

# **Demand Prediction in Retail**

A Practical Guide to Leverage Data and Predictive Analytics

Maxime C. Cohen, Paul-Emile Gras, Arthur Pentecoste, Renyu Zhang

<b>I/ Introduction</b>	<b>5</b>
1. Motivation	5
2. Dataset	7
3. Objective and Scope	12
3.1. Training and Test Data	12
3.2. Prediction Accuracy Metrics	13
3.3. Application	14
<b>II/ Data Pre-Processing and Modeling Factors</b>	<b>16</b>
1. Dealing with Missing Data	16
2. Testing for Outliers	19
3. Accounting for Time Effects	22
4. Price and Lag-Prices	23
5. Featured on Main Page	25
6. Item Descriptive Features	25
7. Additional Features	26
8. Scaling	26
9. Sorting and Exporting the Dataset	27
<b>III/ Common Demand Prediction Methods</b>	<b>29</b>
1. Primer: Basic Linear Regression for One SKU	29
2. Structuring the Dataset	31
3. Centralized Approach	32
4. Decentralized Approach	34
5. Feature Selection and Regularization	35
5.1. Subset Selection	35
5.1.1. Presentation of Subset Selection	35
5.1.2. AIC and BIC	36
5.2. Lasso Regularization	41
5.2.1. Lasso for One SKU	41
5.2.2. Decentralized Lasso	43
5.3. Ridge Regularization	44
5.3.1. Ridge Model	44
5.3.2. Decentralized Ridge	44
5.4. Elastic Net Regularization	45
5.4.1. Elastic Net Model	45
5.4.2. Decentralized Elastic Net	46

6.	Log Transformations	47
6.1.	Log-Transformation on the Price Variable	47
6.2.	Log-Transformation on the Target Variable	50
6.3.	Transformations and Prediction Accuracy	51
7.	Centralized Approach with SKU-Fixed Effects	53
8.	Centralized Approach with Price-Fixed Effects	56
9.	Centralized Approach with SKU-Price-Fixed Effects	59
10.	Decentralized Approach with Aggregated Seasonality	61
11.	Summary and Next Steps	63
<b>IV/</b>	<b>Tree-Based Methods</b>	<b>64</b>
1.	Decision Tree	65
1.1.	Centralized Decision Tree	65
1.1.1.	Selecting the Parameters	66
1.1.2.	Focusing on the Best Model	70
1.1.3.	Example of a Plotted Tree	70
1.2.	Decentralized Decision Tree	72
1.2.1.	Selecting the Parameters	72
1.2.2.	Focusing on the Best Model	73
2.	Random Forest	75
2.1.	Centralized Random Forest	75
2.1.1.	Selecting the Parameters	75
2.1.2.	Focusing on the Best Model	77
2.2.	Decentralized Random Forest	77
2.2.1.	Selecting the Parameters	77
2.2.2.	Focusing on the Best Model	78
3.	Gradient Boosted Tree	79
3.1.	Centralized Gradient Boosted Tree	80
3.1.1.	Selecting the Parameters	80
3.1.2.	Focusing on the Best Model	81
3.2.	Decentralized Gradient Boosted Tree	82
3.2.1.	Finetuning the Parameters	82
3.2.2.	Focusing on the Best Model	83
4.	Methods Comparison	84
<b>V/</b>	<b>Clustering Techniques</b>	<b>85</b>
1.	K-means Clustering	85
1.1.	Description of K-means Clustering	85

1.2.	Clustering using Average Price and Weekly Sales	89
1.3.	Adding Standard Deviations of the Clustering Features	92
2.	DBSCAN Clustering	95
2.1.	Description of DBSCAN Clustering	95
2.2.	Clustering using Average Price and Weekly Sales	100
2.3.	Adding the Standard Deviation of the Clustering Features	104
<b>VI/ Evaluation and Visualization</b>		105
1.	Summary of Results	105
2.	Prediction vs. Actual	107
3.	Varying the Split Ratio	111
<b>VII/ More Advanced Methods</b>		116
1.	The Prophet Method	116
1.1.	What is the Prophet Method?	116
1.1.1.	How it Works	116
1.1.2.	Illustration for One SKU	118
1.2.	Forecasting with Prophet	121
1.2.1.	Univariate Time-Series	121
1.2.2.	Adding Features	124
2.	Data Aggregation and Demand Prediction	128
2.1.	Presentation of the DAC Method	128
2.2.	Fine Tuning the Hyperparameters	130
2.3.	Interpretating the DAC Results	132
<b>VIII/ Conclusion and Advanced Topics</b>		134
<b>References</b>		138
<b>Acknowledgements</b>		141
<b>Authors Bios</b>		141
<b>Endorsements</b>		142