University of Massachusetts, Lowell
Manning School of Business
MS in Business Analytics Capstone Project
Spring 2020

Analyzing Variability in Actual Demand vs forecast of Dell's Hardware Products

Authors: Modupe Ajala, Elahd Hain, and Favour Tejuosho

Table of Contents

Acknowledgement

We would like to use this opportunity to appreciate several persons who made this project possible.

Thank you to Joseph Collins the Director of Global Supply Chain Operations at Dell EMC for sponsoring this Business Analytics capstone project, providing background information about the organization, and presenting the project proposal to the class. We are grateful for his feedback, time, and efforts he put into making this project a success.

We would also like to thank Ram Das, a senior Data Intelligence Engineer at Dell Technologies. Ram walked us through the dataset and scope of the project. He also met with us weekly to answer all our questions and provide insights and leadership for this project. We are grateful for his prompt responses and relentless assistance.

Finally, we would like to thank Professor Luvai Motiwalla of the Operations and Information Systems Department, in the Manning School of Business at the University of Massachusetts Lowell. Professor Motiwalla, who was our project liaison, provided guidance from the pilot to the end of our project and was readily available to answer our many questions. For all of this, we are truly grateful.

Business Understanding

Dell was founded by Michael Dell in 1984 at Austin TX. Dell is an American multinational computer technology company that develops, sells, repairs, and supports computers and related products and services. The company is one of the largest technological corporations in the world, employing more than 145,000 people in the U.S. and around the world.

The company is well known for its innovation in Supply Chain Management and Electronic Commerce. The company was a pure hardware vendor until it acquired Perot Systems in 2009 and entered the market for IT services. Dell aims to expand its portfolio from offering computers only to delivering complete solutions for enterprise customers.

Dell acquired the enterprise technology firm EMC Corporation for the record-breaking sum of $67 billion dollars in 2015, following the completion of the purchase, Dell and EMC became divisions of Dell Technologies.

Dell EMC

Dell EMC is an American multinational technology company that offers products and services across all areas of computing, networking, and storage. The company was formed when Dell acquired EMC Corporation along with its subsidiaries including VMware, RSA Security, and Pivotal.

Dell EMC headquarters are in Hopkinton, Massachusetts. Dell EMC's target market is businesses of all sizes. The merged company combines the two former entities' positions as providers of personal computers, servers, storage, and virtualization products. Areas of focus include software-defined data center, hybrid cloud, converged infrastructure, mobile computing, and security.

Dell's Supply Chain & Data Analytics Center (SCDnA) aims to explore digital supply chain initiatives, including how suppliers build products, how suppliers are integrated into their processes, and how products perform in the field to establish predictive analytics across the entire life cycle of Dells products, thereby enabling customers. At any given time, the program has any number of different analytic initiatives going on. *(Handfield, R. 2018)*

Project Objectives

Dell's Global Operation Storage and Services Supply Chain Data & Analytics is interested in an Analytical model to establish stock-outs risks for purchased parts required to fill customer demand over a rolling 8-week window. The ability to predict a stock-out benefits Dell's customers by allowing supply chain to mitigate or balance expedite expenses vs increasing the lead time of products at the time of customer quote or order. The project team aims to use both descriptive analytics and prescriptive analytics to solve the following objectives laid out by Dell;

- Identify gaps or variability in the predicted quantity of materials demanded (MRP demand) vs actual quantity of materials demanded.
- Analyze various factors affecting the variability.
- Prescribe solutions to fix the variability.

The project team created an analytical model to find the disparity in the predicted quantity of materials demanded (which has not booked) with actual demand (which has booked). They visually represented the variability by Region, Plant, Fiscal year and so on, and gave recommendations based on their findings.

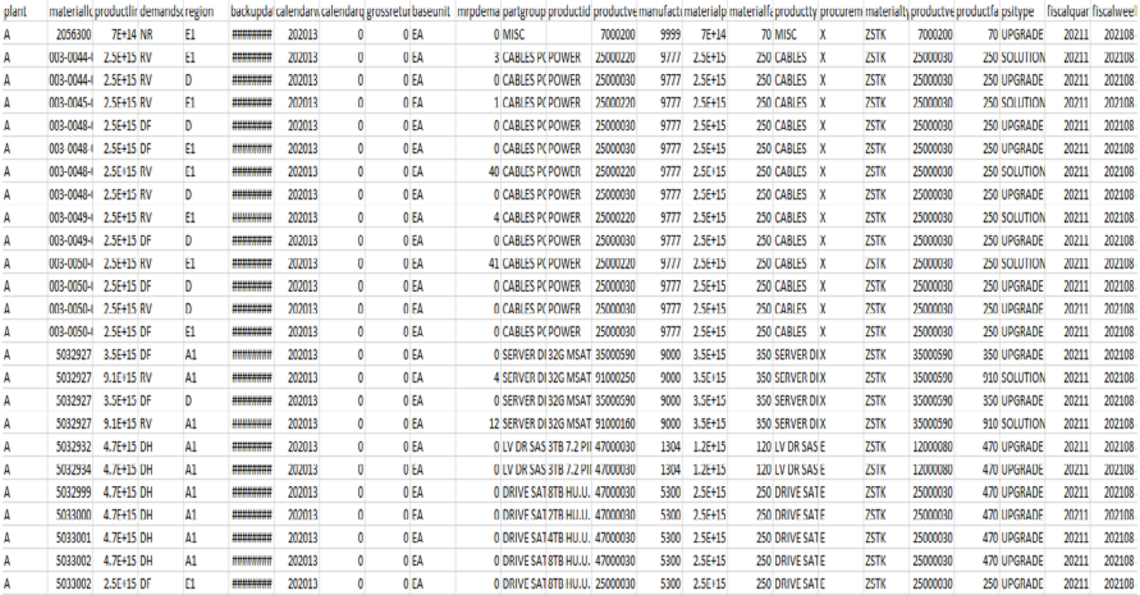
Data understanding

Each week dell provided our team with two different datasets. The first dataset was an MRP dataset. This dataset contained 25 different variables each describing an item's specific characteristic. The MRP dataset contained 350,000 records in each week's file. Each record in this dataset contained information on everything from the product description to what region/plant the product was created in. The most important variable that we were given in the MRP dataset was **gross return**. This variable allowed us to understand exactly how much of each product Dell predicted would be in demand in the coming week.

When looking at the characteristics of this dataset, we noticed that the majority of the 25 variables happened to be categorical, with a few of them being integer or other datatypes. This meant that we were left with the responsibility of deciding which of these variables would be more important for our analysis. This was mostly because our data needed to be converted into dummy variables, since none of the values in each of the categories had any mathematical purpose. On top of this, most of the categories had a very wide range of values, which meant we would have to cut down on the number of category choices.

This process of elimination was necessary, because if we had left everything alone, there would have been an immense number of dummy variables, and it would have required more computing power than what could have been provided. In fact, at one point we had tried running our models through Collab notebooks on google drive without cutting down on the number of values, and it consistently led to our notebook crashing in the middle of the process. We will go into this process of selecting variables in detail in the Data Preprocessing section.

Figure 1 displays a snapshot of the MRP dataset, and gives an example the values within each of the 25 variables

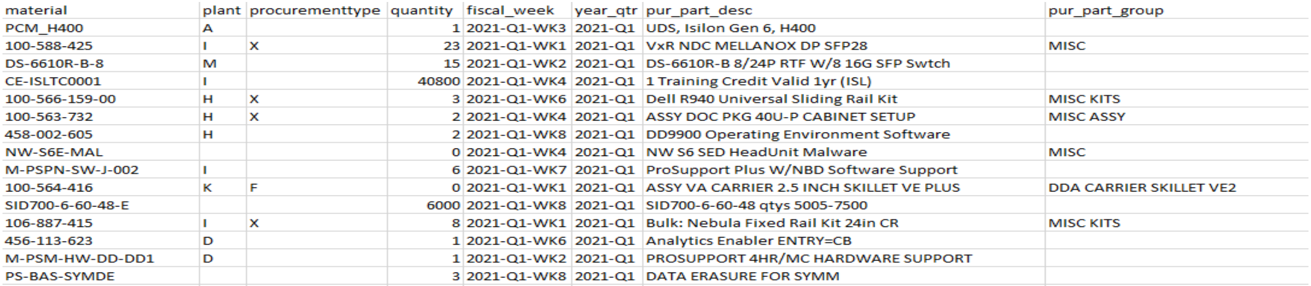


plant	materiallc	productlir	demandsc	region	backupda	calendarw	calendarq	grossretur	baseunit	mrpdema	partgroup	productid	productve	manufact	materialp	materialfa	productty	procurem	materialty	productve	productfa	psitype	fiscalquar	fiscalwee
A	2056300	7E+14	NR	E1	########	202013	0	0	EA	0	MISC		7000200	9999	7E+14	70	MISC	X	ZSTK	7000200	70	UPGRADE	20211	202108
A	003-0044-	2.5E+15	RV	E1	########	202013	0	0	EA	3	CABLES PC	POWER	25000220	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	SOLUTION	20211	202108
A	003-0044-	2.5E+15	RV	D	########	202013	0	0	EA	0	CABLES PC	POWER	25000030	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	UPGRADE	20211	202108
A	003-0045-	2.5E+15	RV	E1	########	202013	0	0	EA	1	CABLES PC	POWER	25000220	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	SOLUTION	20211	202108
A	003-0048-	2.5E+15	DF	D	########	202013	0	0	EA	0	CABLES PC	POWER	25000030	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	UPGRADE	20211	202108
A	003-0048-	2.5E+15	DF	E1	########	202013	0	0	EA	0	CABLES PC	POWER	25000030	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	UPGRADE	20211	202108
A	003-0048-	2.5E+15	RV	E1	########	202013	0	0	EA	40	CABLES PC	POWER	25000220	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	SOLUTION	20211	202108
A	003-0048-	2.5E+15	RV	D	########	202013	0	0	EA	0	CABLES PC	POWER	25000030	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	UPGRADE	20211	202108
A	003-0049-	2.5E+15	RV	E1	########	202013	0	0	EA	4	CABLES PC	POWER	25000220	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	SOLUTION	20211	202108
A	003-0049-	2.5E+15	DF	D	########	202013	0	0	EA	0	CABLES PC	POWER	25000030	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	UPGRADE	20211	202108
A	003-0050-	2.5E+15	RV	E1	########	202013	0	0	EA	41	CABLES PC	POWER	25000220	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	SOLUTION	20211	202108
A	003-0050-	2.5E+15	DF	D	########	202013	0	0	EA	0	CABLES PC	POWER	25000030	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	UPGRADE	20211	202108
A	003-0050-	2.5E+15	RV	D	########	202013	0	0	EA	0	CABLES PC	POWER	25000030	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	UPGRADE	20211	202108
A	003-0050-	2.5E+15	DF	E1	########	202013	0	0	EA	0	CABLES PC	POWER	25000030	9777	2.5E+15	250	CABLES	X	ZSTK	25000030	250	UPGRADE	20211	202108
A	5032927	3.5E+15	DF	A1	########	202013	0	0	EA	0	SERVER DI	32G MSAT	35000590	9000	3.5E+15	350	SERVER DI	X	ZSTK	35000590	350	UPGRADE	20211	202108
A	5032927	9.1E+15	RV	A1	########	202013	0	0	EA	4	SERVER DI	32G MSAT	91000250	9000	3.5E+15	350	SERVER DI	X	ZSTK	35000590	910	SOLUTION	20211	202108
A	5032927	3.5E+15	DF	D	########	202013	0	0	EA	0	SERVER DI	32G MSAT	35000590	9000	3.5E+15	350	SERVER DI	X	ZSTK	35000590	350	UPGRADE	20211	202108
A	5032927	9.1E+15	RV	A1	########	202013	0	0	EA	12	SERVER DI	32G MSAT	91000160	9000	3.5E+15	350	SERVER DI	X	ZSTK	35000590	910	SOLUTION	20211	202108
A	5032932	4.7E+15	DH	A1	########	202013	0	0	EA	0	LV DR SAS	3TB 7.2 PII	47000030	1304	1.2E+15	120	LV DR SAS	E	ZSTK	12000080	470	UPGRADE	20211	202108
A	5032934	4.7E+15	DH	A1	########	202013	0	0	EA	0	LV DR SAS	3TB 7.2 PII	47000030	1304	1.2E+15	120	LV DR SAS	E	ZSTK	12000080	470	UPGRADE	20211	202108
A	5032999	4.7E+15	DH	A1	########	202013	0	0	EA	0	DRIVE SAT	8TB HU.U.	47000030	5300	2.5E+15	250	DRIVE SAT	E	ZSTK	25000030	470	UPGRADE	20211	202108
A	5033000	4.7E+15	DH	A1	########	202013	0	0	EA	0	DRIVE SAT	2TB HU.U.	47000030	5300	2.5E+15	250	DRIVE SAT	E	ZSTK	25000030	470	UPGRADE	20211	202108
A	5033001	4.7E+15	DH	A1	########	202013	0	0	EA	0	DRIVE SAT	4TB HU.U.	47000030	5300	2.5E+15	250	DRIVE SAT	E	ZSTK	25000030	470	UPGRADE	20211	202108
A	5033002	4.7E+15	DH	A1	########	202013	0	0	EA	0	DRIVE SAT	8TB HU.U.	47000030	5300	2.5E+15	250	DRIVE SAT	E	ZSTK	25000030	470	UPGRADE	20211	202108
A	5033002	2.5E+15	DF	E1	########	202013	0	0	EA	0	DRIVE SAT	8TB HU.U.	25000030	5300	2.5E+15	250	DRIVE SAT	E	ZSTK	25000030	250	UPGRADE	20211	202108

Figure 1: Snapshot of MRP Dataset

The second dataset received from Dell was the Bookings dataset. This dataset provides both the characteristics and number of actual sales for each item during the week. Just like the MRP dataset each item was separated based uniqueness among all 8 of the variables. This meant you could potentially see two records that were almost identical, but may have had different plant values, which would have caused a separation in the records. In total there were around 34,000 records in each dataset, with each record containing a total of 8 different variables. In the bookings dataset the most important variable was quantity which was the variable that showed the actual amount sold.

Figure 2 displays a snapshot of the Bookings dataset and gives an example the values within each of the 8 variables.



material	plant	procurementtype	quantity	fiscal_week	year_qtr	pur_part_desc	pur_part_group
PCM_H400	A		1	2021-Q1-WK3	2021-Q1	UDS, Isilon Gen 6, H400	
100-588-425	I	X	23	2021-Q1-WK1	2021-Q1	VxR NDC MELLANOX DP SFP28	MISC
DS-6610R-B-8	M		15	2021-Q1-WK2	2021-Q1	DS-6610R-B 8/24P RTF W/8 16G SFP Swtch	
CE-ISLTC0001	I		40800	2021-Q1-WK4	2021-Q1	1 Training Credit Valid 1yr (ISL)	
100-566-159-00	H	X	3	2021-Q1-WK6	2021-Q1	Dell R940 Universal Sliding Rail Kit	MISC KITS
100-563-732	H	X	2	2021-Q1-WK4	2021-Q1	ASSY DOC PKG 40U-P CABINET SETUP	MISC ASSY
458-002-605	H		2	2021-Q1-WK8	2021-Q1	DD9900 Operating Environment Software	
NW-S6E-MAL			0	2021-Q1-WK4	2021-Q1	NW S6 SED HeadUnit Malware	MISC
M-PSPN-SW-J-002	I		6	2021-Q1-WK7	2021-Q1	ProSupport Plus W/NBD Software Support	
100-564-416	K	F	0	2021-Q1-WK1	2021-Q1	ASSY VA CARRIER 2.5 INCH SKILLET VE PLUS	DDA CARRIER SKILLET VE2
SID700-6-60-48-E			6000	2021-Q1-WK8	2021-Q1	SID700-6-60-48 qtys 5005-7500	
106-887-415	I	X	8	2021-Q1-WK1	2021-Q1	Bulk: Nebula Fixed Rail Kit 24in CR	MISC KITS
456-113-623	D		1	2021-Q1-WK6	2021-Q1	Analytics Enabler ENTRY=CB	
M-PSM-HW-DD-DD1	D		1	2021-Q1-WK2	2021-Q1	PROSUPPORT 4HR/MC HARDWARE SUPPORT	
PS-BAS-SYMDE			3	2021-Q1-WK8	2021-Q1	DATA ERASURE FOR SYMM	

Figure 2: Snapshot of Bookings Dataset

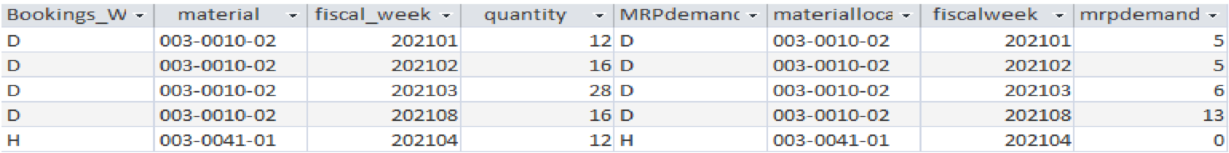
Together these two datasets would provide the information necessary to begin creating the descriptive analytics and modeling that were needed to accomplish the project objectives.

Data preprocessing

One of the lengthiest parts of dealing with these datasets was the preprocessing step. Right after Dell provided the first dataset, we immediately noticed a few things that would need to be cleaned, for our programs to process the data and create visualizations and models. The first step we decided to take was to clean up all the nulls and remove the duplicate values. To accomplish this task, we needed to use multiple tools. For example, we used Python to eliminate null values and we used Access and Python together to handle getting rid of any of the duplicate records. Python was especially helpful when it came to this process, since it provided an automated procedure for analyzing and cleaning the dataset each week as it came in. This meant that a good amount of the cleaning process took part in the first week and it was a much more streamlined process in the following weeks. On the other hand, removing the duplicate records was a less streamlined process. These duplicate records created a larger problem as we continued to compile each week's datasets into the master files for both the MRP and Bookings data. We found that the new MRP datasets would sometimes contain data that could be found in previous weeks, which meant we had to pull just the distinct records through using a *select distinct* query.

After completing the data cleaning process, we were left with the task of merging the MRP and Bookings datasets. A merged file would allow us to accurately compare the actual vs predicted quantity of sales for the week. In order to merge the files, we needed to discover a unique key for each record that could be found in both datasets. Through analyzing the data in python, we were able to identify a few variables that when used together, met the criteria of creating a unique key. These variables were fiscal week, material allocation/material (which was a unique code for each material), and plant. Through setting these variables as the unique key, we were able to merge the two master files to create a single document that contained all the data we needed for each material.

Finally, once the merging process had been completed, we decided the only thing left to do was to eliminate some of the variables. In order to complete this process, we had to meet with our partners at Dell to further investigate the purpose of each column. Through using the information, we collected from speaking with Dell, we were able to separate the variables into two categories, those that provided value and those that just added noise to our dataset. The type of variables that could be found in the bulk category were ones that provided repetitive information (e.g. material product line and material family group) or provided no meaningful information to our task (e.g. calendar week). In order to simplify the dataset, we ended up carefully removing some of these columns so that it would become a much more manageable dataset. After completing this task, we were ready to go ahead with analyzing the dataset and running it through the models.



Bookings_W	material	fiscal_week	quantity	MRPdemanc	materialloca	fiscalweek	mrpdemand
D	003-0010-02	202101	12	D	003-0010-02	202101	5
D	003-0010-02	202102	16	D	003-0010-02	202102	5
D	003-0010-02	202103	28	D	003-0010-02	202103	6
D	003-0010-02	202108	16	D	003-0010-02	202108	13
H	003-0041-01	202104	12	H	003-0041-01	202104	0

Figure 3: Snapshot preprocessed data

Exploratory Analysis

After the data cleaning and preprocessing, we carried out some models and data visualizations in Tableau and Python, which allowed us discover patterns, anomalies, and check assumptions for our dataset. Figures 4 through 9 are visualizations created in Tableau. Note the varying scales in both the Y and X axis's as they alter according to the data shown.

Figure 1 below shows a side by side comparison of the percentage of quantity booked (actual demand) and MRP demand (predicted demand) by plant. For both actual quantities booked and predicted demand, Plant D accounts for the highest percentage of records, with 57.4% percent of materials booked and 45% of predicted demand coming from Plant D.

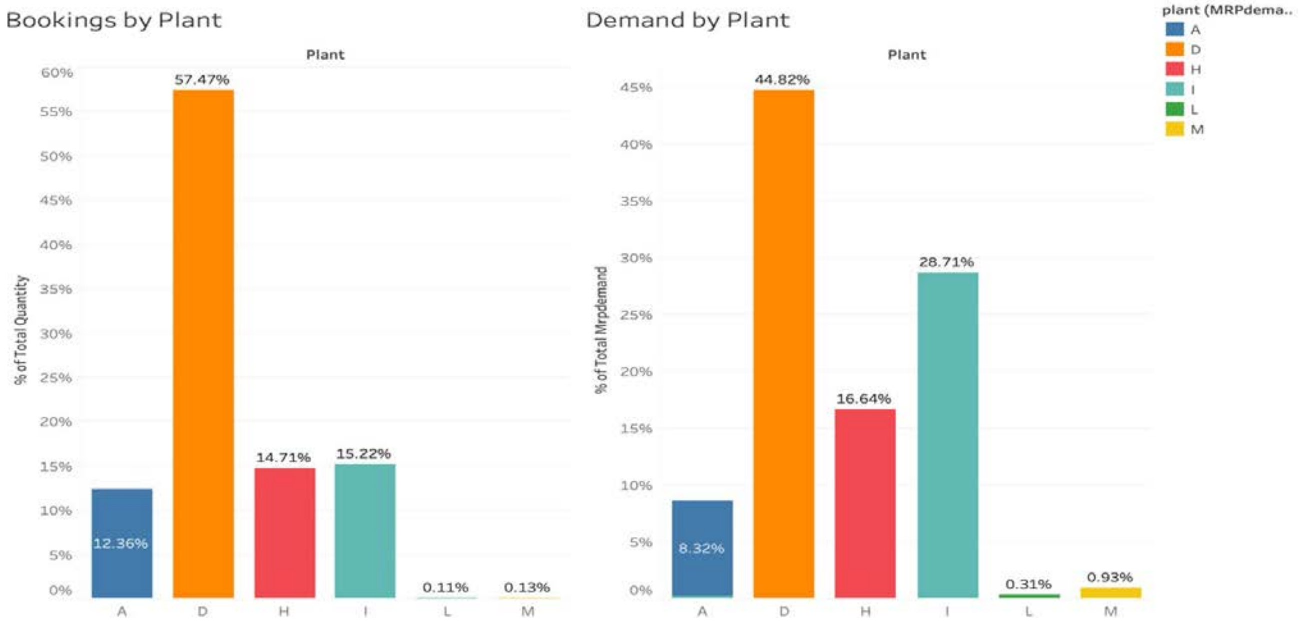


Figure 4

The next six charts gives an overview and breakdown of the variability of MRP demand and quantity booked.

Figure 5 is an overview of MRP demand to quantity booked by fiscal week and plant. What this implies is that; the flatter the plots, the smaller the variance, and the more scattered the plot the greater the variance. This is because, the x-axis and the y-axis are at different scales. Actually, the y-axis has such a large scale that any movement away from the x-axis most certainly means that there is a large gap between the predicted and actual demand for a product. From this graph, we noticed a high variance between MRP demand and quantity booked in fiscal week 1 (202101). This is quite expected in the first week of the year or quarter because it is when the models are first being trained. Consequently, as we move to later fiscal weeks, predicted quantities closely match with actual quantities booked.

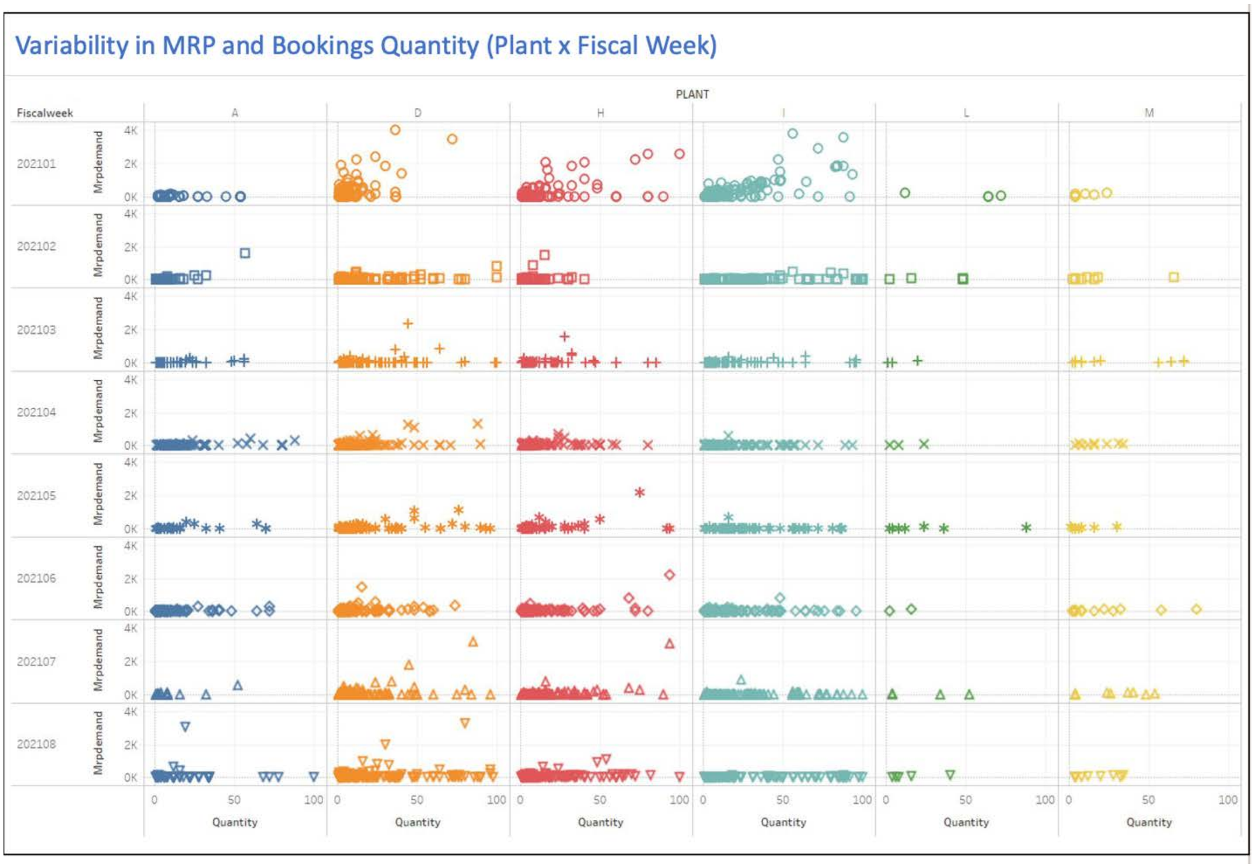


Figure 5

In addition to the fiscal week, figure 5 shows another factor (Plant) which accounts for variability. The next three graphs shed more light on variability by plant.

In figure 6, we changed the scale of the y-axis (MRP demand) from 4,000 to 600, to show a closer look at the variability by plant. From this graph we can see that plant "D" has a more scattered plot even as we move to later fiscal weeks, showing more variability.

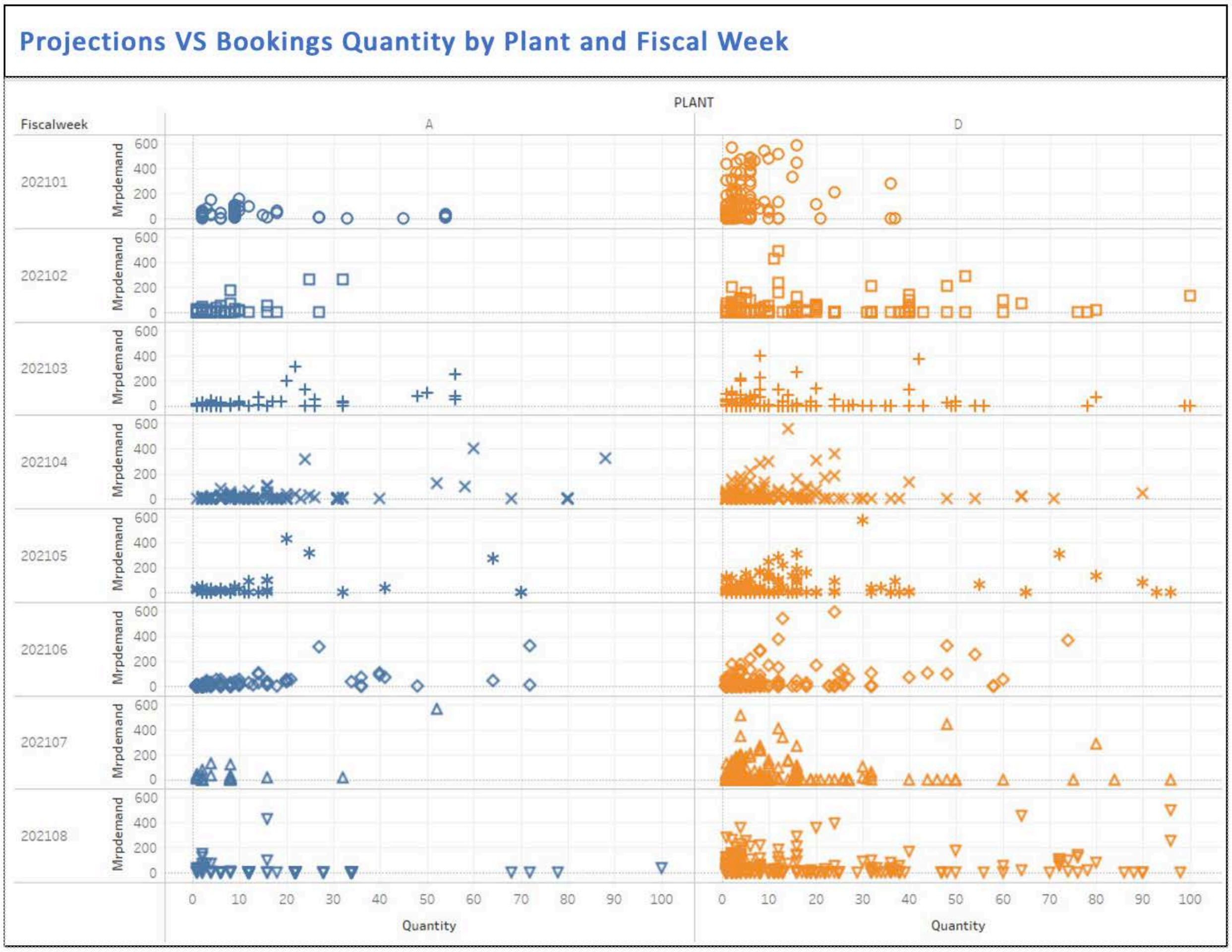


Figure 6

On the other hand, looking at figure 7 below, we see that plant “I”, although having a few outliers flattens out as we progress to later fiscal weeks. While plant H shows larger variances between projected quantity and actual quantity booked, even as we move to later fiscal weeks.

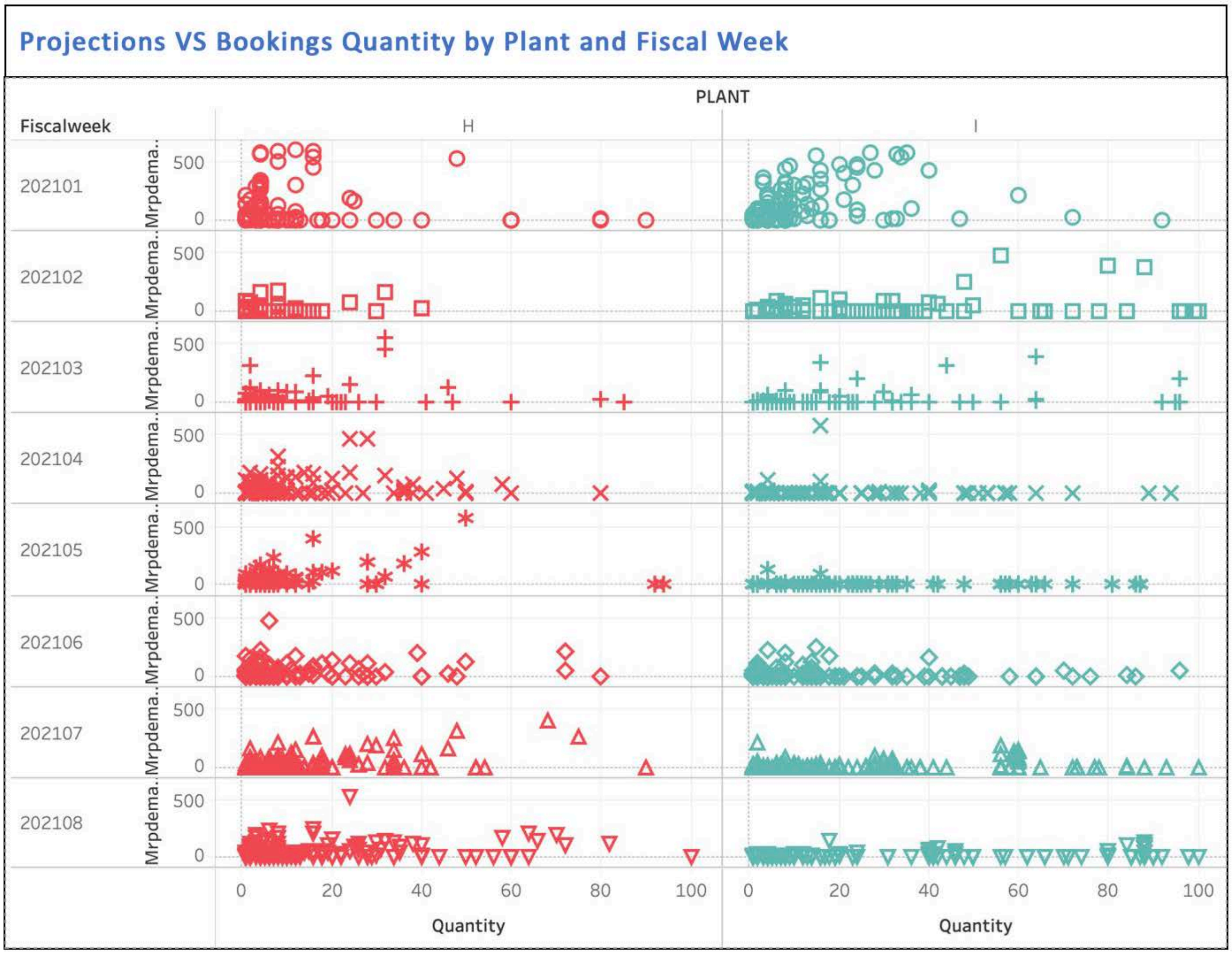


Figure 7

Figure 8 shows us the materials with the largest differences between the projected quantity and the actual quantity booked. Each Box in this graph shows the material number, plant, fiscal week and the amount over predicted. Here the darker the color or bigger the box, the larger the (overprediction) gap.

Note that this graph is filtered to show materials by plant whose predicted (MRP) demand was greater than actual demand or bookings quantity by a thousand or more. Looking at the graph, we can see that material “040-002-652" in plant “D” has the highest variability in fiscal weeks 3-8. An over projected amount could mean overbought inventory and can lead to higher costs for the firm.

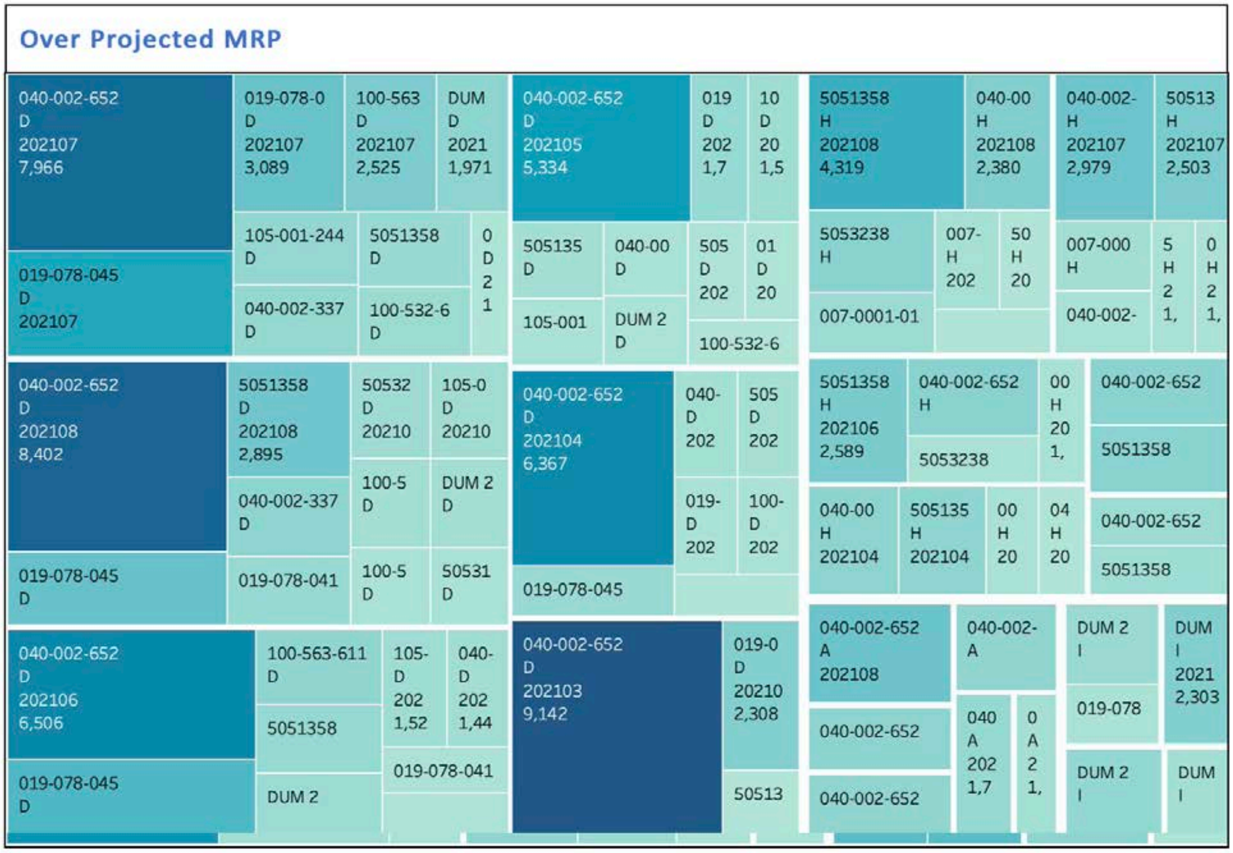


Figure 8

Similarly, figure 9 shows us the under-projected demand. Meaning, actual quantity booked was larger than the predicted quantity. Note that this graph is filtered to show materials by plant whose actual demand or bookings quantity was greater than predicted (MRP) demand by 350 or more. This could potentially mean lost sales, due to insufficient inventory.

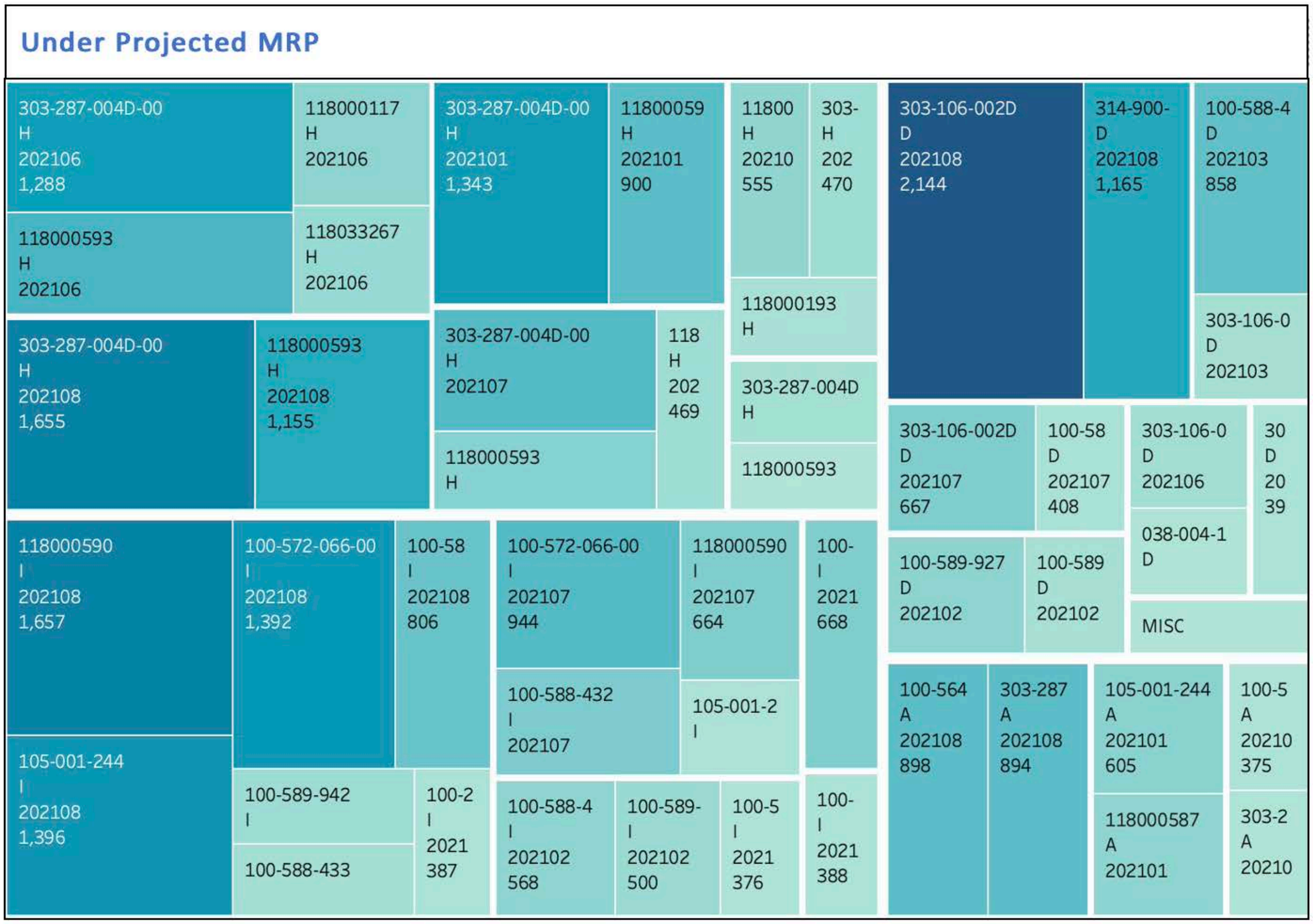


Figure 9

After analyzing where and how much variability exists, we then went on to see if we could identify whether the variables, we were provided with had any impact on how much variability there was and whether or not they could be used to predict quantity booked.

Hence, we decide to use a correlation matrix to see if there were any correlation between the variables and quantity. A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. *(Bock, T. 2018)*

We discovered that there was little to no correlation, which we knew would provide some trouble when creating a model further down the line. We then tried running a VIF function to see if there was any multicollinearity and we also found little to no correlation.

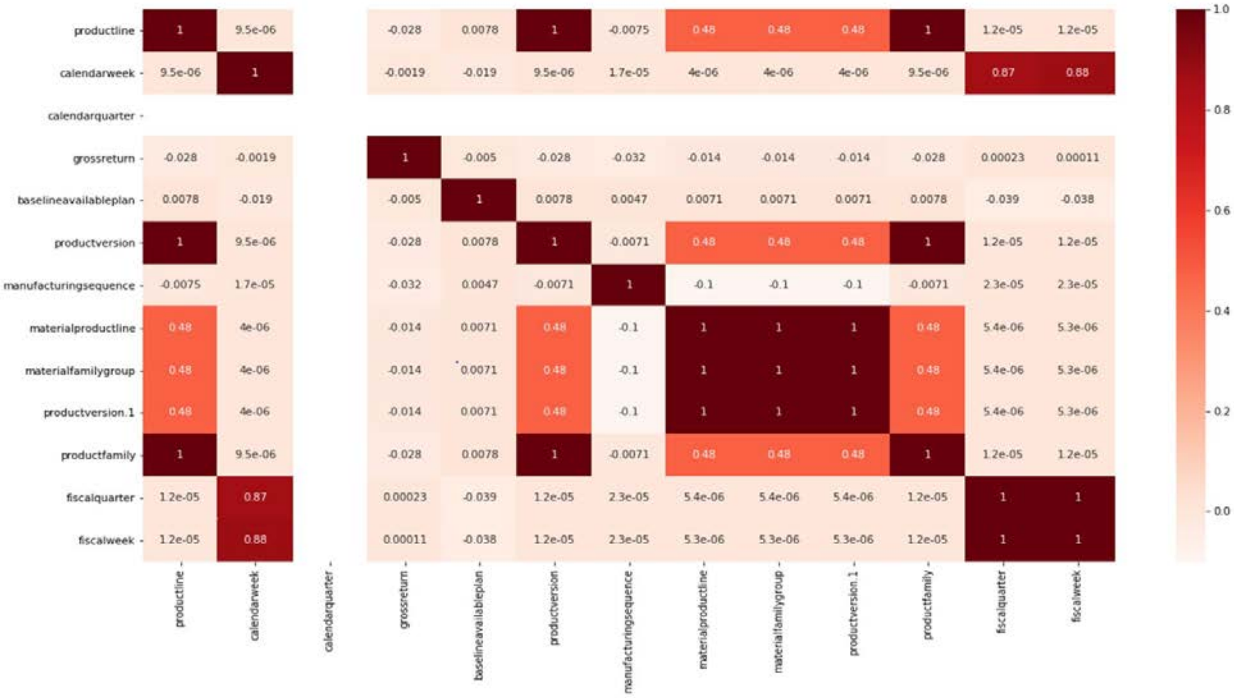


Figure 10

Methodology/Approach

The machine learning technique employed in this model building process is Decision Tree. This technique is used to model the relationship between two or more independent variables and a dependent variable. A decision tree model was selected as it can visually and explicitly capture if there is a match or a mismatch in the quantity booked (actual demand) and MRP demand (predicted demand) and to what extent.

A decision tree is a largely used non-parametric effective machine learning modeling technique for regression and classification problems. To find solutions, a decision tree makes sequential, hierarchical decision about the outcome's variable based on the predictor data. The model acts like a protocol in a series of "if this occurs then this occurs" conditions that produce a specific result from the input data. (*Plapinger, T. 2017)*

Decision trees are incredibly simple to understand due to their visual representation. They require very little data and can handle both qualitative and quantitative data and are quite computationally inexpensive. Decision trees implicitly perform variable screening or feature selection.

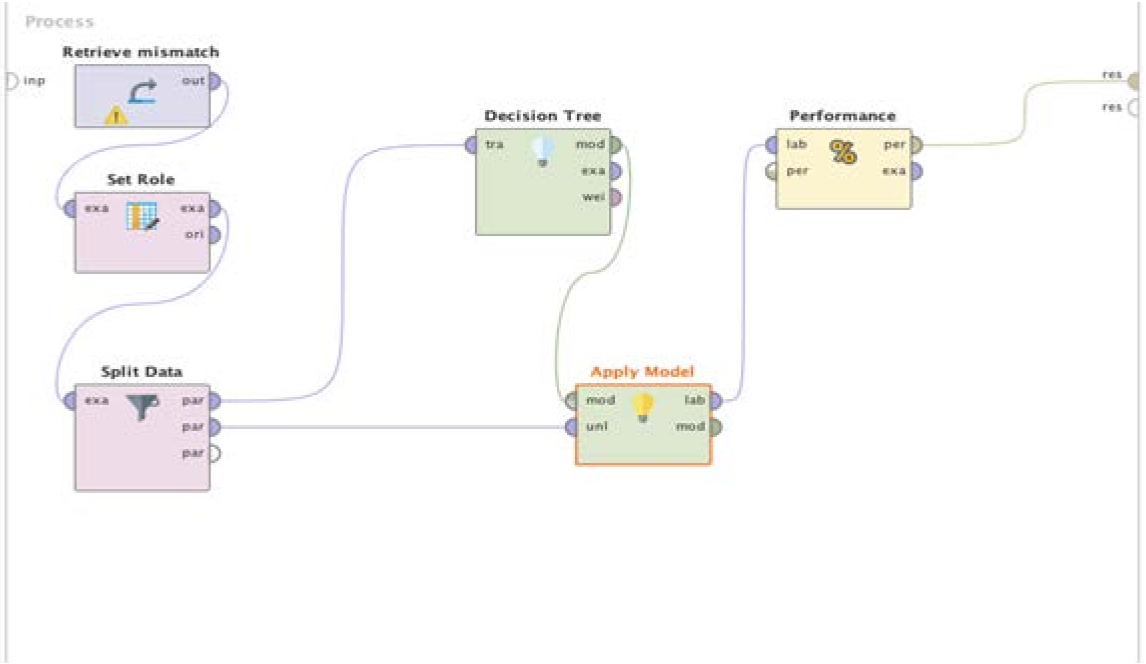


Figure 11: Parameter Tuning

To create the model, we used the merged dataset that had been created earlier in the preprocessing step. We also created a variable called Match or Mismatch to serve as a dependent variable in the model. A classification tree was selected where each node, or leaf, represent class labels while the branches represent conjunctions of features leading to class labels. The Data was partitioned into 70% train set and 30% test set.

A node is 100% impure when a node is split evenly 50/50 and 100% pure when all its data belongs to a single class *(Plapinger, T. 2017)*. To optimize our model, we need to reach maximum purity and avoid impurity. Gini impurity was used as a measure in determining how often a randomly chosen element is labeled incorrectly if it was randomly labeled according to distribution. The goal is to have it reach 0 where it will be minimally impure and maximally pure falling into one category.

Information gain was also used in order to decide what feature to split at each step in the tree. It is calculated; Information Gain = Entropy(parent) - Weighted Sum of Entropy (Children). After running the model a few times, we discovered a high variance at the expense of bias, which meant that the model was overfitting.

In the tree below, the maximum depth was set to 10. Maximum depth shows how many nodes deep the tree will go. In general, the deeper we allow our tree to grow, the more complex our model will become because we will have more splits and will capture more information about the data. This is one of the root causes of overfitting in decision trees because if our model fits perfectly on the training data, it will not generalize well on the test set. So, overfitting can be reduced by reducing the number of maximum depths. *(Mithrakumar, M. 2019)*

The performance of our Decision tree model was further increased by pruning. It involves removing the branches that make use of features having low importance. This way, we reduced the complexity of the tree, and thus increasing its predictive power by reducing overfitting while not creating too much bias in the process.

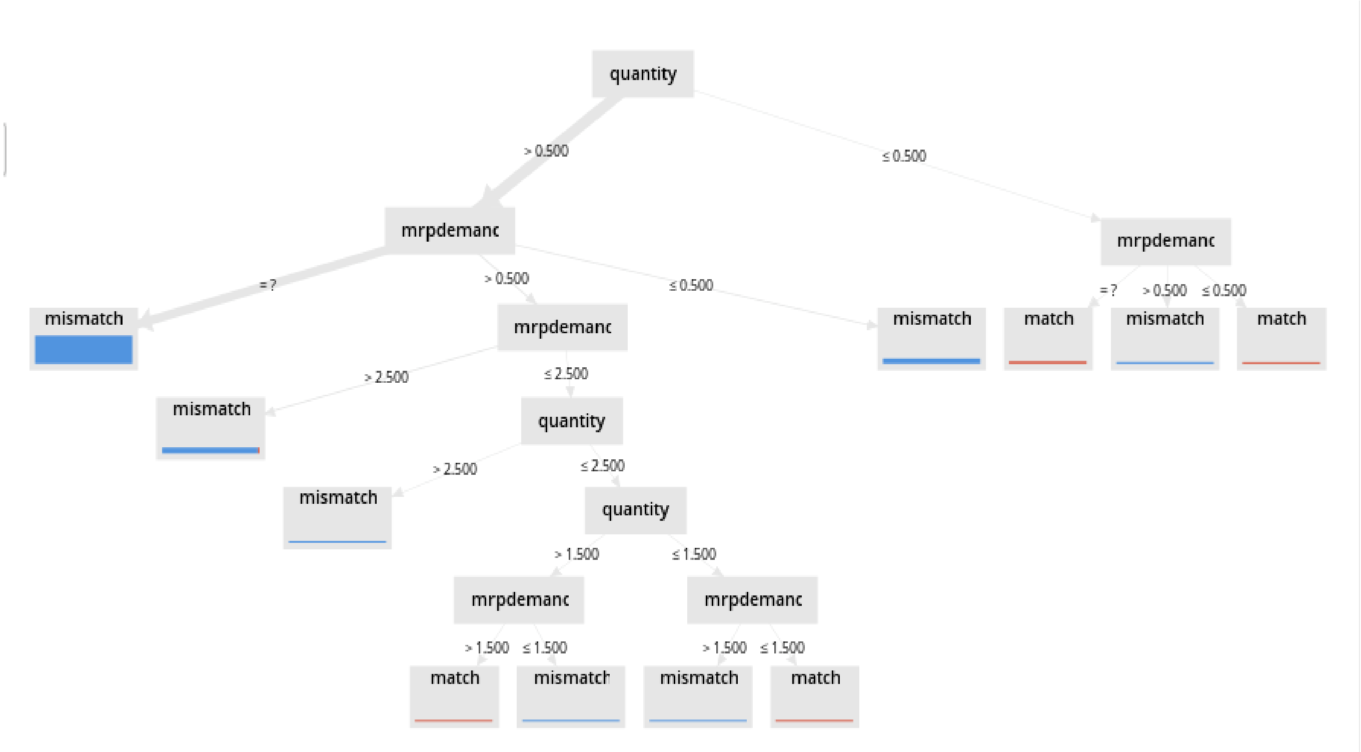
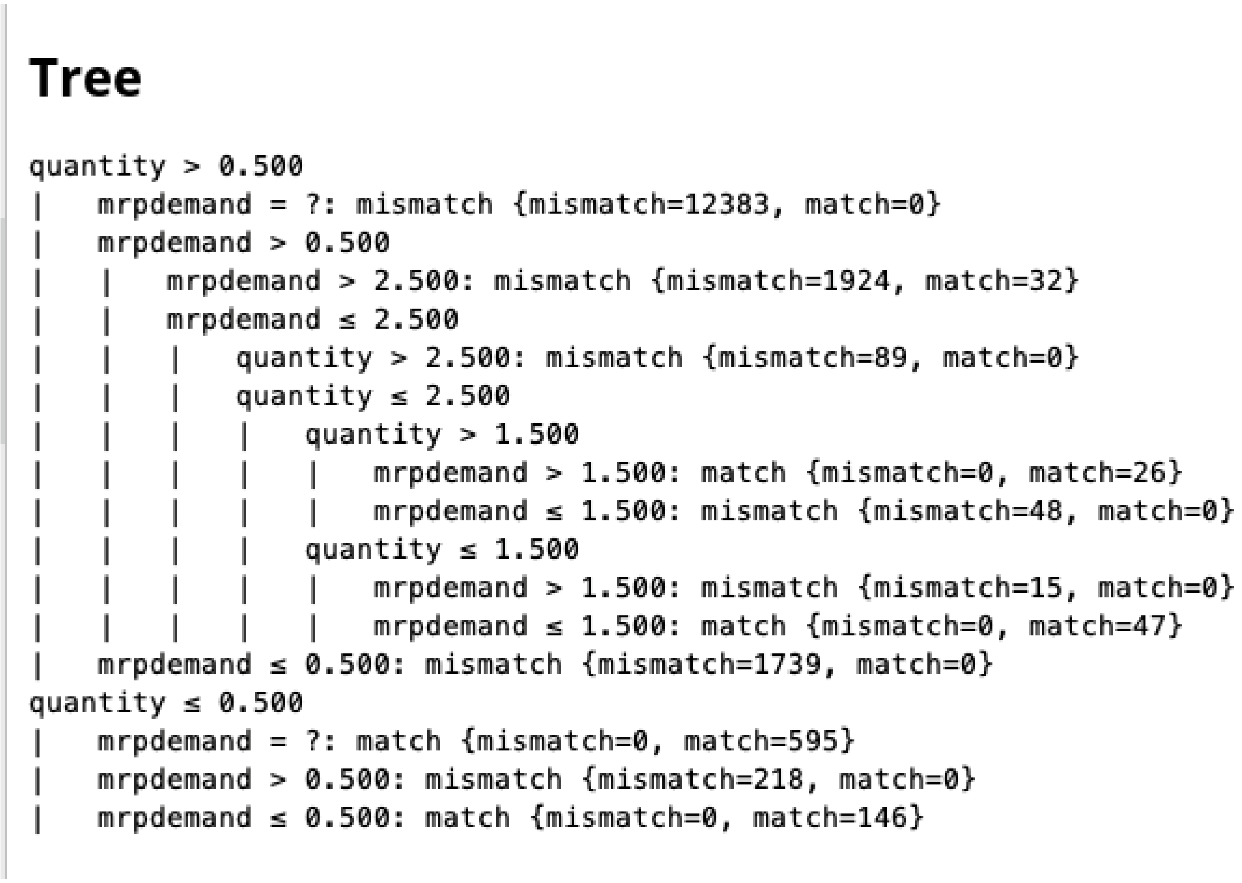


Figure 12: Decision Tree

In this tree, Quantity booked is the predictor variable used for the primary split. The same predictor variable can be used to split many nodes. For example, node 2 is further split using MRP demand. MRP demand is the primary split for node 3. Nodes 2 and 3 were formed by splitting node 1 on the predictor variable Quantity booked. The split point is 0.5. If the splitting variable is continuous (numeric), as in this split, the values going into the left and right child nodes will be shown as values less than or greater than some split point 0.5 in this case. Node 2 consists of all rows with the value of Quantity booked greater than 0.5, whereas node 3 consists of all rows with Quantity booked less than 0.5. The red line represents a match and the blue lines represent a mismatch

The number after the Split Point, for example, for node 2 the first number is 12,383, which indicates the total number of rows in the data that belongs to this node. For node 3 the number is 1924, and for node 4 the number is 89. Quantity booked and MRP Demand are the variables that are actually used to construct the tree. If we look at the decision tree image and at the node descriptions, we can see that splits have occurred on the variables “Quantity booked and MRP Demand”.



Tree

```
quantity > 0.500
|   mrpdemand = ?: mismatch {mismatch=12383, match=0}
|   mrpdemand > 0.500
|   |   mrpdemand > 2.500: mismatch {mismatch=1924, match=32}
|   |   mrpdemand ≤ 2.500
|   |   |   quantity > 2.500: mismatch {mismatch=89, match=0}
|   |   |   quantity ≤ 2.500
|   |   |   |   quantity > 1.500
|   |   |   |   |   mrpdemand > 1.500: match {mismatch=0, match=26}
|   |   |   |   |   mrpdemand ≤ 1.500: mismatch {mismatch=48, match=0}
|   |   |   |   quantity ≤ 1.500
|   |   |   |   |   mrpdemand > 1.500: mismatch {mismatch=15, match=0}
|   |   |   |   |   mrpdemand ≤ 1.500: match {mismatch=0, match=47}
|   mrpdemand ≤ 0.500: mismatch {mismatch=1739, match=0}
quantity ≤ 0.500
|   mrpdemand = ?: match {mismatch=0, match=595}
|   mrpdemand > 0.500: mismatch {mismatch=218, match=0}
|   mrpdemand ≤ 0.500: match {mismatch=0, match=146}
```

Figure 13: Decision Tree

Evaluation

The validation technique used for our model is 10-fold cross validation. Cross-validation is a statistical method used to estimate the skill of machine learning models on unseen data. This means using a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. The cross-validation technique involves; shuffling the dataset randomly, Splitting the dataset into K group, and for each unique group, taking the group as a hold out or test set, taking the remaining groups as a training dataset, fitting a model on the training set and evaluating it on the test set, retaining the evaluation score and discarding the model, finally, summarizing the skill of the model using the sample of model evaluation scores. *(Brownlee J. 2018)*

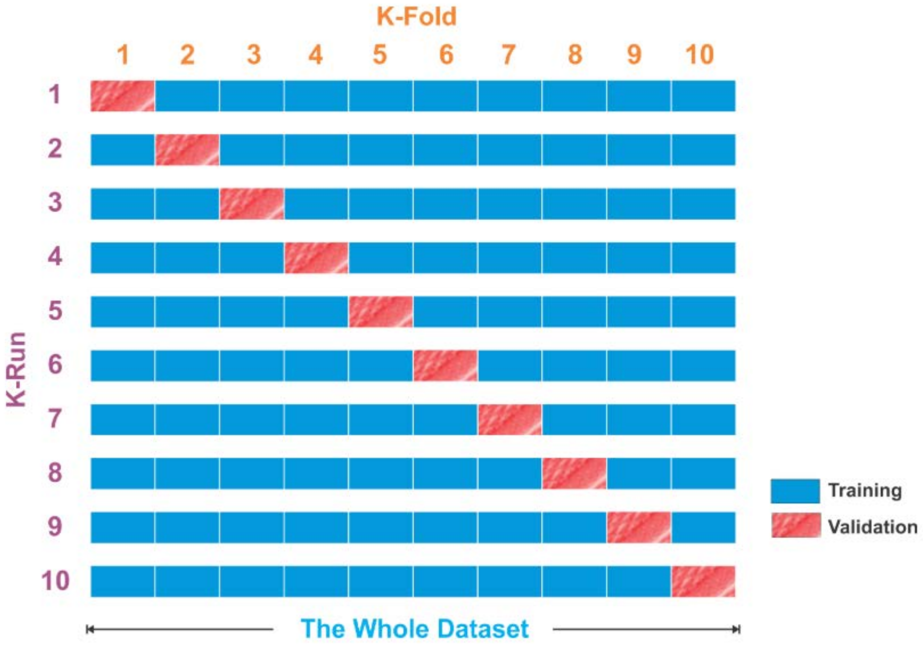


Figure 14: 10-fold Cross Validation

Results

A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It visualizes the performance of an algorithm. The confusion matrix shows the ways in which a classification model is confused when it makes predictions. It gives insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. The terms of a confusion matrix are defined below.

- Positive (P): Observation is positive
- Negative (N): Observation is not positive
- True Positive (TP): Observation is positive and is predicted to be positive.
- False Negative (FN): Observation is positive but is predicted negative.
- True Negative (TN): Observation is negative and is predicted to be negative.
- False Positive (FP): Observation is negative but is predicted positive.

Accuracy:

Accuracy stands for the ratio of correct prediction on the match and mismatch among all observation. The accuracy rate for our model is 99.35%

Accuracy = TP + TN/ (TP+TN+FP+FN)

Recall:

Recall can be defined as the ratio of the total number of correctly classified positive examples divided by the total number of positive examples. In our case, it is the ratio of correct prediction on Match among all Matches. High Recall indicates the class is correctly recognized (a small number of FN). The Recall rate for our model is 86.74%

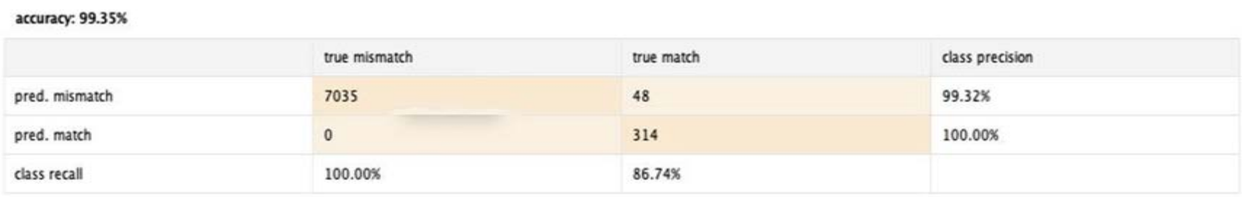
Recall = TP/ (TP + FN)

Precision:

Precision stands for the ratio of correct prediction on Match among all predicted Matches. To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. The Precision rate for our model is 99.32%.

Precision = TP / (TP + FP)

A low recall and a high precision like in our case indicates that we missed a lot of positive examples (high FN) but those we predicted as positive are indeed positive (low FP)



accuracy: 99.35%

	true mismatch	true match	class precision
pred. mismatch	7035	48	99.32%
pred. match	0	314	100.00%
class recall	100.00%	86.74%	

Figure 15: Confusion Matrix

Recommendations

Throughout this process we were able to gather a lot of information based on the datasets that Dell provided. Although we believe that by taking into consideration a few simple steps and recommendations, a lot more information can be discovered.

The first step we recommend is to collect more data on consumer behaviors and economic conditions and compile them into the datasets. Most of the data provided to us, was focused on the intent of analyzing product characteristics and its effect on variability. This data did give us a lot of insight into which types of products were most likely to have a variability, but the data left some questions unanswered on what may have been the reasons for this variability. We believe that with the right collection of new data we could gain a lot more insight into what is causing this variability. For example, consumer and business confidence data could provide us further insight into how market conditions might affect the variability for certain products as these metrics change over time. Overall, Dell could investigate and test many different variables to see which ones may have the biggest impact on the variability.

The second recommendation would be to add in more of a variety of product characteristics into the MRP and Bookings datasets. For example, we would recommend adding in variables that not only show where variability exists, but also show the impact of that variability on the company. For example, one of the variables that could be used to measure this impact of variability is price. Knowing the price of each material would help decision makers to determine how important the materials variability was. This is a variable that would especially be important when it comes to analyzing materials that have large gaps, since some materials can have a gap of 1000 or larger. This size of a gap may matter a lot more when it comes to a server that may cost $10,000, than it would for a headphone jack that may cost 1 cent. Having the price for each product would allow Dell to add weights to each individual material, so that they could better visualize how much of an impact the gaps had. These types of variables would help Dell obtain more in-depth analysis of the data.

The current datasets also did present some big discoveries based on the analysis of the data and provided us with areas that Dell could investigate to further understand what may be causing these gaps directly. As discussed earlier, we had found that some plants continued to have a large amount of variability in the later fiscal weeks, while other plants were able to lower their variability. That is why we recommend that the third step Dell should take is to investigate these plants and figure out what may be causing the high rates of variability. For example, looking into what may be attributing to high variability at plants D and H vs exploring the conditions that exist at plants A and I which have low variability. This would help Dell discover whether these inconsistencies exist because of decisions at the business level or because of decisions made by the consumer.

Finally, after collecting all these data points, Dell would be given the opportunity to create some new and good analysis on why the variability exists. This is because all this data would provide more room to create models that could not only analyze where gaps exists but could potentially provide some insight into what type of factors and variables have the most impact on each products variability. This is all because the more data provided, the more likely it is that different models would have a better chance at finding correlation and have the ability to predict which materials might have larger or smaller gaps in the actual vs predicted sales.

Together all these steps should provide Dell with a footprint for moving forward. These datasets have a lot of potential to create impactful results on the firm’s decisions, and with the right mix of new data, Dell would be well positioned for discovering lot more information. If Dell is to take these next steps, it would not only give them a new set of descriptive analytics that could potentially be insightful to their business decisions, but it could also provide them with new ideas for further research.

References

- Bock, Tim. (2018, August. 16) “What is a Correlation Matrix?”. Retrieved from: https://www.displayr.com/what-is-a-correlation-matrix/
- Factor Analysis | SPSS Annotated Output: Statistical Consulting Group. Retrieved from: https://stats.idre.ucla.edu/spss/output/factor-analysis/
- Handfield, Robert (2018) Supply Chain Resource Cooperative. Retrieved from: https://scm.ncsu.edu/scm-articles/article/dells-supply-chain-data-analytics-center-scdna
- Gupta, Prashant (2017) Decision Trees in Machine Learning. Retrieved from: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
- Plapinger, Thomas (2017) What is a Decision Tree? Retrieved from: https://towardsdatascience.com/what-is-a-decision-tree-22975f00f3e1
- Mithrakumar, Mukesh (2019) How to tune a Decision Tree? Retrieved from: https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680
- Brownlee, Jason (2019) A Gentle Introduction to K-fold Cross Validation. Retrieved from: https://machinelearningmastery.com/k-fold-cross-validation/
- Confusion Matrix in Machine Learning: GeeksforGeeks. Retrieved from: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
- https://corporate.delltechnologies.com/en-us/index.html
- https://en.wikipedia.org/wiki/Dell_EMC