**SUPPLY CHAIN MANAGEMENT**

**MILESTONE 3**

**Goal & Objective: The objective of this exercise is to build a model, using historical data that will determine an optimum weight of the product to be shipped each time from the respective warehouse.**

Milestone 3 - Model preprocessing and Feature engineering stage.

1.Create transformed features in the existing feature set - wherever required and provide reasons for the transformations .

2.Identify new features that can be created from the existing feature set and back that up with business/statistical logic .

3.Drop features that are not required or useful for analysis .

*1.Create transformed features in the existing feature set - wherever required and provide reasons for the transformations* :

***1.Missing values imputation:***

Our dataset contains 24 columns where 8 columns are categorical and 16 are numerical. The target column here is product wt per\_tone and is numerical. We have 3 columns with missing values. They are workers\_numbers, wh\_est\_year , approved\_wh\_govt\_certificate . First 2 are numerical while third one is categorical .

We can see the mean and median of workers\_numbers as well as wh\_est\_year

follows the same. The median of workers\_number is 28 and median of wh\_est\_year is

2009 . So I impute the null values with median .

Now for the approved\_wh\_govt\_certificate column most occurring value is “ C ”. So I impute those missing values with “C ” .

At first there 47.6 % null values in wh estimated year column , 3.9 % null values in workers numbers and 3.6 % null values in approved\_wh\_govt\_certificate column. We imputed those null values with median and mode. Now there is no columns with null values .

**2.Outlier Treatment:**

After the imputation ,box plot is drawn and we can see there are some outliers in it. Here I used IQR method for treating outliers.

IQR method uses the quartiles of the data to determine a range within which the majority of the data points lie. Any data point outside this range is considered an outlier and can be treated or removed as necessary. Q1=25%, Q2=50%, Q3=75% of the distribution. Calculate the first quartile (Q1) and the third quartile (Q3). The first quartile is the median of the lower half of the data, and the third quartile is the median of the upper half of the data.

Calculate the interquartile range (IQR) = Q3 - Q1. Determine the lower and upper bounds for outliers. Typically, any data point below “Q1 - 1.5 \* IQR “or above “Q3 + 1.5 \* IQR” is considered an outliers.

**3.Feature Scaling:**

Feature scaling is a preprocessing step in machine learning that involves transforming input features to a similar scale. In this dataset there are many columns like workers no. which is from 90 and above while electric supply in terms of 0 s & 1 s and product wt is in large values. So before processing this all should be in a common range except the target column. There are types of scaling techniques like minmax scalar, standard scalar, robust scalar etc. Here minmax scaling can be used. We can do this after deriving new significant features and dropping independent features.

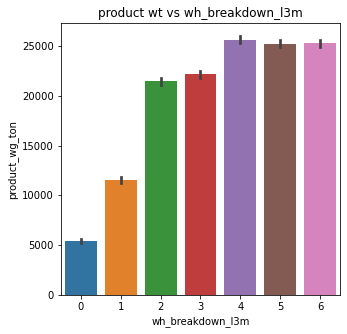
*2. Identify new features that can be created from the existing feature set and back that up with business/statistical logic .*

In the feature engineering process ‘ feature extraction’ is a significant process. It’s the process of finding or deriving new columns from the already excisting ones. Its can be done through combining or dividing present columns or by any other ways . After deriving new columns we should examine its significance with target column through graphically or hypothetically. Implementations of new columns can provide new patterns in that model and through that model can identify the relationship with target column and can perform better accuracy.

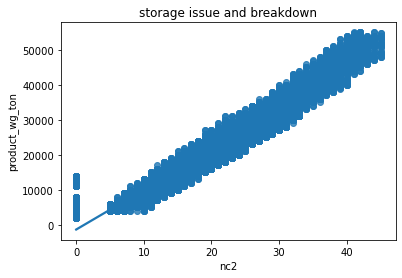
***Feature extraction steps:***

2.Storage issues + warehouse breakdown column(Standard of warehouse):

Both for storage issues reported and warehouse breakdown P\_VALUE = 0.0

****

Storage issues and breakdown of a warehouse indicates the standard of the warehouse. So I try to combine both and created a new column .We got the following regression plot.



**Stats = 0.969546776874581 and p value is0.0 .**

**product weight ton is dependent on storage issue and breakdown.**

Since the derived column shows the p value=0.0 which means those are significant to the product wt per ton and also gives the proper graphs I think we can add that columns to our dataset.

*3.Drop features that are not required or useful for analysis .*

Dropping of unimportant columns is an important step before build the model because abundance of the columns will confuses the model and causes over fitting of the model. So we need to avoid the variance through checking the significance of the columns. Here we already inferred that the columns like 'Ware\_house\_ID' , 'WH\_Manager\_ID' , 'flood\_proof' are independent of the product wt. The first two gives just the ID’s of warehouse and manager while “ Flood Proof “ is a wrong info since it shows the flood effected area as flood proof in many rows. So first we drop these columns totally. Remaining columns have less or strong relationship with our target column. We will consider it through different tests.

*1.Variance Inflation Factor:*

We first checked the VIF tableof included columns :

Variation Inflation Factor assess the multicollinearity (the extent of correlation) between predictor variables in a regression analysis. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, which can cause issues in the analysis.

In general, VIF values greater than 1 indicate some level of multicollinearity, and a commonly used threshold to identify high multicollinearity is a VIF value of 5 or 10. If VIF values exceed this threshold, it suggests that the predictor variables are highly correlated and may need to be addressed in the regression analysis, either by removing some variables, combining them, or using techniques such as principal component analysis to handle multicollinearity.

Here I first treated the the dataset having 1 derived column with Outlier treatment code and with Minmax scalar. Then we got the below table.

FEATURES COLUMN VIF Factor

0 const 1.874059e+07

1 Location\_type NaN

2 WH\_capacity\_size 1.395771e+00

3 zone 1.066843e+00

4 WH\_regional\_zone 1.407454e+00

5 num\_refill\_req\_l3m 1.254118e+00

6 transport\_issue\_l1y 1.083435e+00

7 Competitor\_in\_mkt 1.061840e+00

8 retail\_shop\_num 1.043292e+00

9 wh\_owner\_type 1.075136e+00

10 distributor\_num 1.003183e+00

11 flood\_impacted NaN

12 electric\_supply 1.186702e+00

13 dist\_from\_hub 1.001749e+00

14 workers\_num 1.180381e+00

15 wh\_est\_year 1.044306e+00

16 storage\_issue\_reported\_3m 5.357816e+01

17 temp\_reg\_mach 1.244651e+00

18 approved\_wh\_govt\_certificate 1.084262e+00

19 wh\_breakdown\_3m 1.216370e+00

20 govt\_check\_3m 1.133791e+00

21 storage issue and breakdown 5.348930e+01

Here we can infer that except storage\_issue\_reported and storage issue and breakdown ,all other columns are values less than 5. That 2 columns have values around 5. There is not a high multicollinearity issues.

Nan values can happen bcoz of zero divisional problems. All other columns are free from multicollinearity ,So we can use in the model building.

***2.Mutual Information Gain:***

*Mutual information Gain(MIG)between two random variables is a non-negative value,which measures the dependency between the variables .It is equal to zero if and only if two random variables are independent ,and higher values mean higher dependency.*

1.storage\_issue\_reported\_l3m 1.890083

2.storage issue and breakdown 1.585627

3.wh\_breakdown\_l3m 0.359849

4.wh\_est\_year 0.068363

5.num\_refill\_req\_l3m 0.040762

6.approved\_wh\_govt\_certificate 0.038286

7.temp\_reg\_mach 0.013348

8.electric\_supply 0.012185

9.transport\_issue\_l1y 0.011696

10.retail\_shop\_num 0.011386

const 0.007407

11.WH\_capacity\_size 0.006771

12.wh\_owner\_type 0.001466

13.WH\_regional\_zone 0.000604

14.Location\_type 0.000000

15.distributor\_num 0.000000

16.dist\_from\_hub 0.000000

17.workers\_num 0.000000

18.Competitor\_in\_mkt 0.000000

19.zone 0.000000

20.govt\_check\_l3m 0.000000

21.flood\_impacted 0.000000

It seems upto 10 th column we can accept and after that almost zero importance is showing. Also 14.location type column can accept since from hypothetically and graphically it’s shows significance.

--Thank you--