Predicting Sales of Retail Items using Bigmart Sales Data

Adriana Rotaru and Dionisie Nipomici

Stat 139

1 Abstract

Sales forecasting is a key element in conducting a business. Sales forecasting can help manage all aspects of a business which determined the emergence of many marketing enterprises which provide sales forecasting as a service. This project tries to find an efficient way of predicting sales using the most appropriate parameters.

This paper aims to look at the parameters which determine the amount of sales in retail stores and use the findings to come up with improvement and solutions for a better supply chain management and inventory controls.

Using BigMart Sales Data, the main factors that we have studied in sales prediction are related to the size, the type and the location of the store, as well as the some internal factors related to the store including establishment year and Item visibility in the store.

Relying on the linear regression results, we have found that the best predictive model for Sales includes price of the item, the store features and the interaction between the two. Our best model, which we called **Modelstep** (a stepwise procedure in both directions) has an R-squared of **69.26 percent** and the minimun AIC of **35885**. The results of analysis indicate that the most influential factors in sales forecasting are the store features and price.

2 Introduction

Sales is the lifeblood of a business. It is what helps cover expenses, grow the economy, pay employees, market new products and buy new inventory. With the huge success of big retailers such as Walmart, Kroger, Walgreen's, smaller grocery stores have given up many customers and part of that reason is a poor chain and inventory management.

On the other hand, some small grocery stores such as Trader Joe's or Wholefoods have sustained a huge perpetual success over the years, even though they

do not offer a huge variety of products. Many of these stores generate huge amounts of waste daily because items are not displayed in an efficient way for the customer to access them and in a way that encourages an effective marketing of the products. Another issue is that many of the stores are located in non-strategic areas where the flux of customers is weak for those specific type of products.

Much effort goes into understanding which factors that help increase sales and which ones are less influential in determining the amount of sales. In this study, we aim to investigate additional factors which significantly influence the amount of sales in retail stores and thus help small and large businesses create better strategies for optimizing inventory, save money and practice effective marketing. The aim is to build a predictive model and find out the sales of each product at a particular store.

Using the BigMart Sales Data, we specifically wish to examine 3 types of variables that can be easily quantified and put in practice by businesses:

1. `Outlet_Identifier`- the outlet id
2. `Item_MRP` - maximum retail price
3. `Item_Visibility` - an index in percent of how visible and easily accessible is the product in the store

3 Methods

3.0.1 Data. Description.

Big Mart World is franchise retail business with stores all over the world. Because Bigmart has digitalized the tracking process of sales, it has put at public disposal lots of data collected from the stores. The dataset we are using for this study was collected in 2016 and contains over 8000 rows and 12 variables.

We have performed data cleaning in order to make the data set functional and easy to use. Data cleaning included dealing with empty values, reassigning names to some variables, splitting them as well as factorizing them appropriately. We have also narrowed down our list of predictors and used different predictors for different models and purposes. The response variable across all models is `Item_Outlet_Sales` (See **Figure 1** below) for a complete set of the predictor variables and their description).

Predictor Variables

Variables	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

Figure 1: Predictors Description

3.0.2 Data. Pre-processing and feature engineering.

The first thing we did before starting the analysis was to regroup some of the values in Item.Fat.Content which were referring to the same category. For instance, we have regrouped "Low Fat" and "LF" together. We also noticed that the `Item_Type` has too many categories, so we have created a separate column - `Item_Cat2` - which has 3 categories for food, drinks and non-consumable based on the `Item_Identifier` initials from the first column. We have transformed `Outlet_Establishment_Year` from a categorical variable into a numerical one and made a new column `Outlet_Age` , by subtracting the establishment year from the year 2016, when the data was collected. Furthermore, after these modifications, there were some missing values from our dataset, most of them in the column `Item_Weight` . Interestingly enough, based on a boxplot of the Outlet ID's (**Figure 10**, 2nd row), these weight values were only missing for Out027 and Out019, which suggested that those stores do not track the item weights. Since, items with the same ID, should weigh the same across different stores, if they have the same identifier, we assigned an mean weight value to the items with missing weight from the same ID category. 4 Items had unique ID's so we decided to remove them from the data to facilitate the analysis, which we will discuss thoroughly in the Limitations section. The last thing that we did,in our data pre-processing was to create a new column which contains the sales

volume, calculated as the ratio of `Outlet_Sales` and price `Item_ MRP`.

3.0.3 Exploratory Data Analysis (EDA) and Data Transformation.

We visually inspected the distributions of the response and numeric predictor variables (histogram plots, **Figure 6**) to determine whether transformations are needed to correct any skewness. We also examined possible violations of linearity between numeric and difference in means among categorical predictors (boxplots and plots **Figure 10**) and homoskedasticity of residuals for the fitted models. Finally, we explored the relationship between the significance of the association of the response versus categorical variables and numeric predictors using boxplots and scatter plots, respectively.

An important visualization of the quantities in the data that we looked at before conducting the analyses is the summary table for both numerical and categorical data, generated using the **describeBY** function in the psych library in R:

```
 16  Variables      8519  Observations
--------------------------------------------------------------------------------------------------------------
----------------
Item_Identifier
       n  missing distinct
    8519        0     1555

lowest : DRA12 DRA24 DRA59 DRB01 DRB13, highest: NCZ30 NCZ41 NCZ42 NCZ53 NCZ54
--------------------------------------------------------------------------------------------------------------
----------------
Item_Weight
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
.90      .95
    8519        0      415        1    12.88    5.362    5.940    6.675    8.785   12.650   16.850
19.350   20.250

lowest :  4.555  4.590  4.610  4.615  4.635, highest: 21.000 21.100 21.200 21.250 21.350
--------------------------------------------------------------------------------------------------------------
----------------
Item_Fat_Content
       n  missing distinct
    8519        0        3

Value       Low Fat not_food  Regular
Frequency      3917     1599     3003
Proportion    0.460    0.188    0.353
--------------------------------------------------------------------------------------------------------------
----------------
Item_Visibility
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
.90      .95
    8519        0     7876        1  0.06611  0.05579  0.00000  0.01204  0.02698  0.05392  0.09456
0.13949  0.16357

lowest : 0.000000000 0.003574698 0.003589104 0.003597678 0.003599378
highest: 0.309390255 0.311090379 0.321115010 0.325780807 0.328390948
--------------------------------------------------------------------------------------------------------------
----------------
Item_Type
       n  missing distinct
    8519        0       16

Baking Goods (647, 0.076), Breads (251, 0.029), Breakfast (110, 0.013), Canned (649, 0.076), Dairy
(681, 0.080),
Frozen Foods (855, 0.100), Fruits and Vegetables (1232, 0.145), Hard Drinks (214, 0.025), Health and
Hygiene (520,
0.061), Household (910, 0.107), Meat (425, 0.050), Others (169, 0.020), Seafood (64, 0.008), Snack
Foods (1199,
0.141), Soft Drinks (445, 0.052), Starchy Foods (148, 0.017)
--------------------------------------------------------------------------------------------------------------
----------------
Item_MRP
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
.90      .95
    8519        0     5936        1      141    71.55    42.51    52.80    93.84   143.05   185.68
231.21   250.78

lowest :  31.2900  31.4900  31.8900  31.9558  32.0558, highest: 266.1884 266.2884 266.5884 266.6884
```

Figure 2: Predictors Description

3.0.4 Models, Model Assumptions and Comparisons.

The models of our choice were mainly based on the hypotheses we have stated at the beginning of the paper. We used 2 categories of models to investigate the relationship between predictor variables and Sales: 1) multiple and simple linear regression models using Sales cube rooted as a response. 2) a general least squares model for Cross-Validation

The first type of analysis allowed us to test the originally generated hypotheses. We then explored several variations for the baseline model. Our strategy was to compare whether a specific set of additional terms or interaction terms improve the current best predictive model using Extra-Sum-of-Squares F test. We fitted a total of 7 linear regression models (details in Results section), which we visually examined model assumptions by a residual vs. fitted scatter plot and a QQ plot. We also looked at the R-squared improvement with each model we fitted. For the best predictive regression model (as defined by the best R-squared and AIC performance and estimated by cross-validation MSE), we used bootstrap to perform inference on the fitted coefficients permuting residuals, as we did not trusted the normality test for residuals. Finally, we fitted lasso, ridge and elastic net models in attempt to reduce overfitting and improve model predictive ability.

3.1 Hypothesis Generation

Before beginning our analysis, we have thought of some important hypothesis that could be interesting to test. In this study, we have managed to analyze some of them, whilst the rest are here for further reflection and improvement. In order to avoid any sort of data dredging as well as leave some space for unexpected results, we have stated two types of hypotheses the first related to the features of the stores and the 2nd related to the features of the product.

Store related Hypotheses:

1. Location type: Retail stores located in urban areas(represented as Tier 1 in our data) would generally generate more sales on average because of a higher influx of customers.
2. Store Capacity: Stores of Size Large (Outlet.Size) are expected to benefit from higher sales, for the same reason why stores in more populated areas do. We also expect an interaction between store size and store location, given that larger stores populate more urban areas in general.
3. Neighbourhood: Stores located within very close to other big marketplaces might suffer from lower sales, given that there is increased competition. If our data included geographical coordinates, it would allow us to study how location and competitors interact in association with sales, using a machine-learning based approach and measuring a competitiveness factor, for instance.
4. Ambiance: Stores which are well-maintained and managed by polite and humble people are expected to have higher footfall and thus higher sales.

Product Related Hypotheses:

1. Marketing Index: Branded Items are more prone to be attractive to customers. In our data, we used visibility as a marker of marketing.

2. Branding: The customers are more willing to buy items of the same type which have more colorful labels, a better design, a better smell.

3. Utility: Daily use products, such as milk, bread, fruit and vegetables have higher sales and replacement. This is reflected under the Item.Type in our data, and specifically diary, fruits and vegetables

4. Ambiance: New and well designed stores (reflected probably in the Outlet.Establishment.Year variable) are expected to have higher sales.

3.1.1 Exploratory Data Analysis (EDA)

We first checked the distribution of the response variable (Item.Sales) to determine if there is any skewness to be corrected with transformation. As shown in **Figure 6**, the distribution of Sales. Sales is fairly right skewed and such right-skewness is usually corrected by applying the log transformation. However, in our case log-transformation was not the best solution, so we found that cube-root is the right transformation for our response **(Figure 7)**. Note that we decided to sacrifice some interpretability by $(Sales)^{1/3}$ over some other transformations for a more symmetrically distributed response variable. We then inspected the association of numeric and categorical predictor variables with the transformed response in order to choose the best type of OLS (linear or polynomial) **(Figure 10)**.

Specifically, we found that there is linear association between all the numerical variables and our response, Sales. The association with Item Visibility is approximately flat, probably because of the extreme small values of the visibility factor (a one-to-one simple model, would be appropriate to test this association). Additionally, for the categorical variables, the boxplots show enough variation in their distribution, more prominent for Item.Type and Outlet.Identifier.

3.1.2 Multiple Linear Regression Model Performance, Interpretation, Assumptions and Comparisons

Based on our hypotheses related to the store features, we were expecting to see an association between `Outlet_Type` and Sales, given that stores of higher capacity have a higher volume of sales and also higher prices compared to grocery stores, so we fitted a model predicting sales from `Item_MRP`+`Outlet_Type`(**Model 2**) and then another one which has the interaction between these two predictors (**Model 2b**). The associations were significant and positive for all predictors and interaction terms. The second model, which included the interaction terms had a 1 percent higher R-squared (from **68** to **69**) and can be trusted more, since there is significant interaction between prices and outlet types (level 1 stores have more expensive products compared to small grocery stores). The association between price and Sales can be visualized using the plot:

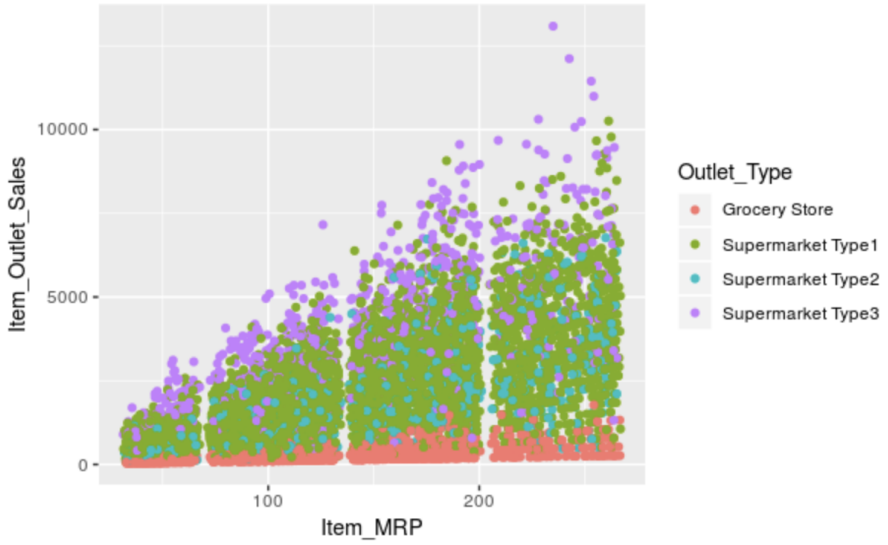

Figure 3: Item MRP vs Sales

Next, we added to the previous model, `Outlet_Identifier` and we excluded `Outlet_Type` because of multicollinearity with Outlet ID (**Model 3b**). `Outlet_Identifier` is one of our most important predictors, as stated in the Hypothesis section. Since `Outlet_Identifier` includes the information about the store's size, store's type and location, it is the most informative predictor out of all. For this model, all predictors were significant (except for Outlet ID 19 and the interaction term between it and price). We concluded, that there is something special about Outlet 19 and that it might be of a different type since there is no significant interaction between its price and type in predicting sales. There was no significant improvement in R-squared for this model.

The last important association that we tested in our analysis was the one with `Item _Visibility`. For advertising purposes, `Item_Visibility` is the most important factor in predicting the sales volume for a product. Many companies pay retailers for displaying their products in specific locations in the store. For instance, snacks are usually displayed in easily accessible spots, usually in proximity to the counter. Therefore, the model we fitted to predict how influential is `Item_Visibility` in forecasting sales included the `Item_Visibility`, `Item_MRP` and `Outlet_Identifier` and the interaction between those, given that stores with higher quality inventory offer higher visibility to their products. The results of this model (**Model.big**) were surprising and suggested that our hypothesis might not be true. The model suggests that `Item_Visibility` and its interaction with `Outlet_Identifier` is not significant, meaning that there is no strong association between the visibility of a product in a store and the sales amount, while there is a very significant association (pvalues less than $2 * 10^{-16}$) between sales and `Outlet_Identifier`. This is a very interesting result, which

suggests that customers might be more determined what to buy and that the inventory and the store ambiance is more important.

Our last step in the analysis was to conduct a **stepwise** procedure to perform model selection. We started with a model with all the main effects and specified the intercept-only model(Model0) as a lower-limit model, and the full model including all two-way interactions of all possible predictor variables as the upper limit. The stepwise procedure would start with the main-effects model and either add or subtract predictors based on their significance and performance. We called this model, **Modelstep**. The Adjusted R-squared for this model was **69.26 percent** and the significant associations based on it are `Item_MRP`, `Outlet_Identifier` and their interaction.We called this model- our best predictive model and we will discuss what are some reasons why we think it is best model in the following section. The summary of the beta coefficients for this model is as follows:

3.1.3 Best Predictive Model

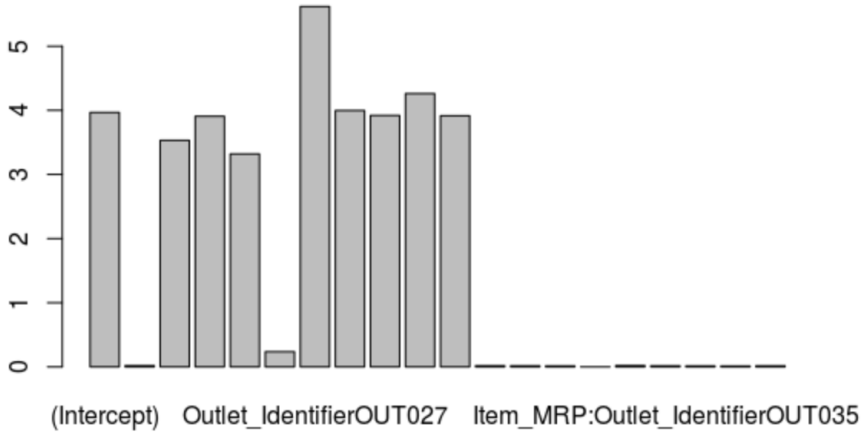

Figure 4: Coefficients of the Best Predictive Model

AIC summary of all models

Model1	Model2	Model2b	Model3	Model3b	Modelfull	**Modelstep**
42822	36145	35906	36130	35885	36135	**35885**

3.1.4 Interpretation of the Best Stepwise procedure-Selected Predictive Linear Regression Model

To test whether **Modelstep** is our best model, we used the method of a Permutation F-test for testing the model overall. The procedure had the ultimate goal of determining whether our coefficients were different from zero, which is another way of testing if there is indeed any association between `Item_MRP`, `Outlet_Identifier` and their interaction. The 95 percent quantile of the F-statistic distribution has a value of 1.65 (see **Figure 18**), while our F-test from the **Modelstep** was 1011. This means that the result is extremely significant and we can trust our association.

In fact, based on our original hypotheses, `Item_MRP` and `Outlet_Identifier` were the main predictors that the analysis relied on. Furthermore, **Modelstep** had the highest R-squared of **69.26 percent** and the lowest AIC of **35885** (a full table of AIC and R-squared can be found in the Appendix).

The issue with **Modelstep**, which might make it less strong is that the residuals are not perfectly homoskedastic. To further test for this assumption, a bootstrap for the slope of the `Item_MRP` was performed, using a residuals approach wich resulted in an 95percent estimated confidence interval which did not contain zero. Therefore, we concluded that the beta can be trusted.

Finally, we have performed a Cross-Validation using the Elastic Net method on **Modelstep** to count for overfitting and see if **Modelstep** can be improved.

3.1.5 Variable Selection and Testing Using Elastic Net Models

To further improve **Modelstep** we decided to regularize it using Ridge and Lasso regressions with various parameters of lambda. We used cross validation to tune for the lambda hyperparameter and also to estimate the mean SSE out of sample. We concluded that as lambda increases beyond 2, Ridge and Lasso both have similar mean SSE (7989.5 and 7988.5), slightly above mean SSE (7988.1) of model step (see **Figure 19**). In order to see if we could further improve the **Modelstep** we decided to use an elastic net model (a combination between Lasso **L1** norm and Ridge **L2** norm penalty). Since, the elastic net regularization is a mix of Lasso and Ridge weighted differently, the best alpha was obtained using multiple trials and testing for SSE. The elastic net model had the following formulation:

$$\beta elastic = argmin \frac{1}{2n} * || \, Y - X\beta \, ||_2^2 + \alpha z_1 + (1\alpha)/2 z_2,$$

where best $\alpha = 0.8$ (see **Figure 19**)
The reason for using the mixed model was because although we found that our model slightly improved it still didn't manage to improve the **Modelstep**. Finally, we have concluded that the best model is **Modelstep**.

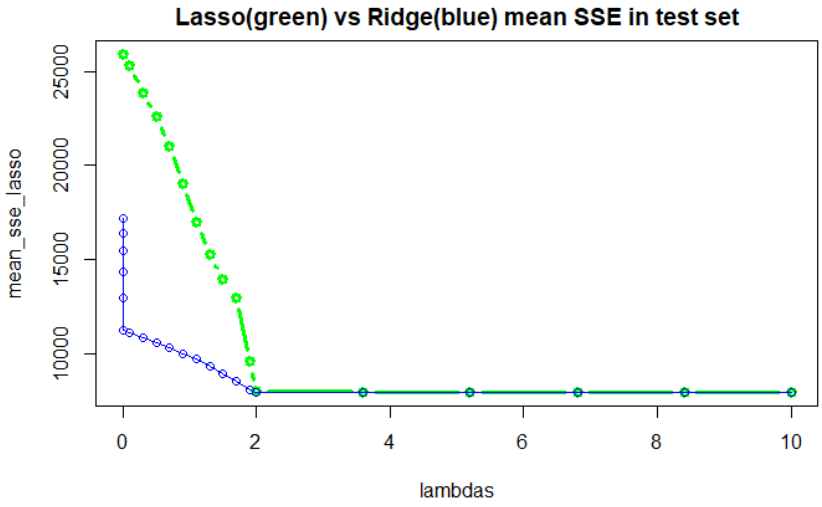

Figure 5: Coefficients of the Best Predictive Model

4 Limitation

One of the main limitations of our analysis is that we dropped observations with missing data and all observations with an answer of either "refused" or "unknown" in the process of data cleaning. Since some of the observations which we have dropped might have an important influence on the distribution of predictors and their association to Sales, there might be some bias in our analysis. There are generally 2 strategies in handling missing values in Data Science: deletion and imputation. However, data imputation is beyond the scope of this course and is covered in Time Series

Furthermore, although we have started our analysis with some hypothesis generation and tried to make sense of which predictors are the most meaningful in predicting sales, there might be some associations and confounding factors that are not considered in this paper. For instance, there might be some confounding factors between `Outlet_Type` and `Outlet_Location_Type`, since large stores tend to populate urban areas or more hectic places, while small groceries are generally found in suburbans.

Moreover, an important predictor variable that was missing from our data is time. This is an imperative factor in determining the type of items sold and the quantity, since sales are cyclical. If we had a time variable we could have created a more complex model, which involves time trends of sales and identify periods of enhanced consumerism as well as identify `Item_Types` which are more sensitive to fluctuations in sales and thus, more elastic from an economic perspective. However, our data source did not provide additional information on the collected data so it does not allow for such study.

Other weaknesses of our analysis include: 1) the data was collected from an observational study instead of a randomized experimental trial, so it is difficult to provide causal inference between the predictors and Sales. The analysis reduces to observing associations.

5 Conclusion

Overall we conclude that the results of the linear Model constructed through a stepwise procedure in both direction has the best performance (measured by mean MSE, R-squared and minimum AIC). Although Lasso and Ridge and a mix between them approach Modelstep performance-wise, additional shrinkage doesn't seem to add more predictive power, this further supports the idea that Modelstep already selected for the best predictors.

Our best predictive model can be helpful in making strategic plans for starting a retail business. Our analysis suggests the importance of pricing and providing inventory for businesses, as well as providing a high-quality service within the store in terms of ambiance, and access to products are the most important factors in increasing sales and revenue from a retail businesses. Our study can set the platform for new questions that sales forecasting specialists can investigate. For instance, how important and effective advertising actually is in increasing the sales volume? This would be a more complex question to answer and in future analysis which requires a more rigorous modelling and larger data.

Retail is another industry which extensively uses analytics to optimize business processes. Tasks like product placement, inventory management, customized offers, product bundling, etc. are being smartly handled using data science techniques.

6 Acknowledgement

We would like to thank professor Kevin Rader, preceptor Julie Vu and our Teaching Fellow Kathryn McKeough for all their contribution to this project, for their guidance and support throughout the cour.

7 References

1. *Prabhakaran, Selva. "Missing Value Treatment." DataScience , 25 Apr. 2016, datascienceplus.com/missing-value-treatment/.*

2. *"Practice Problem: Big Mart Sales III — Knowledge and Learning." DataHack : Biggest Data Hackathon Platform for Data Scientists, datahack.analyticsvidhya.com.*

3.*Ramsey, Fred L., and Daniel W. Schafer. The Statistical Sleuth: a Course in Methods of Data Analysis. Brooks/Cole, Cengage Learning, 2013.*

4. *Roos, Dave. "How Sales Forecasting Works." HowStuffWorks, HowStuffWorks, 20 June 2008, money.howstuffworks.com/sales-forecasting1.htm.* 5. *Libraries in R: "jtools", "broom", "Hmisc"*

8 TABLES

8.1 Predictors

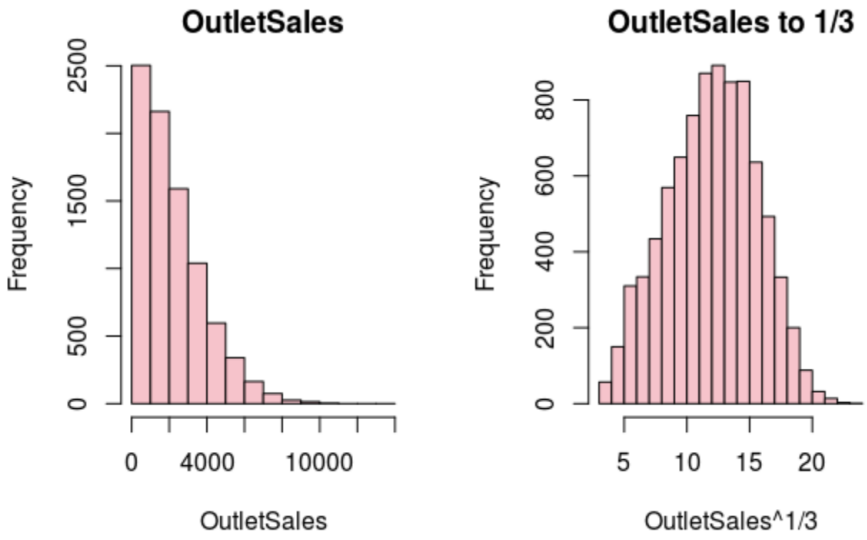

Figure 6: Outlet Sales Un-transformed vs Transformed

8.2 Normality and Symmetry

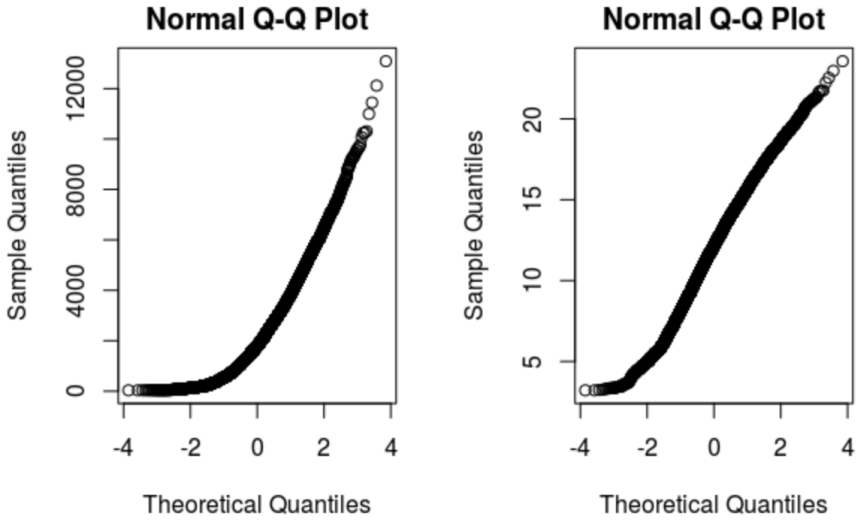

Figure 7: QQnorm Plot of Outlet Sales

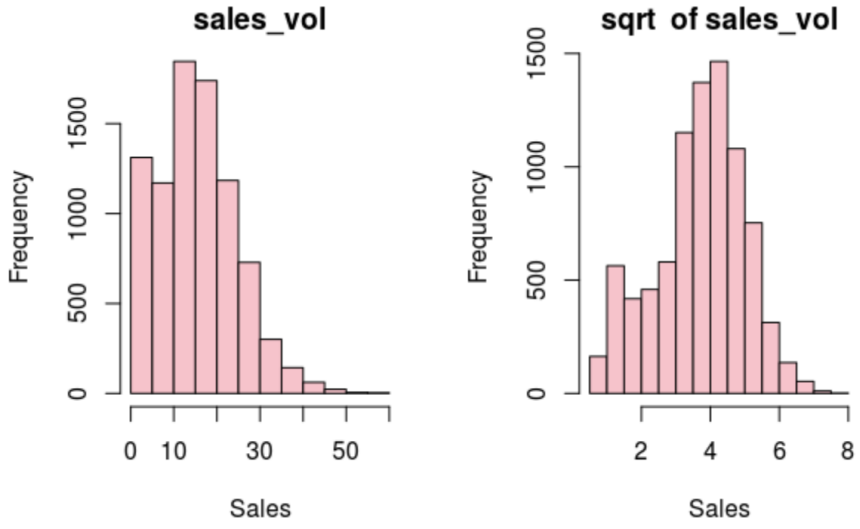

Figure 8: Distributions of Sales Volume

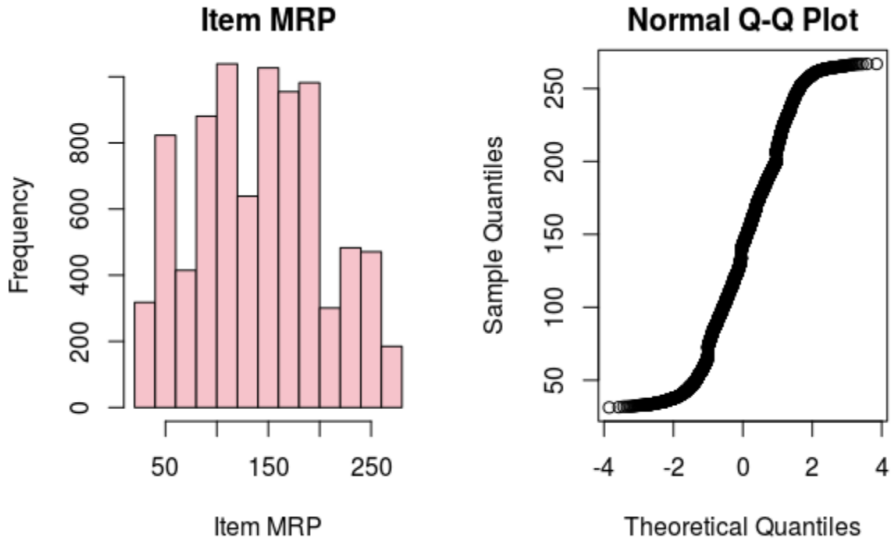

Figure 9: Item MRP distribution

8.3 Linearity

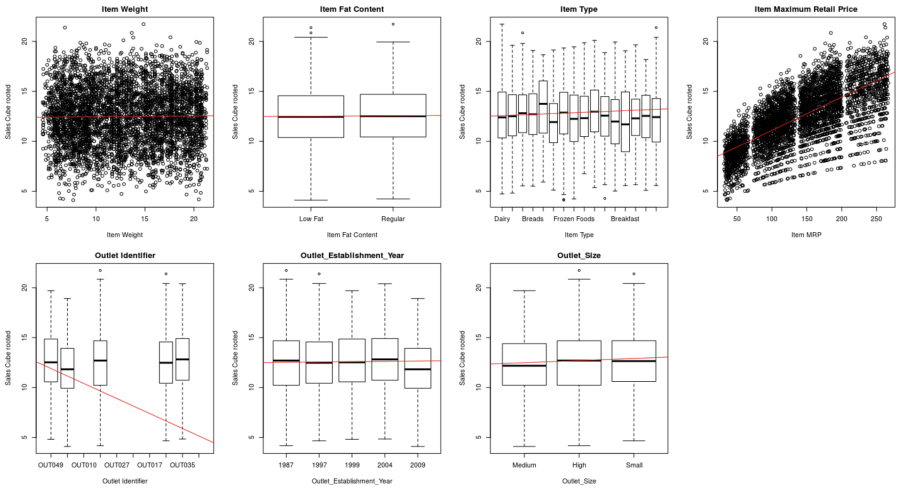

Figure 10: Linearity between all predictors and response

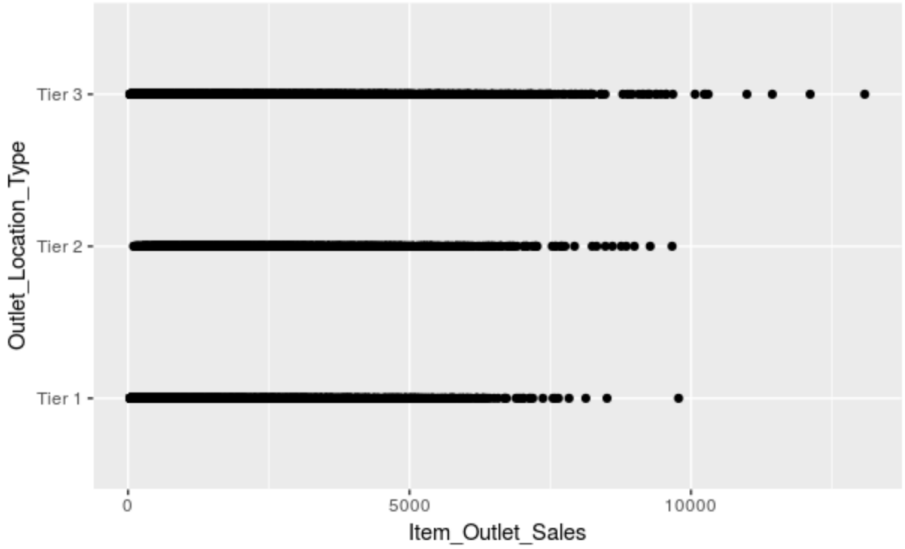

Figure 11: Location and Sales

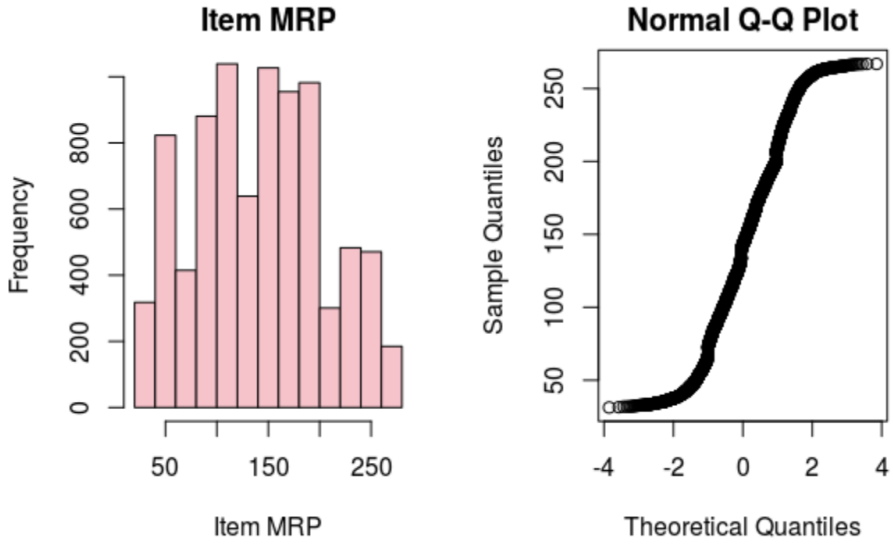

Figure 12: MRP and Sales

```
 16  Variables      8519  Observations
--------------------------------------------------------------------------------------------------------------
----------------
Item_Identifier
       n  missing distinct
    8519        0     1555

lowest : DRA12 DRA24 DRA59 DRB01 DRB13, highest: NCZ30 NCZ41 NCZ42 NCZ53 NCZ54
--------------------------------------------------------------------------------------------------------------
----------------
Item_Weight
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
.90      .95
    8519        0      415        1    12.88    5.362    5.940    6.675    8.785   12.650   16.850
19.350   20.250

lowest :  4.555  4.590  4.610  4.615  4.635, highest: 21.000 21.100 21.200 21.250 21.350
--------------------------------------------------------------------------------------------------------------
----------------
Item_Fat_Content
       n  missing distinct
    8519        0        3

Value       Low Fat not_food  Regular
Frequency      3917     1599     3003
Proportion    0.460    0.188    0.353
--------------------------------------------------------------------------------------------------------------
----------------
Item_Visibility
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
.90      .95
    8519        0     7876        1  0.06611  0.05579  0.00000  0.01204  0.02698  0.05392  0.09456
0.13949  0.16357

lowest : 0.000000000 0.003574698 0.003589104 0.003597678 0.003599378
highest: 0.309390255 0.311090379 0.321115010 0.325780807 0.328390948
--------------------------------------------------------------------------------------------------------------
----------------
Item_Type
       n  missing distinct
    8519        0       16

Baking Goods (647, 0.076), Breads (251, 0.029), Breakfast (110, 0.013), Canned (649, 0.076), Dairy
(681, 0.080),
Frozen Foods (855, 0.100), Fruits and Vegetables (1232, 0.145), Hard Drinks (214, 0.025), Health and
Hygiene (520,
0.061), Household (910, 0.107), Meat (425, 0.050), Others (169, 0.020), Seafood (64, 0.008), Snack
Foods (1199,
0.141), Soft Drinks (445, 0.052), Starchy Foods (148, 0.017)
--------------------------------------------------------------------------------------------------------------
----------------
Item_MRP
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
.90      .95
    8519        0     5936        1      141    71.55    42.51    52.80    93.84   143.05   185.68
231.21   250.78

lowest :  31.2900  31.4900  31.8900  31.9558  32.0558, highest: 266.1884 266.2884 266.5884 266.6884
```

Figure 13: Summary of Categorical and Numerical Predictors

8.4 Models

8.4.1 Model on all main effects

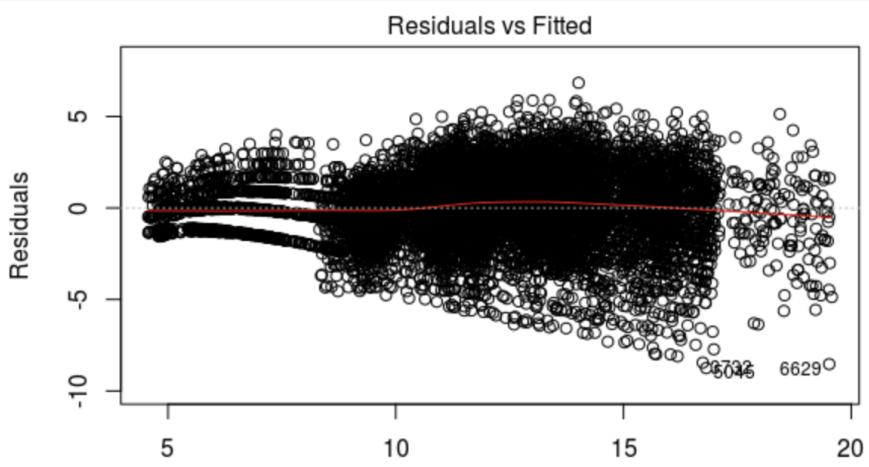

Figure 14: Residuals of Model Main Effects

8.4.2 Cross Validation

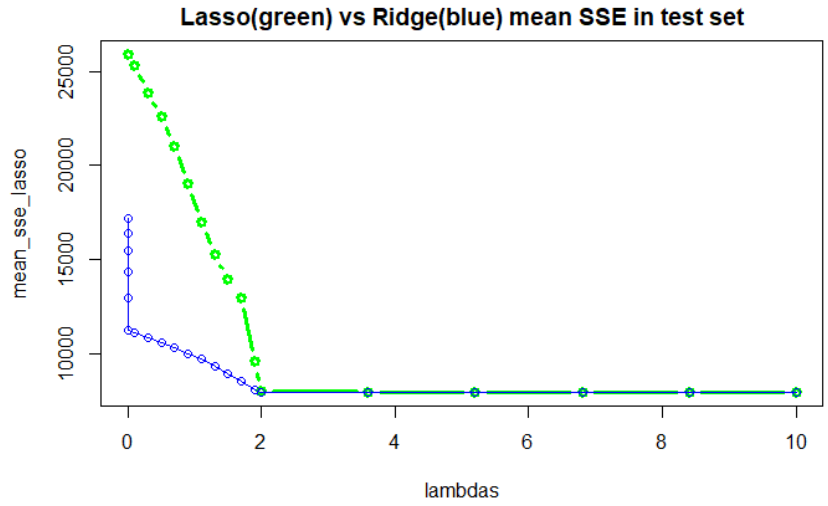

Figure 15: Cross Validation Output

8.4.3 Best Predictive Model

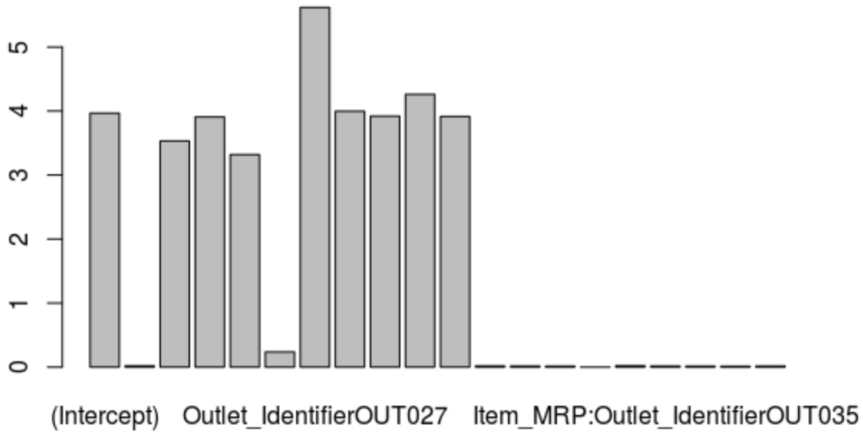

Figure 16: Coefficients of the Best Predictive Model

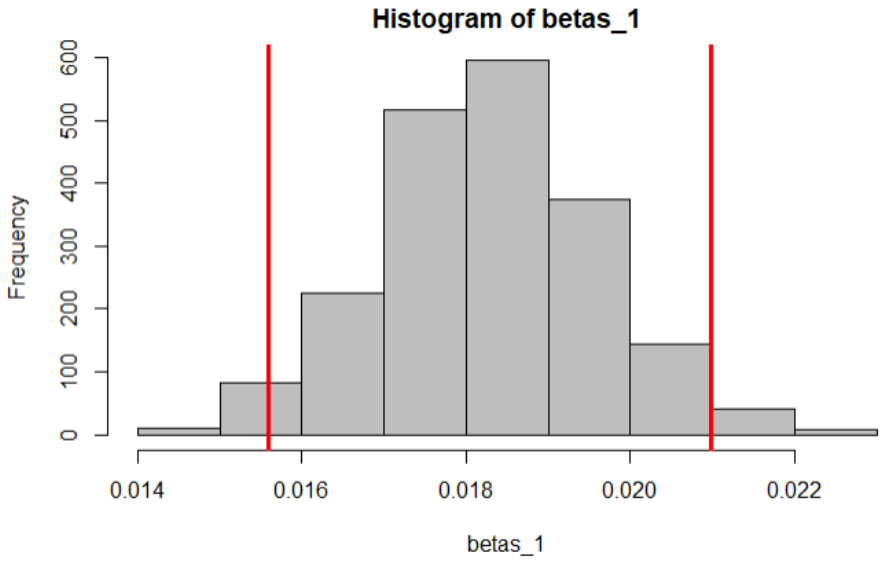

Figure 17: Beta of Item.MRP distribution in Modelstep

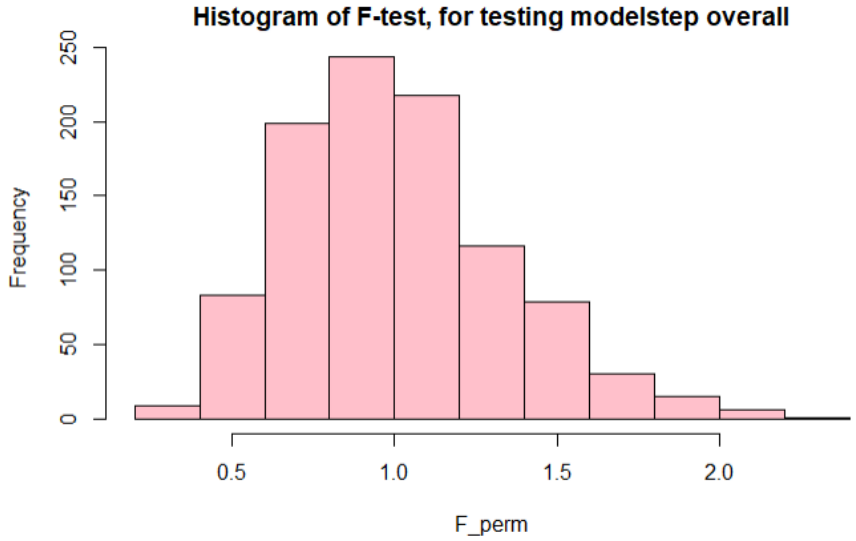

Figure 18: Permutation F-test for testing **Modelstep** overall

```
[1] 8233.432 7988.113
       lambdas mean_sse_ridge mean_sse_el2 mean_sse_el4 mean_sse_el6 mean_sse_el8 mean_sse_lasso
 [1,]  0.00010      17201.849     25917.509     25919.746     25919.746     25919.746     25919.746
 [2,]  0.00208      16415.011     25118.990     25919.746     25919.746     25919.746     25919.746
 [3,]  0.00406      15478.852     24035.638     25919.746     25919.746     25919.746     25919.746
 [4,]  0.00604      14350.608     21763.276     25919.746     25919.746     25919.746     25919.746
 [5,]  0.00802      12972.565     18549.360     23477.261     25919.746     25919.746     25919.746
 [6,]  0.01000      11254.386     15162.427     17656.028     20886.366     23450.934     25912.553
 [7,]  0.10000      11131.551     14957.285     17255.454     20339.455     22973.004     25321.513
 [8,]  0.30000      10878.369     14535.484     16465.812     19061.481     21956.288     23869.508
 [9,]  0.50000      10613.867     14075.912     15707.838     17816.177     20365.220     22579.089
[10,]  0.70000      10335.149     13580.831     15004.299     16669.187     18759.630     21010.338
[11,]  0.90000      10039.251     13013.131     14362.433     15608.618     17109.017     19059.664
[12,]  1.10000       9720.003     12425.224     13792.098     14644.020     15695.080     16970.489
[13,]  1.30000       9370.042     11595.440     13231.346     13825.068     14525.682     15268.525
[14,]  1.50000       8979.404     10531.963     12397.968     13165.995     13536.911     13984.845
[15,]  1.70000       8541.375      9321.885     10631.423     12052.234     12720.788     13008.403
[16,]  1.90000       8104.713      8239.536      8443.096      8736.223      9151.515      9587.715
[17,]  2.00000       7993.533      7993.814      7995.934      7999.382      8003.466      8008.469
[18,]  3.60000       7992.603      7992.260      7993.342      7995.580      7998.178      8001.444
[19,]  5.20000       7991.721      7990.904      7991.302      7992.583      7994.079      7995.997
[20,]  6.80000       7990.975      7989.896      7989.782      7990.341      7991.031      7991.986
[21,]  8.40000       7990.259      7989.228      7988.870      7988.951      7989.077      7989.351
[22,] 10.00000       7989.592      7988.931      7988.640      7988.612      7988.599      7988.595
```

Figure 19: Table of Mean Squared Error for Lasso-Ridge Model **Modelstep** overall