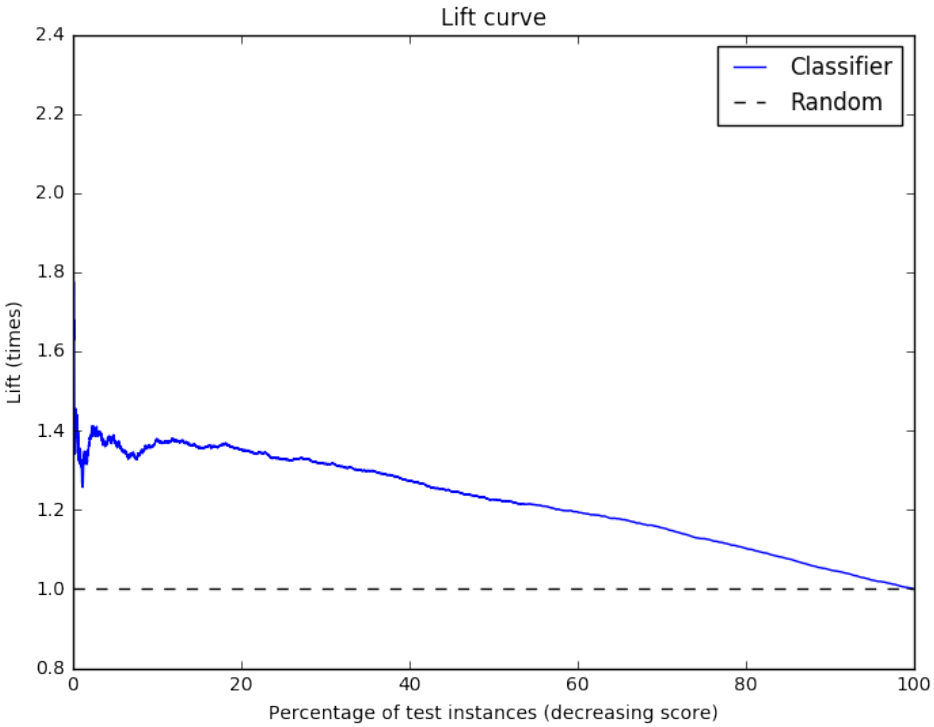
Contributions

- Business Understanding - Jessica Mondon
- Data Understanding - Matthew Aaron
- Data Preparation - Doug Meyers, Xinya Zhao
- Modeling - Doug Meyers, Xinya Zhao
- Evaluation - All
- Deployment - Jessica Mondon, Matthew Aaron

Logistic Regression ROC Curve



Logistic Regression Lift Curve



Logistic Regression Learning Curve