MEF UNIVERSITY

STEEL PRODUCT CLUSTERING AND FEATURE-BASED PRODUCT PRICE ESTIMATION FOR FLAT SECONDARY MATERIALS

Capstone Project

Meryem Kemerci

İSTANBUL, 2018

MEF UNIVERSITY

STEEL PRODUCT CLUSTERING AND FEATURE-BASED PRODUCT PRICE ESTIMATION FOR FLAT SECONDARY MATERIALS

Capstone Project

Meryem Kemerci

Advisor: Prof. Dr. Özgür Özlük

İSTANBUL, 2018

MEF UNIVERSITY

Name of the project: PRODUCT CLUSTERING AND FEATURE-BASED PRODUCT PRICE ESTIMATION FOR FLAT SECONDARY MATERIALS

Name/Last Name of the Student: Meryem Kemerci
Date of Thesis Defense: 10/08/2018

I hereby state that the graduation project prepared by Meryem Kemerci has been completed under my supervision. I accept this work as a "Graduation Project."

02/07/2018
Prof Dr. Özgür Özlük

I hereby state that I have examined this graduation project by Meryem Kemerci which is accepted by his supervisor. This work is acceptable as a graduation project, and the student is eligible to take the graduation project examination.

02/07/2018

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of Meryem Kemerci and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Prof Dr. Özgür Özlük	……...
2. ……..	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Meryem Kemerci Date Signature

EXECUTIVE SUMMARY

STEEL PRODUCT CLUSTERING AND FEATURE-BASED PRODUCT PRICE ESTIMATION FOR FLAT SECONDARY MATERIALS

Meryem Kemerci

Advisor: Prof Dr. Özgür Özlük

AUGUST 2018, 44 pages

Machine Learning replaces manual and repeatable processes every day, none of the industries can resist these developments. Older systems were rule-based which would bring some level of automation, but all had their limits. One of the goals of Machine Learning is prediction, and it can be used to obtain higher accuracy and better forecasts.

Price predictions are made by hand according to market expectations and countries' conjuncture in the past, but it is changing fast with the developments of Artificial Intelligence tools.

In steel Industry, price levels are determining based on human intuition and simpler statistics. Profits are directly connected to the right pricing for the right time, machine learning algorithms may do the quotation of the steel properly to increase the company profits.

This study aims to classify items as per quality and estimate the price level for the products. Feature selection preprocessing steps are used to enhance the performance and scalability of the classification method. The second part is feature-based product price estimation for the secondary products and selects the predictors of each quality under the product family.

Key Words: Product Clustering, Price Estimation, Linear Regression, K-Means Clustering

ÖZET

STEEL PRODUCT CLUSTERING AND FEATURE-BASED PRODUCT PRICE ESTIMATION FOR FLAT SECONDARY MATERIALS

Meryem Kemerci

Tez Danışmanı: Prof Dr. Özgür Özlük

AĞUSTOS 2018, 44 sayfa

Makine Öğrenimi her gün manuel ve tekrarlanabilir süreçlerin yerini almaya başladı, endüstri kollarının hiçbiri bu gelişmelere karşı koyamaz hale gelmektedir. Eski sistemler, bazı otomasyon seviyelerini getirecek kurallara dayalıydı, ancak hepsinin sınırları vardı. Makine Öğreniminin amaçlarından biri tahminidir ve daha yüksek doğruluk ve daha iyi tahminler elde etmek için kullanılabilir. Fiyat beklentileri geçmişte piyasa beklentileri ve konjonktürel gelişmelere göre elle yapılıyordu, ancak Yapay Zeka araçlarının gelişmeleri ile birlikte bu durum hızla değişmektedir. Sektörler, makinelerin tahminlerinin eski yöntemlerden daha güvenilir olduğunu fark etmeye başladıkça, yapay zeka sistemlerine geçişleri daha hızlı oluyor. Çelik Endüstrisinde fiyat seviyeleri, insan sezgisine ve daha basit istatistiklere dayanarak belirleniyor. Kârlar doğru zaman için doğru fiyatlandırma ile doğrudan bağlantılıdır, makine öğrenimi algoritmaları, şirketin karını arttırmak için çeliğin teklifini doğru bir şekilde yapabilir.

Bu çalışma, kaliteye göre ürün sınıflandırmayı ve ürünler için fiyat seviyesini tahmin etmeyi amaçlamaktadır. Özellik seçimi ön işlem adımları, sınıflandırma yönteminin performansını ve ölçeklenebilirliğini geliştirmek için kullanılır. İkinci kısım, ikincil ürünler için özellik bazlı ürün fiyat tahminidir ve ürün ailesi altında her bir kalitenin tahmin edicilerini seçer.

Anahtar Kelimeler: Ürün Segmentasyonu, Fiyat Tahminleme, Linear Regresyon, K-Means

TABLE OF CONTENTS

.

1. INTRODUCTION

Due to new computing technologies, machine learning is becoming much more critical (Robert D. Hof, 2013) according to the past. The idea was to teach computers without being programmed to apply specific tasks; explorers curiosity was to try if computers can learn from data. (Larry Hardesty, 2013) Increasing the repetition attribute of machine learning is significant because as models exposed to new data, they can adapt independently. They learn from past applications to create reliable, acceptable decisions and results. Machine learning is not modern science but has a long way to go with exciting developments.

Heavy Industries may adopt slower all those developments in the matter of machine learning or technological innovation of each part of the business, but when it returns to automation, the area creates new wants and more flexibility. (Gutierrez and Khaytin, 2016) Most of the manufacturers can make the fresh produced goods pricing easily according to market demand, seasonality or other factors. Steel prices are always affected by oil prices, raw material and scrap prices movements, market demand. Each market has to be evaluated in its own factors. While fresh produced steel products pricing is a huge process with several factors, secondary products pricing is much more complex issue.

If fresh steel products cannot pass quality control (chemically or physically), producers separate that materials and call them “non-prime” or “secondary choice” (Picture 2). It means that you cannot sell them in a market price, you should do the quotation of the product again according to main defect or problem.

Each product creates its own market and demand, therefore no-prime market has a huge potential and profitability for companies if you do the right pricing for the right material. The non-prime market becomes more attractive for the buyers due to price benefits and a possibility to sell or use them as a prime for some cases. Producers must create a new sales channel for the goods that could not pass the quality control.

ArcelorMittal Flat Carbon Europe has an auction system for non-prime/secondary materials through www.steeluser.com. SteelUser offers the chance to use ArcelorMittal's online auction system to see the latest offers of available content and to submit bids. Every week new non-prime material list loading to the steeluser system for weekly auctions. There is a big demand in Europe market for secondary materials. Auction system receives the lists from several mills with their own references and categories. Some mills have their grades

which is specific to this mill and cannot find from other mill grades. There are two reasons for this situation, first is mills can have their own grade code for some of them, secondly there are several types of world steel standards and there are no such thing as equivalent steel standards. Therefore, biggest problem is to make all different grades in basic groups. In other words, clustering of the different source of similar products.

Second problem is to make a proper pricing to increase the company profit.

Picture 1 : How steel coils look like?

Description of Coil

A finished steel product such as sheet or strip which has been wound or coiled after rolling.

Flat products

A flat product can be described rolls produce that with smooth surfaces and ranges of width, varying in thickness. We can define the flat products under two categories, flat products (between 1mm and 10mm in thickness) and plates (between 10mm and 200mm thick and used for large welded pipes, shipbuilding, construction and usage of principal works).

Picture 2 : NoN-Prime coil appearance

1.1.Purpose

The purpose of this study is to cluster the products correctly and find the price predictors to get a correct pricing per product by applying the machine learning algorithms based on past three years data. In other words, the study aims to develop a system that automatically tries to determine the right value of a coil according to the characteristics of each product.

For original products, quotation is easy as per chemical contents and physical particulars with market factors and cyclical effects for each market separately. There is an exact rule for pricing; a price cannot be lower than the production cost and slab prices for steel business. Each additional process increases the price of the product, and all are known by the marketing and sales teams to give the best price for the company and customers.

Secondary materials pricing is not very easy; there is no specific rule or set prices for non-prime materials. There are several factors to consider like main defect of the product, actual grade of the material and physical features (width, thickness and length). Following, category of the product must be determined correctly with all combination of the factors This is the first problem to be solved to proceed with further steps for price estimation and

customer segmentation. When companies have a customer segmentation for a product, they can predict the customer demand and price levels that customers are willing to pay.

2. LITERATURE REVIEW

This study aims to make a product classification and price estimation according to product features. Accomplished marketing in the modern industrialized world cannot be done without separation or segmentation of the current customers and potential customers. Companies need to understand the customer, its heterogenic needs, and desires for products and services (Weinstein 2014b p.7)

Segmentation of the market of a company should not be a marketing function, but it should be a determinant of every corporate function. (Malcolm and Dunbar 2012b p.9)

Market segmentation is one of the critical instrument in marketing as other actors in the market share both their customers and prospects into sub-segments which are categorized by sharing such characteristics that are important for the actor. (Kotler & Armstrong 2010b p. 391) Market segments can be qualified in several ways on the way is to describe the preferences of the target customers; similar preferences, referring to customers that roughly have the same choices. Secondly, there are diffused preferences which mean that the customers vary in their preferences and finally clustered preferences which indicate that the natural market segments emerge from groups of consumers with shared choices (Kotler and Keller, 2009: 249).

Before starting customer segmentation work, two steps to be appropriately done: how many classes do we have for the products and how are the borders of between levels determined. Our target is to figure out what factors drive choices in classifications and what techniques are appropriate in different circumstances. We consider that much can be gained by searching and apply by knowing these factors and their relationships. (van Kampen, Tim J.; Akkerman, Renzo; van Donk, Dirk Pieter, 2012)

The primary target of product classification is to use the similarity of products with regards to different properties to classify products systematically. Krishnan and Ulrich (2001) determined four standpoints within the academic society from which product characteristics are studied: marketing, organizations, engineering design, and operations management. In this paper, we focus on the classification of products from the organizations and activities management perspective.

One of the approaches for product classification is a feature-based summary. In this paper, product features are first identified, then physical characteristics are selected.

Secondly, we will estimate the price as per product features. The pricing models described in the literature (P.T. Fitzroy,J.P. Guiltiman,1976) were not felt to apply to essential commodity markets such as steel and did not adequately recognize the uncertainty of demand. The objective of a simple approach, which systematically made maximum use of the market knowledge at a personal level was considered essential, and the method described here evolved from first principles in discussions with marketing/finance officers.

The translation of qualitative customer knowledge into increased revenue is no longer believed to be a simple process, and the numerical representation of the market situation requires access to one or more of the three sources: market "feel" based on constant contact with customers, the new order position, and published market information. Statistical analysis of past orders/deliveries. Econometric studies and market research of the steel industry.

In this paper, we will work on the product features that are given in data, how they are affecting to the prices.

3. METHODOLOGY

Firstly, we should classify the products correctly and see the clusters. Therefore, the K-Means Clustering method will be used for product classification. After that, PCA applied to check the components and correlation of components with all features for product classification. Linear Regression method is employed for price estimation according to product features. Some feature selection methods are applied such as forward selection to increase the model scores to select strong predictors to the price.

3.1. K-Means Clustering

K-means (MacQueen, 1967) is one of the most straightforward unsupervised learning algorithms that solve the well-known clustering problem. The method performs an essential and primitive way to make a classification a given data set through an absolute number of clusters (assume k clusters) fixed a priori. The basic idea is to describe k centroids, one for each group. These centroids should be placed cunningly because of different location causes a different result. The best way is to make the clusters as much as far away from each other.

The following action should be to take each point belonging to the data and associate it to the nearest centroid. When we have no pending point, the first step completed, and the first grouping is done. Then we should re-calculate new centroids as barycenter's of the clusters results from the previous level. When we have known centroids, same steps must continue between the same data points and the nearest centroids. This process will continue until we have no centroids move anymore.

3.2. Agglomerative Hierarchical Clustering

Agglomerative Hierarchical clustering algorithm effectively used for grouping the data one by one based on the nearest space measure of all the pairwise distance between the data points. Hierarchical Clustering is a way to cluster the data with dendrogram to represent data to each group is connected two or more successor groups. All these groups are conglomerated and organized as a tree that expected to end till a meaningful classification diagram (Benjamin C. M. Fung, Ke Wang, and Martin Ester, Simon Fraser,2014).

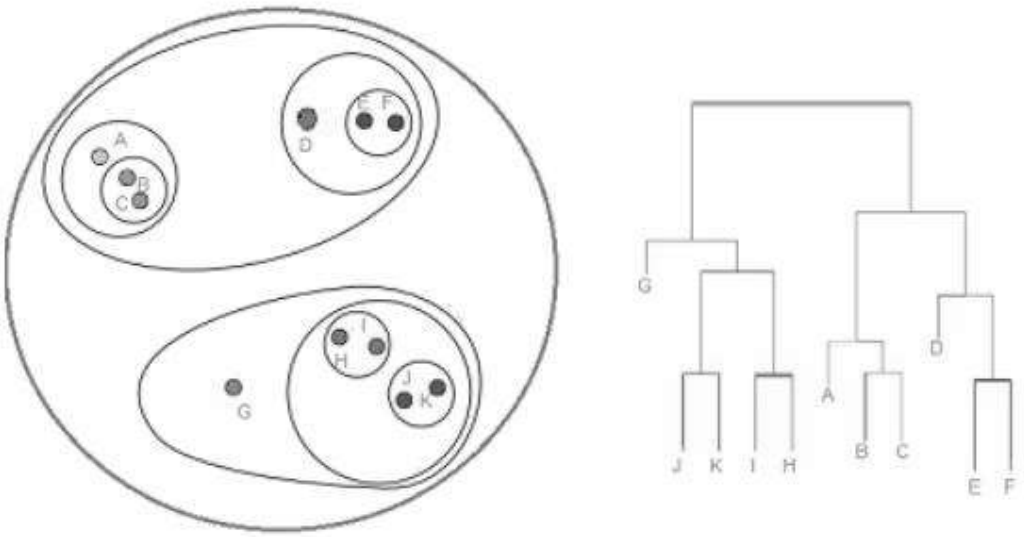


Figure 1: What is the Dendrogram?

3.3. Principal Component Analysis

Big datasets are increasingly common and are often hazardous to exposition. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, boosting interpretability but at the same time reducing the information loss. It does so by building new uncorrelated variables that successively maximize variety. Creating such

new variables, the principal components, decreases to solving an eigenvalue/eigenvector problem, and the new variables are determined by the dataset at hand, not a priori, hence making PCA an adaptive data analysis technique. (Ian T. Jolliffe and Jorge Cadima)

In other words, PCA is a dimensionality reduction technic to reduce the dimensions that can capture the most variation of the data. When we check the PCA results, we can decide how many components you will select by PCA or how my new components will explain the information that the original data.

3.4. Linear Regression

Linear regression is an essential and commonly used type of predictive analysis (Statistics Solutions, 2013). The overall idea of regression is to examine two things: (1) does a set of predictor variables do an excellent job in predicting an outcome (dependent) variable? (2) Which factors, in particular, are significant predictors of the outcome variable, and in what way do they–indicated by the importance and sign of the beta estimates–impact the outcome variable? To explain the relationship between one dependent variable and one or more independent variables, a regression method is used. The formula y defines the purest form of the regression equation with one dependent and one independent variable = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable; the technique calculates the ties by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line.

3.5. Forward Selection

The most straightforward data-driven model building approach is called forward selection (B. Rollin, 2014). In this approach, one adds variables to the model one at a time. At each step, each variable that is not already in the model is tested for inclusion in the model. The most significant of these variables are added to the model, so long as it's P-value is below some pre-set level. It is customary to set this value above the conventional .05 level at say .10 or .15, because of the exploratory nature of this method.

Thus, we begin with a model including the variable that is most significant in the initial analysis and continues adding variables until none of the remaining variables are "significant" when added to the model. Note that this multiple use of hypothesis testing means that the real type I error rate for a variable (i.e., the chance of including it in the model given it isn't essential), does not equivalent the critical level we select. In fact, because of the complication that arises from the complex nature of the procedure, it is virtually impossible to control error rates, and this procedure must be viewed as exploratory.

4. DATASET

4.1. Features and explanation of features

Data is including two separate database datasets merge, one is bidsa, and the other is itemsa both data merged with YWB column which is showing YearWeekBundle details and each row is showing specifically the coil identification.

Itemsa data is mostly including product features such as chemicals (Aluminum, Bore, Carbon, Sulphur, Vanadium, Titanium..etc) and grade (DD11, S355, DD13..etc), product family(Hot Dip Galvanized Steel Coils, Hot Rolled Coils..etc), bundle details, and produced mill country, city, and central defect. Bidsa data content is bidsa auction process details such as auction name, coil is awarded or not, number of offers, Business Division, CMO, product category and class, comments for each row according to status of the coil, customer bid per item and Price Guideline price, customer details as name and country, Market , product family and subfamily, delivery term, product thickness, width and length, weight of the coils.

Price Guideline prices are based on market senses, human perception and basic statistics and it is updated every two weeks.

4.2. EDA Analysis

In our dataset we have 3 group of features, first is about item details such as product family, chemicals, width, thickness, category, grade and produced mill. The second group is about customer details who have the bid for the items, bid price, country, sold to part and the last group is about coil awarded or not and in which price level, auction number, any restriction to confirm the sales of the coil …etc.

We will see some visualizations about related features for our work.

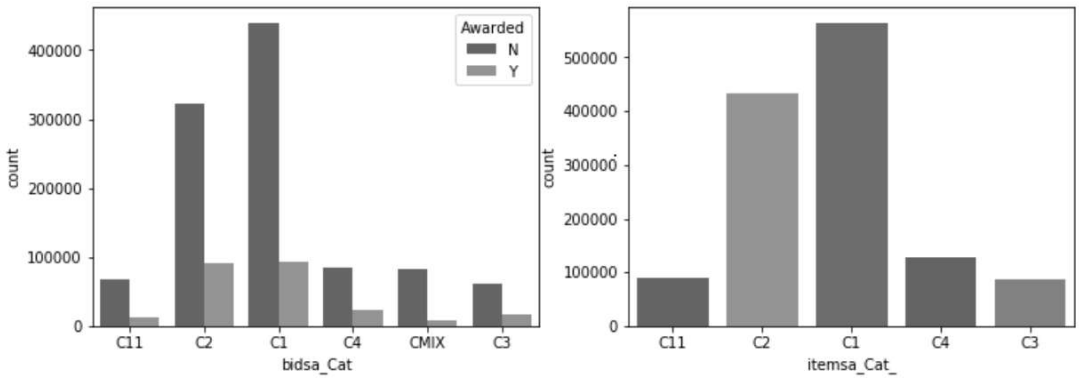


Figure 2: Awarded Coils Category and Items Categories / Left figure (bidsa category from C1 to C11, hue is awarded or not), Right figure (items category from C1 to C11)

We can see clearly in the figure 2 that C1 and C2 is the most significant part of the coils and their sales rate is the highest of all categories.

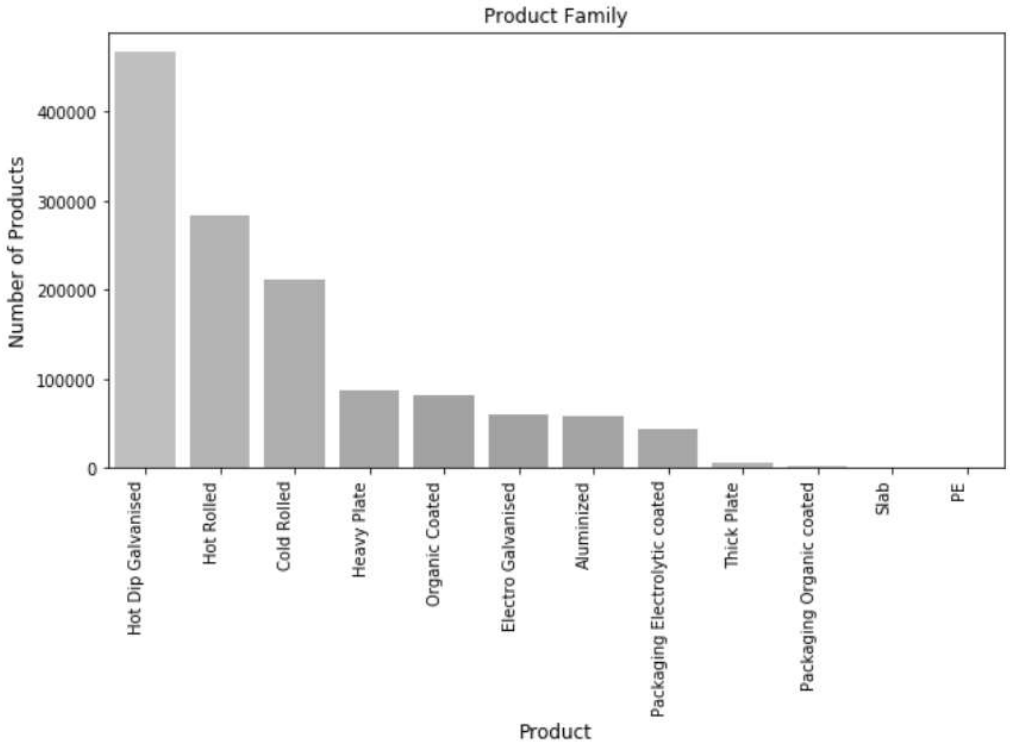


Figure 3: Product Family Breakdown

There are 12 different type of products we have in the data; Hot Dip Galvanized is the first, Hot Rolled is following HDG products. Hot Dip Galvanized products have more physical process than Hot Rolled Coils and we do not have all necessary data in our dataset. Due to incomplete features for HDG we will concentrate on Hot-Rolled material in our work.

4.3. Pre-Processing (Cleaning of Data) and Feature Selection

As we will start with product classification, we need to check product characteristics, and which contents determine the quality of the product. Chemicals and physical features will be essential to classify the products; we will start first with chemicals. Below the table is showing the Chemical values that we have in the data set and the importance of each to decide how to clean NaN or use for classification.

We know that missing values are important as they indicate how much we don't know about our data. Making inferences based on just a few cases is often unwise. Besides, many modeling procedures break down when missing values are involved, and the corresponding rows will either have to be removed entirely or fill with the statistical values based on the feature.

Chemicals	Important for Quality	Alloy Element	Fill or delete NaN?
Carbone	yes	no	delete
Bore	yes	yes	delete
Molybdenum	yes	yes	fill NaN by 0
Chromium	yes	yes	fill NaN by 0
Manganese	yes	no	delete
Silicium	yes	no	delete
Nobium	yes	yes	fill NaN by 0
Titanium	yes	yes	fill NaN by 0
Vanadium	yes	yes	fill NaN by 0
Nitrogen	no	no	fill NaN by 0
Cuivre	no	no	fill NaN by 0
Phosphorus	no	no	fill NaN by 0
Sulfur	no	no	fill NaN by 0
Aluminum	no	no	fill NaN by 0

Table 1: Chemicals essential features

Table 1 is created based on Technical advisor guidance to be able to understand the each chemical value importance and how to use in our analysis.

After cleaning and filling the NaN values (according to Table 1) in chemicals, we need to check if each Grade chemical values are in similar rates. To be able to perform this, all substances are grouped according to the mean and standard deviation to decide if there are outliers or different values in each chemical feature. To be more specific, if we select one grade, we expect to have similar chemical values in all group. Therefore, we decided to

use median value of each chemical. Otherwise, our results may be affected by outliers or wrong entries.

We will concentrate Hot Rolled Steel Coils product, when we filter this product we found 274.618 rows and 120 columns. When we look at the product specification columns; as Grade, we have 792 different Grades and 6 Categories. Category dispersion is as below for Hot Rolled Coils,

Category	Count	Percentage
C1	141047	0,51
C2	45177	0,17
C3	10210	0,04
C4	2987	0,01
CMIX	33946	0,12
C11	41251	0,15

We have six categories given by the mill according to physical conditions for non-prime products. Biggest portion is Category 1 and the lowest is Category 4.

As we explained above that same grades chemical compositions should be similar and better to use median values to prevent outliers, median value of the chemicals for each grade under Hot Rolled Coils are selected to make clustering according to chemicals. For instance; Table 2 in below, first line is showing Grade 0153/C35, we have grouped all 0153/C35 grades according to median values of chemical compositions. It means that Carbone median value is 0,36 and we will use this value for our further analysis.

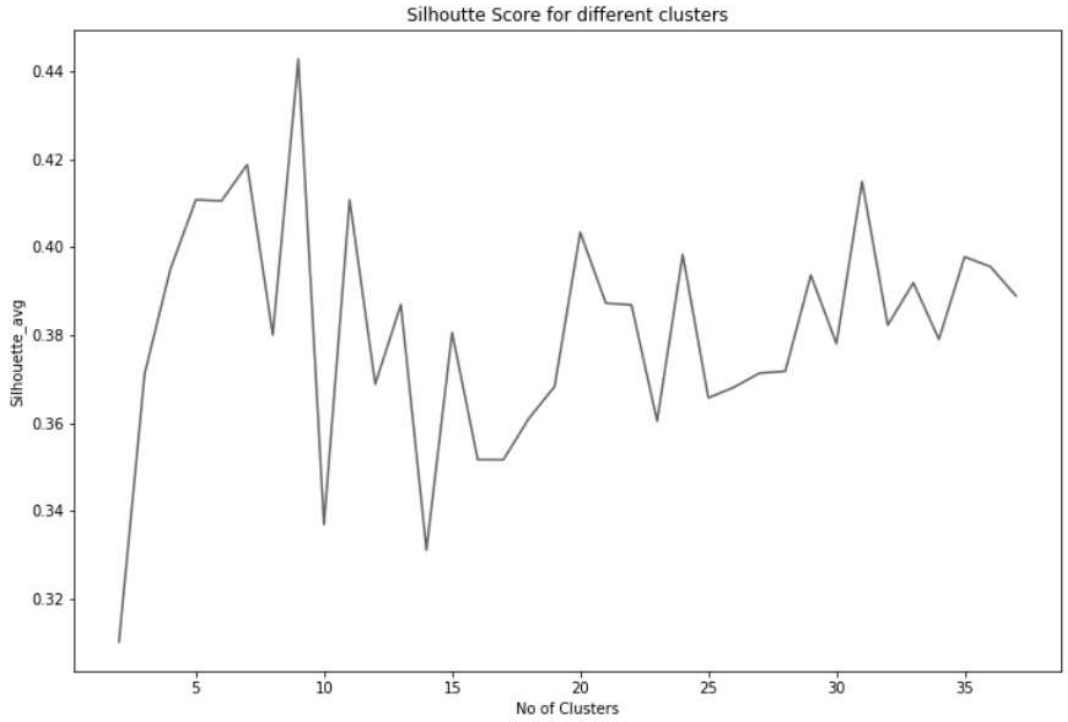
Table 2 : Median Values for each Grade

itemsa_Grade	Aluminium median	Bore median	Carbone median	Chromium median	Cuivre median	Manganese median	Molybdenum median	Nobium median	Phosphorus median	Silicium median	Sulfur median	Titanium median	Vanadium median
0153/C35	0.0180	0.0000	0.3643	0.2865	0.0105	0.7047	0.0019	0.0004	0.0134	0.2426	0.0008	0.0012	0.0010
0162/S600MC	0.0352	0.0001	0.0679	0.0273	0.0140	1.5945	0.0036	0.0511	0.0175	0.0265	0.0022	0.0735	0.0038
01621/S600MC	0.0365	0.0002	0.0705	0.0275	0.0183	1.6114	0.0028	0.0612	0.0185	0.2150	0.0015	0.0741	0.0032
0241/DR14	0.0365	0.0030	0.0352	0.0191	0.0114	0.1903	0.0013	0.0003	0.0132	0.0028	0.0075	0.0002	0.0002
0352/22MNB5	0.0242	0.0001	0.2243	0.0280	0.0165	1.5133	0.0027	0.0017	0.0137	0.3043	0.0013	0.0012	0.1023

5. PRODUCT CLASSIFICATION

5.1. K-Means Clustering

We will use k-means clustering method to classify the goods, due to the availability of the product features in hand, we must start from chemical values to see the determinants of the grade. Biochemical values are scaled between 0 and 1 to prevent outliers effect to the clusters; first results are received as below for 38 groups,



Graph 1: K-Means Clustering for k=38

The silhouette plot shows the that the silhouette coefficient was highest when k = 9, suggesting that's the optimal number of clusters. Above plot shows us nine groups best captures the segmentation of this data set. New clusters are added to the data to check the results with an expert if we have meaningful clusters that is showing different grades.

When we show cluster results to the Metallurgical Engineer about the product clusters, he advised that clustering is not done according to Grades or Carbon. According to him, we should have a new feature or way to catch the Carbon related groups as Carbon is the main determinative to split the quality.

We decided to calculate Carbon Equivalent as a new feature to have a better result.

Carbon Equivalent (CE) is an empirical rate in weight percent, relating the combined effects of different alloying elements used in the creating of carbon steels to an equal amount of carbon. A mathematical equation can be used to calculate carbon equivalent. By varying the amount of carbon and other alloying elements in the steel, the desired strength levels can be obtained by appropriate heat treatment. Better weldability and low-temperature notch toughness can also be obtained. Concerning welding, the Carbon Equivalent governs the hardenability of the parent metal. It is a rating of weldability related to carbon, manganese, chromium, molybdenum, vanadium, nickel and copper content. There are many commonly used equations to calculate the Carbon Equivalent. One example of such a mathematical formula is:

CE = C + Mn/6 + (Cr + Mo + V)/5 + (Ni + Cu)/15

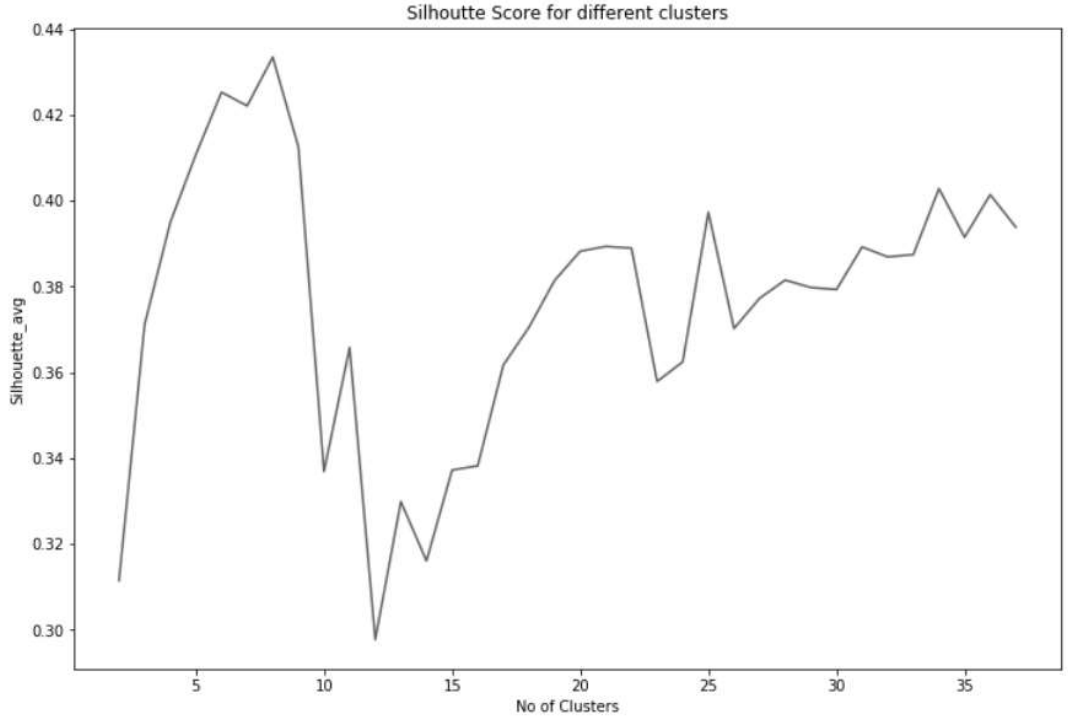
After creation of the new feature, we could make a new clustering by K-means to check if the clusters diverted than the first clusters.

An additional column is created as per above formula; chemical values are scaled between 0 to 1 then values are multiplied by 10 to increase the weight of the column. When we create the carbon clusters, again we showed the results to expert if clustering is significant or not. He advised that CE clusters are closer to the grade groups but better to use below grouping and check the differences between both groups.

Group 1	<= 0,005
Group 2	> 0,005 and <= 0,2
Group 3	> 0,2 and <= 0,33
Group 4	> 0,33 and <= 0,45
Group 5	> 0,45 and <= 0,65
Group 6	> 0,65

K-means clusters and Label Carbon clusters cross table in below is showing that we do not have completely parallel groups,

cluster	0	1	2	3	4	5	6
label_carbon							
1	10	0	0	0	0	0	0
2	295	0	0	73	23	0	0
3	41	28	0	66	2	19	0
4	1	76	1	2	0	16	0
5	0	72	23	1	0	1	5
6	0	2	29	0	0	6	0

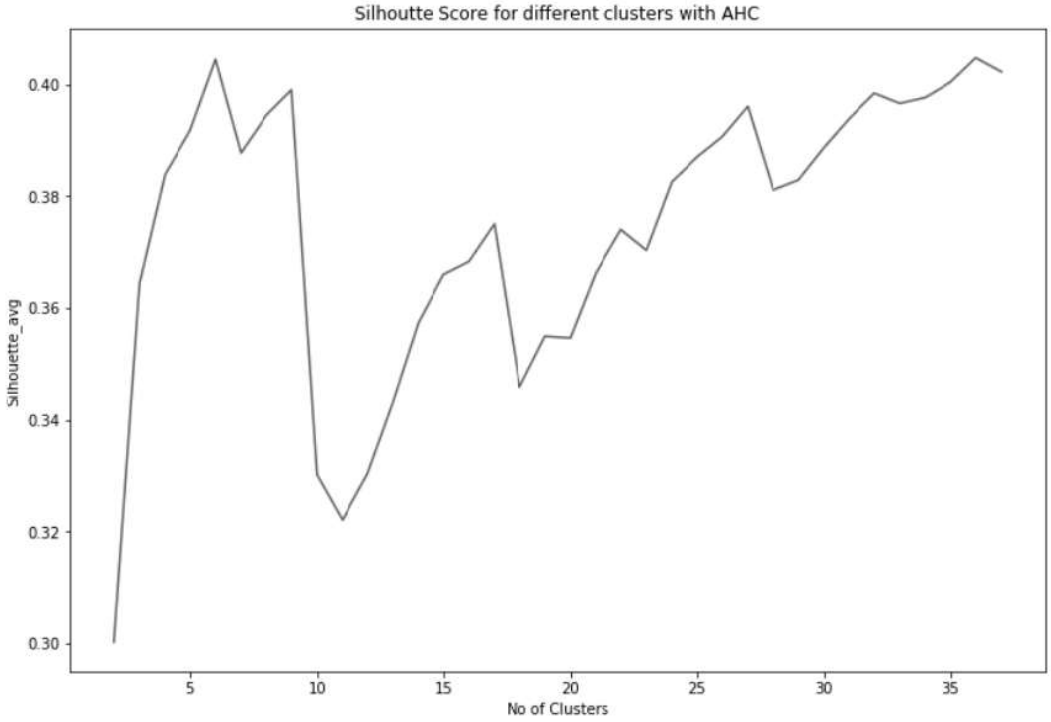


Graph 2: K-Means Silhouette Plot for k=38, CE value added

Graph 2 is showing us that silhouette coefficient is highest when k = 8, looks that it's the optimal number of clusters. As a result, we added Carbon Equivalent feature and applied K-means with a new feature. Separately, new grouping is done according to expert advises and saved label_carbon. Both clusters are compared to each other and showed to expert, he found label_carbon clusters closer to the steel grades.

5.2. Agglomerative Hierarchical Clustering

After k-means clustering, it is better to cross check the results by another clustering method to see if we will embrace any differences, we tried Agglomerative Hierarchical Clustering.



Graph 3 : Agglomerative Hierarchical Clustering for k = 38

Plot result is showing us that the silhouette coefficient is highest when k = 9 which is similar to k-means results for the optimal number of clusters.

To sum up, we found that K-means outcomes with a raw data and with Carbon Equivalent feature are different, we will check PCA plots to understand both results components.

5.3. PCA

PCA will help us to understand which components are explaining more to our model and comparing the K-means, grouping by CE value and Agglomerative Clustering similarities with each other.

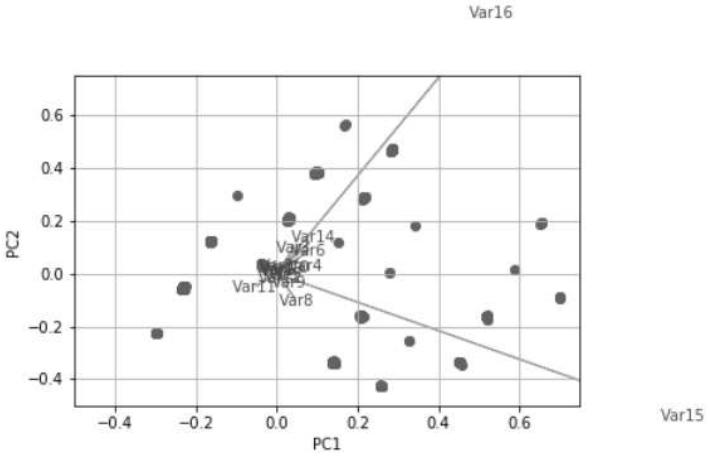


Figure 4: PCA for two components

```
pca=PCA(2)
pca.fit(X)
pca.explained_variance_ratio_
```

```
array([ 0.63614261,  0.31983312])
```

Two components explained ratio is 0,95 we can be sure that two components are enough to justify the variance.

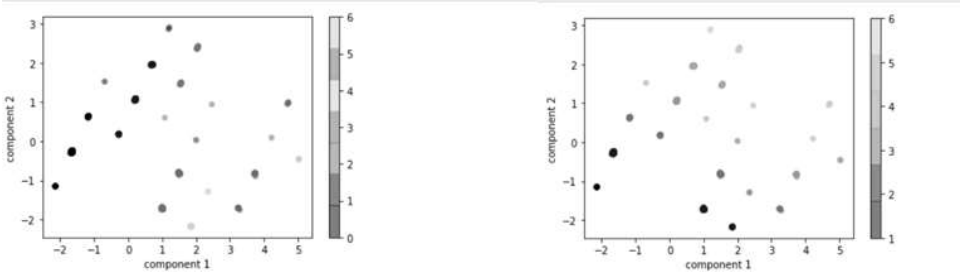


Figure 5: PCA for created clusters by k-means with raw data (left), Carbon Equivalent PCA (right)

We can commentate Figure 5 that k-means clustering clusters and CE groups have different components while left figure color directions are from left to right, other figure color directions are from right to left. Their components are different than each other.

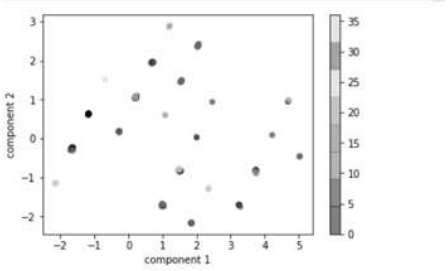


Figure 6: PCA for Agglomerative Hierarchical Clustering

Agglomerative clustering PCA plot has a similarity with K-means clustering, color directions are same.

We found that K-means clustering, and Agglomerative Clustering results show similarity while Carbon Equivalent groups are showing the different pattern. Both results can be used for future analysis for customer segmentation and see which cluster the best is to use for customer segmentation.

6. LINEAR REGRESSION FOR PRICE ESTIMATION

As a second step, Linear Regression model will be used for price estimation to find the powerful predictors within product chemicals, physical features and categories. We will try to find an answer to “Which features are more important for price estimation?”.

Linear Regression will be applied to all Hot Rolled Coils data with all 792 grades to see the regression results and continue our model with the specific qualities to compare the price estimators and results.

Quotation for steel, there are two metrics; chemical contents and physical features. We used 13 chemicals, and one column created according to chemical value weights that is called "Carbon Equivalent," as physical features; we selected thickness, width, and category of the coils from our data for the regression. There are six different categories like C1, C2, C3, C4, CMIX, and C11. All categories are converted to dummy variables as 0 and 1 and concatenated data table for better regression results. All variables are selected as train except PGL price, PGL is our target to predict.

We used two ways for NaN values, first is “drop” them, second is “fill with median”. First, we have dropped NaN values and found the regression result 0,19, it is very low. When we have low linear regression results, it is obvious that some features are affecting the result negatively and we should find the most important features. There are several methods for feature selection, we will use forward selection to find best features. After we applied forward selection adjusted R-square result is increased to 0,75 for 11 predictors which is acceptable. These predictors are Width, Sulfur, CMIX, C11, C1, Titanium, Molybdenum, C2, Silicium, C3, and Carbon Equivalent features. When we add one more predictor to the model, our result decreased to 0,18. The additional predictor is C4 that is strange that our score dropped dramatically. [1]

We will go deeper and select the most repeated grades that are DD11 and S355. DD11 is the first on the list with 66,968 rows than S355 is following with 39,576 rows. Linear regression will be applied for both grades; first, we will go with DD11 analysis.

[1] Appendix V – HRC NaN’s are dropped regression results

6.1. DD11

Before we start with linear regression modeling, DD11 data checked if we have any NaN rows and how to deal with these items. As we mentioned above two ways are tried, first is dropping the NaN values, the second way is to fill the Nan's with median value.[2]

Regression Score found 0,13 which is very low. This result shows us that some features should not be in the model to increase the model score. Forward selection applied, the adjusted R-squared result is increased severely for six predictors to 0,865 (Table 3), and selected features are the width, C1, C11, C2, C3, CMIX. When we add one more predictor, the adjusted R-squared result decreased to 0,11(Table 4) with C4. It is evident that C4 is affecting the outcome negatively and it is one of the predictors that is reduced to R square score.

OLS Regression Results

Dep. Variable:	bidsa_PGL_Price	R-squared:	0.865
Model:	OLS	Adj. R-squared:	0.865
Method:	Least Squares	F-statistic:	4.834e+04
Date:	Thu, 23 Aug 2018	Prob (F-statistic):	0.00
Time:	23:12:30	Log-Likelihood:	21405.
No. Observations:	45377	AIC:	-4.280e+04
Df Residuals:	45371	BIC:	-4.274e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
itemsa_Width__mm_	5.566e-05	2.38e-06	23.376	0.000	5.1e-05	6.03e-05
['cat']_C1	0.3435	0.003	112.101	0.000	0.337	0.349
['cat']_C11	0.3101	0.003	101.865	0.000	0.304	0.316
['cat']_C2	0.2653	0.004	75.165	0.000	0.258	0.272
['cat']_C3	0.2140	0.006	35.584	0.000	0.202	0.226
['cat']_CMIX	0.1339	0.004	31.724	0.000	0.126	0.142

Table 3 : Linear regression Result for 6 predictors

[2] Appendix I - DD11 NaN's are dropped, box plot of categories with the regression features

OLS Regression Results

Dep. Variable:	bidsa_PGL_Price	R-squared:	0.119
Model:	OLS	Adj. R-squared:	0.119
Method:	Least Squares	F-statistic:	1020.
Date:	Thu, 23 Aug 2018	Prob (F-statistic):	0.00
Time:	23:12:30	Log-Likelihood:	21730.
No. Observations:	45377	AIC:	-4.345e+04
Df Residuals:	45370	BIC:	-4.338e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
itemsa_Width__mm_	2.935e-05	2.58e-06	11.384	0.000	2.43e-05	3.44e-05
['cat']_C1	0.3757	0.003	114.106	0.000	0.369	0.382
['cat']_C11	0.3397	0.003	104.965	0.000	0.333	0.346
['cat']_C2	0.2964	0.004	79.910	0.000	0.289	0.304
['cat']_C3	0.2446	0.006	40.161	0.000	0.233	0.256
['cat']_CMIX	0.1649	0.004	37.799	0.000	0.156	0.173
['cat']_C4	0.1851	0.007	25.589	0.000	0.171	0.199

Table 4 : Linear regression Result for 7 predictors

NaN values are filled by the median of each feature (it is applied only for chemicals) and regression score found 0.14 that is very low.[3] Forward selection is applied like previous to see the change in the adjusted R-squared results. Score for nine predictors is 0,963 with Width, Aluminum, Sulfur, Manganese, C1, C11, C2, C3 and CMIX.[4]

If we add one more predictor to see the change in the result, we got 0,12 adjusted R-squared result. The 10th predictor is affected negatively by our effect; when we check the predictor, we found C4 again.[5]

6.2. S355

As we calculated our model for grade DD11, it is better to check another grade to be sure and crosscheck the model reliability if the results will be similar. For S355 we have 39,576 rows, and we will select the same features that we have chosen for DD11, and the same process will be followed for NaN values.[6]

First, we will check the regression results of Nan's dropped; we have got 0,18 regression score. Forward selection is implemented to test the features effect as a predictor. As we can see in below (Table 5) result that adjusted R-squared is 0,97 which is quite

[3] Appendix II - DD11 NaN's are filled by median, box plot of categories with the regression features
[4] Findings are detailed in Appendix V
[5] Findings are detailed in Appendix VI.
[6] Appendix III – S355 NaN's are dropped, box plot of categories with the regression features

acceptable for eight predictors such as Aluminum, Width, C1, C11, C2, CMIX, C3 and Titanium.

OLS Regression Results

Dep. Variable:	bidsa_PGL_Price	R-squared:	0.972
Model:	OLS	Adj. R-squared:	0.972
Method:	Least Squares	F-statistic:	1.076e+05
Date:	Thu, 23 Aug 2018	Prob (F-statistic):	0.00
Time:	17:01:13	Log-Likelihood:	28046.
No. Observations:	24656	AIC:	-5.608e+04
Df Residuals:	24648	BIC:	-5.601e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
itemsa_Aluminium	0.5768	0.074	7.840	0.000	0.433	0.721
itemsa_Width__mm_	1.392e-05	1.76e-06	7.897	0.000	1.05e-05	1.74e-05
['cat']_C1	0.4402	0.003	126.925	0.000	0.433	0.447
['cat']_C11	0.4649	0.004	124.961	0.000	0.458	0.472
['cat']_C2	0.3952	0.004	103.072	0.000	0.388	0.403
['cat']_CMIX	0.3320	0.004	83.948	0.000	0.324	0.340
['cat']_C3	0.3347	0.005	73.011	0.000	0.326	0.344
itemsa_Titanium	-0.7526	0.035	-21.583	0.000	-0.821	-0.684

Table 5 : Linear regression Result for 8 predictors

OLS Regression Results

Dep. Variable:	bidsa_PGL_Price	R-squared:	0.183
Model:	OLS	Adj. R-squared:	0.182
Method:	Least Squares	F-statistic:	688.4
Date:	Thu, 23 Aug 2018	Prob (F-statistic):	0.00
Time:	17:01:13	Log-Likelihood:	28146.
No. Observations:	24656	AIC:	-5.627e+04
Df Residuals:	24647	BIC:	-5.620e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
itemsa_Aluminium	0.4767	0.074	6.476	0.000	0.332	0.621
itemsa_Width__mm_	1.154e-05	1.76e-06	6.543	0.000	8.09e-06	1.5e-05
['cat']_C1	0.4467	0.003	128.207	0.000	0.440	0.454
['cat']_C11	0.4716	0.004	126.262	0.000	0.464	0.479
['cat']_C2	0.4022	0.004	104.461	0.000	0.395	0.410
['cat']_CMIX	0.3384	0.004	85.359	0.000	0.331	0.346
['cat']_C3	0.3413	0.005	74.375	0.000	0.332	0.350
itemsa_Titanium	-0.7430	0.035	-21.390	0.000	-0.811	-0.675
['cat']_C4	0.3921	0.028	14.221	0.000	0.338	0.446

Table 6 : Linear regression Result for 9 predictors

When we add one more predictor, our score is dropped to 0,18 when we check the 9th predictor; we found C4 again.

Last part we check the results for S355 after we fill Nan's of chemical features with median values, the same path will be followed to find the best scores.[7] Regressor score is 0,19, and it needs to apply forward selection to pick best predictors for the price.

As per below results, we found the best adjusted R-square score for seven predictors, 0,96. When we check the features, we found Aluminum, Width, C1, C11, C2, CMIX and C3. It is better to add other features one by one and check the results vary.[8]

One more predictor added, the result is crashed to 0,17 with an additional feature, when we check last added feature we found that it is C4 as expected.[9]

As a conclusion of our regression model, it was evident that there are robust features which are affecting pricing, width is the substantial factor for pricing for both grade. S355 is affected more from Aluminum content for price estimation with width. C1, C2, C3, CMIX, and C11 are the following variables that influenced pricing.

C4 is a negative factor for all scenarios and both grades; we suspect that there is something wrong with the content of this category. There are two possibilities came to our mind, first is C4 is the lowest part in the data and effect of the price is not much, second is this category is not categorized properly and it needs to be focused to fix it. This must be clarified with the data providers to understand the reason and improve the category assortment.

7. LIMITATIONS

Steel is a combination(alloy) of iron and carbon, and Carbon value makes a big difference to classify the steel grade. Steel quality is not only measured with chemicals but also hardness, tensile strength, and another physical process gives the final idea about steel grade. In our dataset, we have only chemical contents and bodily form as thickness and width. Therefore, we cannot get a specific or final classification with that information. We worked with limited sources and limited time. This work may continue further for customer segmentation and fix price recommendation.

[7] Appendix IV – S355 NaN's are filled by median, box plot of categories with the regression features
[8] Results are detailed in Appendix VII.
[9] Findings are detailed in Appendix VII.

8. CONCLUSION

This Project is aimed to make a proper clustering and price estimation for the non-prime products according to given features, when we cluster the products correctly and find the strong predictors of the price, customer segmentation per product can be done smoothly. The following step can be the auction recommendation system for non-prime materials according to customers tendencies and purchase histories; we can develop a model which is catching the right customer of the product with the fix price level. If the project and model can identify the fixed price for the product for the corresponding customer, the company can increase the profit levels for non-prime materials. On the other hand, if you set a good model this can improve the customer satisfaction as they will get the right products with their acceptable price levels.

For these reasons, project first step is started with the classification of the products as we have only mills determined categories in our data that is showing six groups according to defect and product features. Labeling of the products is affected with three factors for secondary materials; chemical contents, physical characteristics/process and central defect of the coils. Our data had chemical substances, width, and thickness, and we decided to cluster the products according to chemicals to check if we will get a rational result.

K-Means Clustering method is used to find the clusters to have the groups and compare with given grades or check with a professional if the results are acceptable. Because of data size, product and grade variety, we decided to select one product family; Hot Rolled Coils is chosen. K-means clustering results showed us that nine clusters best capture the segmentation of the data as the coefficient silhouette was highest when $k = 9$. When we show the results to the product expert, he checked each cluster chemical content levels and advised us that Carbon is the primary determinant to describe the grade of the material. For this reason, he referred to use Carbon Equivalent formula and create a new feature with the CE results and make the clustering again. When we finished the clustering with the new feature, we found that clusters did not undifferentiated than the first K-means clusters. To have different clusters, we decided to create a new group with only Carbon Equivalent values.

To crosscheck the k-means results, another classification model is used which is Agglomerative Hierarchical Clustering. We found the parallel results with k-means.

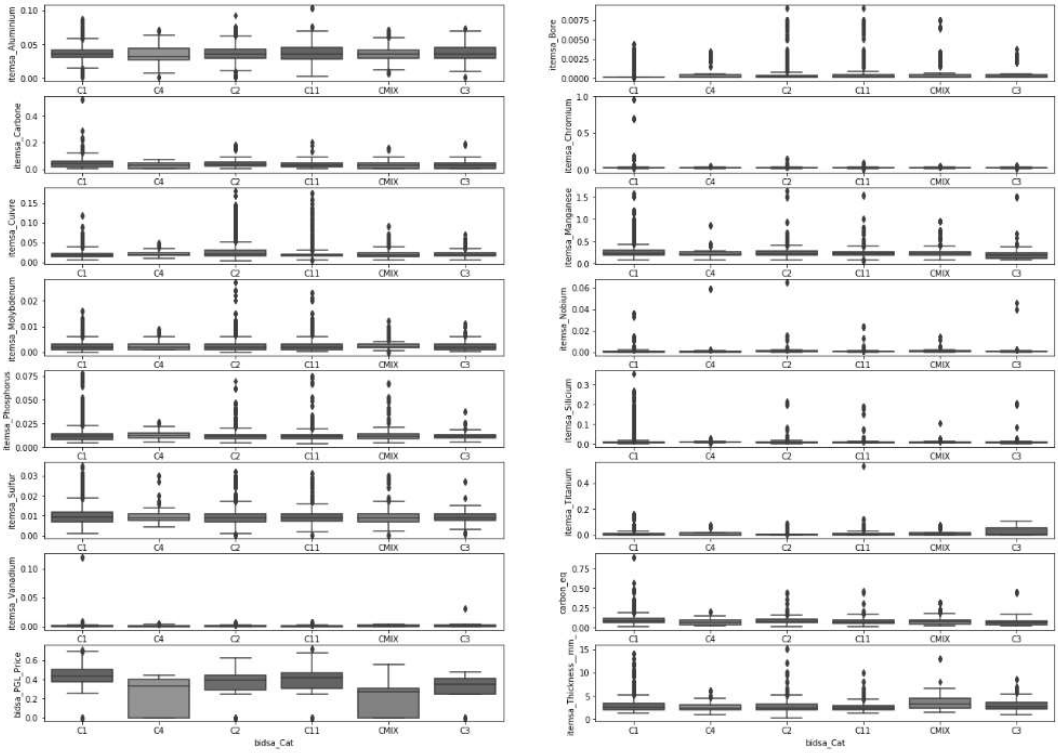
The second part of the project is the estimation of the products price and identify the strong predictors of the price estimation. Hence, the linear regression method selected as the best estimation technique for more predictors. First, we have applied the regression model to Hot Rolled Coils, we have got best score 0,75 for 11 predictors with forward selection method. When we add one more predictor, the result is crashed to 0,18. Last combined predictor was C4. To have better results, we go more in-depth with the product family and choose the most repeated grade which is DD11 from hot rolled products. Two different ways are followed, first is to drop NaN values under chemical contents, second is to fill the NaN rows with the median value for each chemical to compare the results to select best scores.

As we explained all details under Linear Regression part, Width is our main predictor for both grades and C4 is the predictor break for all regression results. Accordingly, we can use deductive that C4 category has a big difference about the products and classification of the C4 should be considered again for the pricing. Price Guideline Prices are created by marketing teams for each region according to market expectations may be PGL prices can be supported by the Machine Learning models to have better results and to increase the profitability.

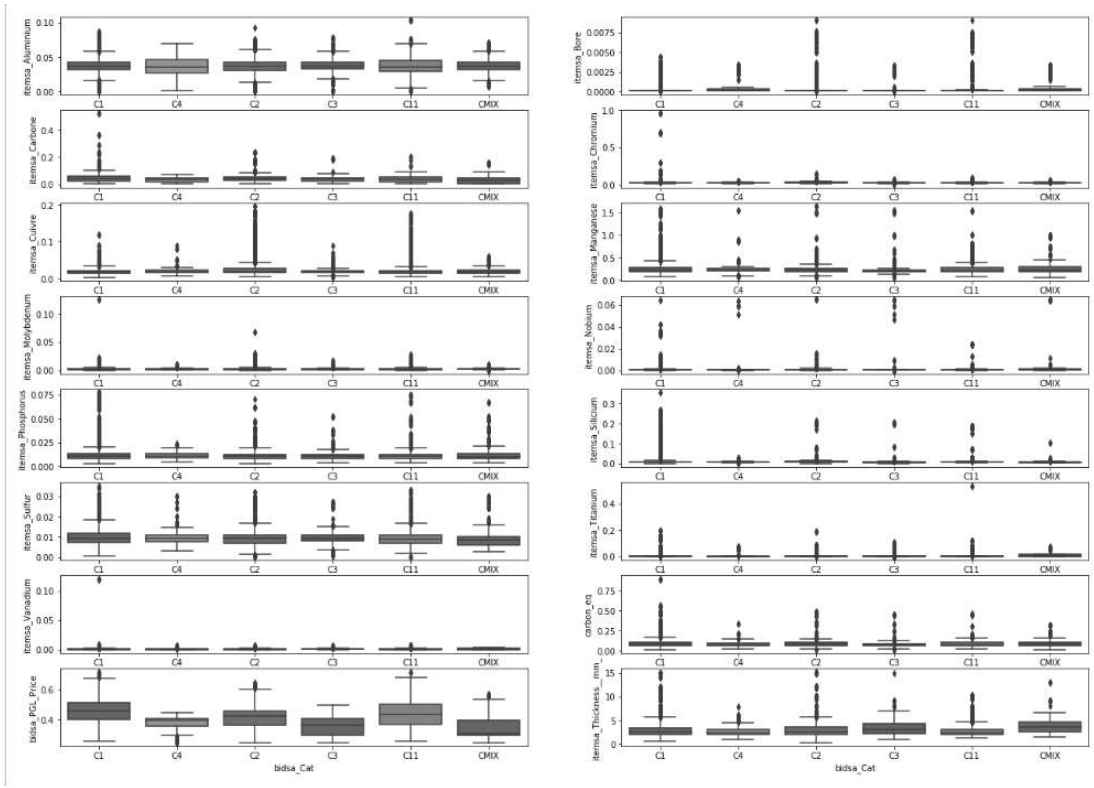
If we sum up the results of the project, product clustering and price estimation according to product features are the basic steps to go further analysis as customer segmentation and fix price recommendation system. This is the first time that I worked on a real dataset which I had to work a lot to understand and decide on which model to use. I could be able to finish basic two steps in a short time period, but this project can be extended for deeper analysis.

9. APPENDIX

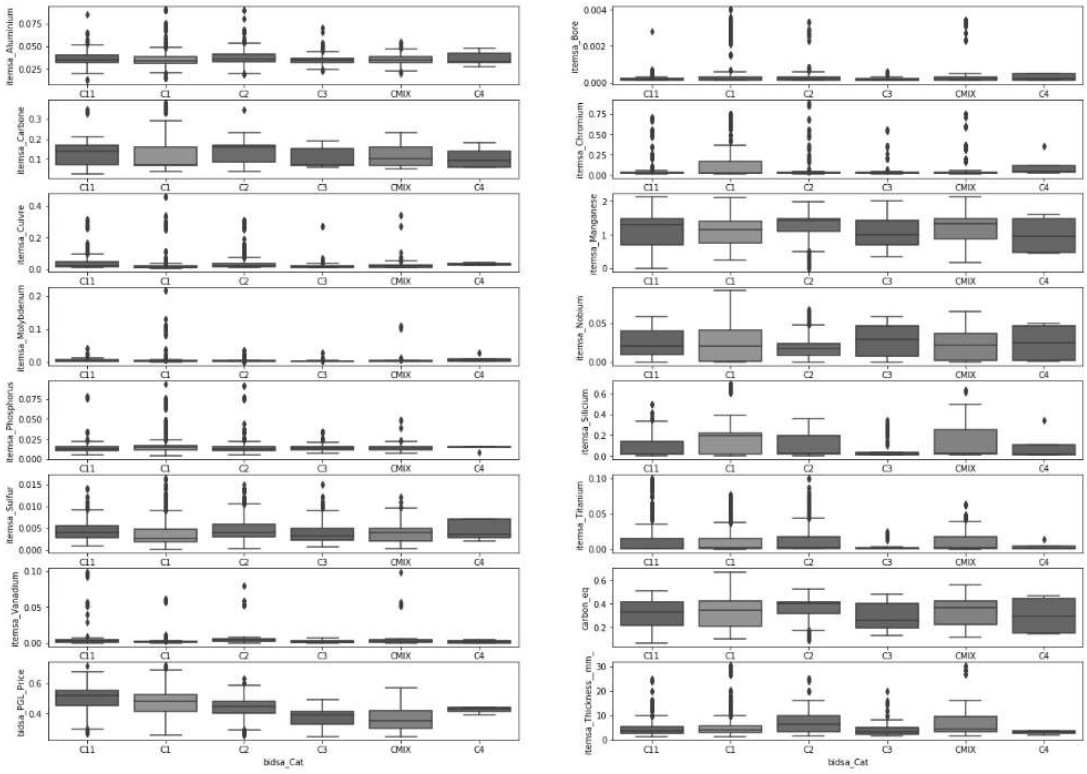
Appendix I – DD11 NaN's Dropped



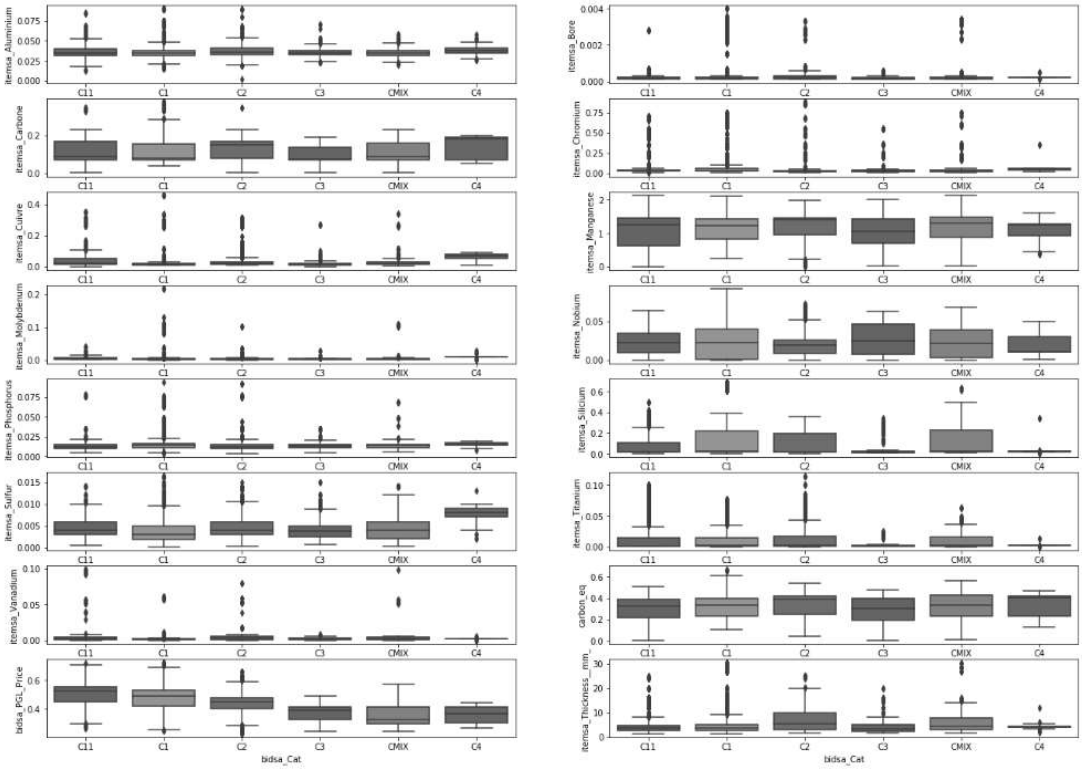
Appendix II – DD11 NaN's Filled by Median



Appendix III-S355 NaN's Dropped



Appendix IV-S355 NaN's filled by Median



Appendix V – HRC Regression Results for 11 and 12 Predictors

OLS Regression Results

Dep. Variable:	bidsa_PGL_Price	R-squared:	0.746
Model:	OLS	Adj. R-squared:	0.746
Method:	Least Squares	F-statistic:	5.488e+04
Date:	Wed, 29 Aug 2018	Prob (F-statistic):	0.00
Time:	16:20:38	Log-Likelihood:	52471.
No. Observations:	205577	AIC:	-1.049e+05
Df Residuals:	205566	BIC:	-1.048e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
itemsa_Width__mm_	8.194e-05	1.45e-06	56.605	0.000	7.91e-05	8.48e-05
itemsa_Sulfur	-1.5735	0.139	-11.286	0.000	-1.847	-1.300
['cat']_CMIX	0.2359	0.003	87.961	0.000	0.231	0.241
['cat']_C11	0.3460	0.002	139.640	0.000	0.341	0.351
['cat']_C1	0.3391	0.002	141.241	0.000	0.334	0.344
itemsa_Titanium	-1.1221	0.014	-77.491	0.000	-1.150	-1.094
itemsa_Molybdenum	-0.3423	0.007	-46.240	0.000	-0.357	-0.328
['cat']_C2	0.2763	0.003	105.246	0.000	0.271	0.281
itemsa_Silicium	-0.0759	0.001	-57.644	0.000	-0.078	-0.073
['cat']_C3	0.2139	0.003	63.255	0.000	0.207	0.221
carbon_eq	-0.2017	0.003	-57.959	0.000	-0.209	-0.195

OLS Regression Results

Dep. Variable:	bidsa_PGL_Price	R-squared:	0.184
Model:	OLS	Adj. R-squared:	0.184
Method:	Least Squares	F-statistic:	4223.
Date:	Thu, 06 Sep 2018	Prob (F-statistic):	0.00
Time:	19:05:15	Log-Likelihood:	53211.
No. Observations:	205577	AIC:	-1.064e+05
Df Residuals:	205565	BIC:	-1.063e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
itemsa_Width__mm_	6.629e-05	1.5e-06	44.240	0.000	6.34e-05	6.92e-05
itemsa_Sulfur	-2.9717	0.144	-20.697	0.000	-3.253	-2.690
['cat']_CMIX	0.2719	0.003	96.052	0.000	0.266	0.277
['cat']_C11	0.3805	0.003	144.891	0.000	0.375	0.386
['cat']_C1	0.3747	0.003	146.104	0.000	0.370	0.380
itemsa_Titanium	-1.1692	0.014	-80.746	0.000	-1.198	-1.141
itemsa_Molybdenum	-0.3411	0.007	-46.239	0.000	-0.356	-0.327
['cat']_C2	0.3133	0.003	112.435	0.000	0.308	0.319
itemsa_Silicium	-0.0799	0.001	-60.703	0.000	-0.082	-0.077
['cat']_C3	0.2492	0.003	71.376	0.000	0.242	0.256
carbon_eq	-0.2277	0.004	-64.455	0.000	-0.235	-0.221
['cat']_C4	0.2510	0.007	38.536	0.000	0.238	0.264

Appendix VI - DD11 Regression Results for nine Predictors

OLS Regression Results

Dep. Variable:	bidsa_PGL_Price	R-squared:	0.963
Model:	OLS	Adj. R-squared:	0.963
Method:	Least Squares	F-statistic:	1.745e+05
Date:	Thu, 23 Aug 2018	Prob (F-statistic):	0.00
Time:	23:12:49	Log-Likelihood:	62936.
No. Observations:	60665	AIC:	-1.259e+05
Df Residuals:	60656	BIC:	-1.258e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
itemsa_Width__mm_	6.221e-05	1.19e-06	52.145	0.000	5.99e-05	6.45e-05
itemsa_Aluminium	0.6005	0.032	18.927	0.000	0.538	0.663
itemsa_Sulfur	0.8109	0.093	8.724	0.000	0.629	0.993
itemsa_Manganese	0.0354	0.003	12.627	0.000	0.030	0.041
['cat']_C1	0.3372	0.002	163.708	0.000	0.333	0.341
['cat']_C11	0.3237	0.002	157.097	0.000	0.320	0.328
['cat']_C2	0.2941	0.002	134.941	0.000	0.290	0.298
['cat']_C3	0.2432	0.003	94.042	0.000	0.238	0.248
['cat']_CMIX	0.2287	0.003	79.879	0.000	0.223	0.234

Appendix VII - DD11 Regression Result for 10 Predictors

OLS Regression Results

Dep. Variable:	bidsa_PGL_Price	R-squared:	0.126
Model:	OLS	Adj. R-squared:	0.126
Method:	Least Squares	F-statistic:	971.4
Date:	Thu, 23 Aug 2018	Prob (F-statistic):	0.00
Time:	23:12:49	Log-Likelihood:	65811.
No. Observations:	60665	AIC:	-1.316e+05
Df Residuals:	60655	BIC:	-1.315e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
itemsa_Width__mm_	3.787e-05	1.18e-06	32.089	0.000	3.56e-05	4.02e-05
itemsa_Aluminium	0.0573	0.031	1.847	0.065	-0.004	0.118
itemsa_Sulfur	-0.7856	0.091	-8.633	0.000	-0.964	-0.607
itemsa_Manganese	-0.0065	0.003	-2.380	0.017	-0.012	-0.001
['cat']_C1	0.4139	0.002	188.239	0.000	0.410	0.418
['cat']_C11	0.3965	0.002	182.104	0.000	0.392	0.401
['cat']_C2	0.3684	0.002	160.998	0.000	0.364	0.373
['cat']_C3	0.3170	0.003	119.934	0.000	0.312	0.322
['cat']_CMIX	0.3032	0.003	104.753	0.000	0.298	0.309
['cat']_C4	0.3442	0.004	77.654	0.000	0.336	0.353

Appendix VIII- S355 Rgression Results for 7 Predictors

OLS Regression Results

Dep. Variable:	bidsa_PGL_Price	R-squared:	0.969
Model:	OLS	Adj. R-squared:	0.969
Method:	Least Squares	F-statistic:	1.464e+05
Date:	Thu, 23 Aug 2018	Prob (F-statistic):	0.00
Time:	17:19:15	Log-Likelihood:	35041.
No. Observations:	32304	AIC:	-7.007e+04
Df Residuals:	32297	BIC:	-7.001e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
itemsa_Aluminium	1.0186	0.064	15.895	0.000	0.893	1.144
itemsa_Width__mm_	2.362e-05	1.51e-06	15.597	0.000	2.07e-05	2.66e-05
['cat']_C1	0.4081	0.003	139.475	0.000	0.402	0.414
['cat']_C11	0.4309	0.003	138.164	0.000	0.425	0.437
['cat']_C2	0.3584	0.003	111.646	0.000	0.352	0.365
['cat']_CMIX	0.2917	0.003	83.594	0.000	0.285	0.299
['cat']_C3	0.3050	0.004	78.168	0.000	0.297	0.313

Appendix IX - S355 Regression Results for 8 Predictors

OLS Regression Results

Dep. Variable:	bidsa_PGL_Price	R-squared:	0.176
Model:	OLS	Adj. R-squared:	0.176
Method:	Least Squares	F-statistic:	986.3
Date:	Thu, 23 Aug 2018	Prob (F-statistic):	0.00
Time:	17:19:15	Log-Likelihood:	35924.
No. Observations:	32304	AIC:	-7.183e+04
Df Residuals:	32296	BIC:	-7.176e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
itemsa_Aluminium	0.1553	0.066	2.368	0.018	0.027	0.284
itemsa_Width__mm_	6.96e-06	1.52e-06	4.565	0.000	3.97e-06	9.95e-06
['cat']_C1	0.4598	0.003	148.555	0.000	0.454	0.466
['cat']_C11	0.4835	0.003	147.563	0.000	0.477	0.490
['cat']_C2	0.4134	0.003	122.302	0.000	0.407	0.420
['cat']_CMIX	0.3430	0.004	95.201	0.000	0.336	0.350
['cat']_C3	0.3570	0.004	89.509	0.000	0.349	0.365
['cat']_C4	0.3463	0.008	42.588	0.000	0.330	0.362

10.REFERENCES

- Hof, Robert D. (2013). Artificial intelligence is finally getting smart. Retrieved from https://technologyreview.com/s/513696/deep-learning/
- Hardesty, Larry. (2013, May 29). How computers can learn better. Retrieved from [http://news.mit.edu/2013/machine-learning-algorithm-outperforms-predecessors-0529]
- Gutierrez, Daniel. (2016, Sept 20). Three Barriers to Machine Learning Adoption. Retrieved from https://insidebigdata.com/2016/09/20/three-barriers-to-machine-learning-adoption/
- An Efficient k-means Clustering Algorithm: Analysis and Implementation by Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu.
- 'P. T. FITZROY (1976) Analytical Methods for Marketing Management. McGraw-Hill, England.
- J. P. GUILTIMAN (1976) Risk-aversive pricing policies: problems and alternatives. J. Marketing
- van Kampen, T. J., Akkerman, R., & van Donk, D. P. (2012). SKU classification: A literature review and conceptual framework. International Journal of Operations and Production Management, 32(7), 850-876. DOI: 10.1108/01443571211250112
- Pepall, L. (1990). Market Demand and Product Clustering. The Economic Journal, 100(399), 195. doi:10.2307/2233603
- Hu M and Liu B. Mining and summarizing customer reviews. in Proceedings of SIGKDD. 2004.168-177
- Pang B and Lee L, Opinion Mining and Sentiment Analysis. Foundations and Trends in IR. 2008. 1-135.
- A Model for the Evaluation of Steel Pricing Options Subject to Administrative Guidelines Author(s): R. W. Bednarz and B. J. Garner
- Statistics Solutions. (2013). What is Linear Regression? Retrieved from http://statisticssolutions.com/what-is-linear-regression/

- J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297
- Benjamin C. M. Fung, Ke Wang, and Martin Ester, Simon Fraser (2014). Hierarchical Document Clustering. Retrieved from https://www.researchgate.net/publication/314455559_Hierarchical_Document_Clustering
- Statistics Solutions. (2013). What is Linear Regression. Retrieved from http://www.statisticssolutions.com/what-is-linear-regression/