SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

STATISTICAL ANALYSIS AND MODELING FOR THE MATERIAL RECOVERY OF THE ALUMINUM PRODUCTION PROCESS

By

Kekempanos Aggelos, Nikolaou

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
July 2016

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΑΝΑΚΤΗΣΗ ΥΛΗΣ ΓΙΑ ΤΗΝ ΠΑΡΑΓΩΓΙΚΗ ΔΙΑΔΙΚΑΣΙΑ ΚΑΤΕΡΓΑΣΙΑΣ ΑΛΟΥΜΙΝΙΟΥ

Κεκεμπάνος Άγγελος, Νικολάου

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Ιούλιος 2016

DEDICATION

I dedicate this thesis to my parents because they are always stand by me and support my decisions.

ACKNOWLEDGEMENTS

I would first like to express my gratitude to my thesis supervisor Prof Dimitris Karlis because whenever I ran into a trouble spot or had a question about my research or writing, he always was there to help and support me. Also, I would like to thank Dr Kaskavelis Elefterios and generally the members of the ELVAL Company who helped me to complete this thesis and gave me a very interesting subject for analysis with a scholarship.

VITA

I was born in Karystos, Greece, on December 26, 1989. I finished high school in 2007 and I was accepted by the Mathematics department of the university of Ioannina in 2009. After four years and a half I got my Mathematics Bachelors Degree with specialization in Statistics and Operation Research. Also, I finished my compulsory lessons of MSc in Statistics to Athens University of Economics and Business in 2015.

ABSTRACT

KEKEMPANOS AGGELOS

STATISTICAL ANALYSIS AND MODELING FOR THE MATERIAL RECOVERY OF THE ALUMINUM PRODUCTION PROCESS

July 2016

The ranges of topics that can be studied using statistical methods grow as statistics science evolves. A very interesting topic that we will try to analyze statistically and modeling in this thesis is the material recovery of the aluminum production process as this done in the ELVAL's factory. This process is very complex and has many steps which affect the quality and the value of the final product. Our data consists of measures which have been made in these steps. Namely, the most important measures are those which describe the initial weight divided by the produced which is called Return Index and the produced weight divided by the initial which called Recovery Index. The Return Index will be used in the descriptive analysis of this thesis and the Recovery Index in the model analysis because is located between zero and one. Our main goal is to find where the most problems occurred during the production process, which factors affect more the final product and how can we predict the Recovery Index better having the minimum error. It is important to emphasize that the descriptive part of our research will be based on all data whereas modeling will be done separately for the two types of process, the Hot Rolling and the Continuous Casting because there are variables which are not defined in both cases. Furthermore, for the descriptive analysis of this thesis which will give us a first view of our data and some very significant results, will be used simple statistics measures such as Pearson Correlation Coefficient, Spearman Correlation Coefficient, Skewness, Kurtosis etc. and figures such as Barplots, Pareto Charts, Box Plots etc. Due to the fact that the diagrammatic representation is a very important part of a statistical analysis, the graphs in this thesis will be done with the ggplot library which gives us great flexibility. Moreover, the modeling part of this thesis in which we will try to find which factors affects more the final result and how can we predict it, will be included multiple linear models, BIC criterion, penalized regression (Lasso), CART algorithms (Regression Tree) and Cross Validation methods. Finally, for the implementation of this thesis the statistical package which has been used is the programming language R. Further analysis and particularities of the problem references to chapter Introduction.

ΠΕΡΙΛΗΨΗ

ΚΕΚΕΜΠΑΝΟΣ ΑΓΓΕΛΟΣ

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΑΝΑΚΤΗΣΗ ΥΛΗΣ ΓΙΑ ΤΗΝ ΠΑΡΑΓΩΓΙΚΗ ΔΙΑΔΙΚΑΣΙΑ ΚΑΤΕΡΓΑΣΙΑΣ ΑΛΟΥΜΙΝΙΟΥ

Ιούλιος 2016

Το εύρος θεμάτων που μπορούν να μελετηθούν χρησιμοποιώντας στατιστικές μεθόδους μεγαλώνει όσο η στατιστική σαν επιστήμη εξελίσεται. Ένα τέτοιο θέμα είναι και η ανάκτηση ύλης για την παραγωγική διαδικασία κατεργασίας αλουμινίου όπως αυτή συμβαίνει στο εργοστάσιο της ΕΛΒΑΛ. Η συγκεκριμένη διαδικασία είναι αρκετά πολύπλοκη και έχει στάδια τα οποία επηρεάζουν την ποιότητα και την αξία του τελικού προιόντος. Τα δεδομένα που θα χρησιμοποιηθούν στην έρευνα αποτελούνται απο μετρήσεις που έγιναν σε αυτά τα βήματα. Ειδικότερα οι πιο σημαντικές μετρήσεις είναι αυτές που περιγράφουν το παραγώμενο βάρος πρός το τελικό και το αντίστροφο και ονομάζονται Συνετεστής Επιστροφής και Συντελεστής Ανάκτησης αντίστοιχα. Ο Συντελεστής Επιστροφής θα χρησιμοποιθεί για το περιγραφικό κομμάτι της ερευνάς μας και ο Συντελεστής Ανάκτησης για το κομμάτι της μοντελοποίησης επειδή βρίσκεται μεταξύ του μηδενός και του ένα. Κύριως στόχος μας είναι να εντοπίσουμε που εμφανίζονται τα περισσότερα προβλήματα κατά την διάρκεια της παραγωγικής διαδικασίας, ποιοί παράγοντες επηρεάζουνε περισσότερο το τελικό αποτέλεσμα και να βρούμε ένα μοντέλο που προβλέπει τον Συντελεστή Ανάκτησης με το μικρότερο δυνατό σφάλμα. Είναι πολυ σημαντικό να επισημάνουμε οτι στο περιγραφικό κομμάτι της ερευνάς μας θα είναι βασισμένο σε όλα τα δεδομένα ενω η μοντελοποίηση θα γίνει ξεχωριστά για τα δύο είδη επεξεργασίας που ακολουθούνται, την Θερμή Έλαση και την Συνεχή Χύτευση αφού υπάρχουν μεταβλητές που δεν ορίζονται και στις δύο περιπτώσεις. Επιπλέον στο κομμάτι της περιγραφική ανάλυση μέσα απο το οποίο θα έχουμε μια πρώτη επαφή με τα δεδομένα μας και μερικά πολύ σημαντικά συμπεράσματα, θα χρησιμοποιηθούν απλά μέτρα περιγραφικής στατιστικής Pearson Correlation Coefficient, Spearman Correlation Coefficient, Skewness, Kurtosis κτλ και διαγράμματα Barplots, Pareto Charts, Box Plots κτλ. Επειδή η διαγραμματική απεικόνιση είναι ένα πολυ σημαντικό κομμάτι για μία στατιστική έρευνα, τα διαγράμματα που περιέχονται στην διπλωματική έρευνα θα υλοποιηθούν με την βιβλιοθήκη ggplot η οποία μας δίνει ιδιαίτερη ευελιξία. Επιπροσθέτως στο κομμάτι της μοντελοποίησης στο οποίο θα προσπαθήσουμε να βρούμε ποιοί παράγοντες επηρεάζουν περισσότερο το τελικό αποτέλεσμα και πως μπορούμε να προβλέψουμε αυτό καλύτερα, θα περιέχονται πολλαπλά γραμμικά μοντέλα, το κριτήριο BIC, ποινικοποιημένη παλινδρόμιση (Lasso), αλγόριθμοι CART (Regression Tree) και μεθόδοι Cross Validation. Τέλος, για την υλοποίηση της διπλωματικής έρευνας το στατιστικό πακέτο που χρησιμοποιήθηκε ήταν η γλώσσα προγραμματισμού R. Περαιτέρω ανάλυση και ιδιαιτερότητες του προβλήματος αναφέρονται στο κεφάλαιο της Εισαγωγής.

Table of Contents

Table of Figures

List of Tables

CHAPTER 1

INTRODUCTION

The main goal of this thesis is to analyze statistically and modeling the material recovery of the aluminum production process in ELVAL Company, based on regression models, penalized regression and CART models. Before starting this analysis and applying these methods, it is very important to understand the nature of this problem. For this reason we have to comprehend the structure of the production process in the ELVAL's company. ELVAL is an aluminum processing factory which imports aluminum in primary form or from recycling (Scrap) and through a production process adds value to it. So, let's take a look at the steps of the production process.

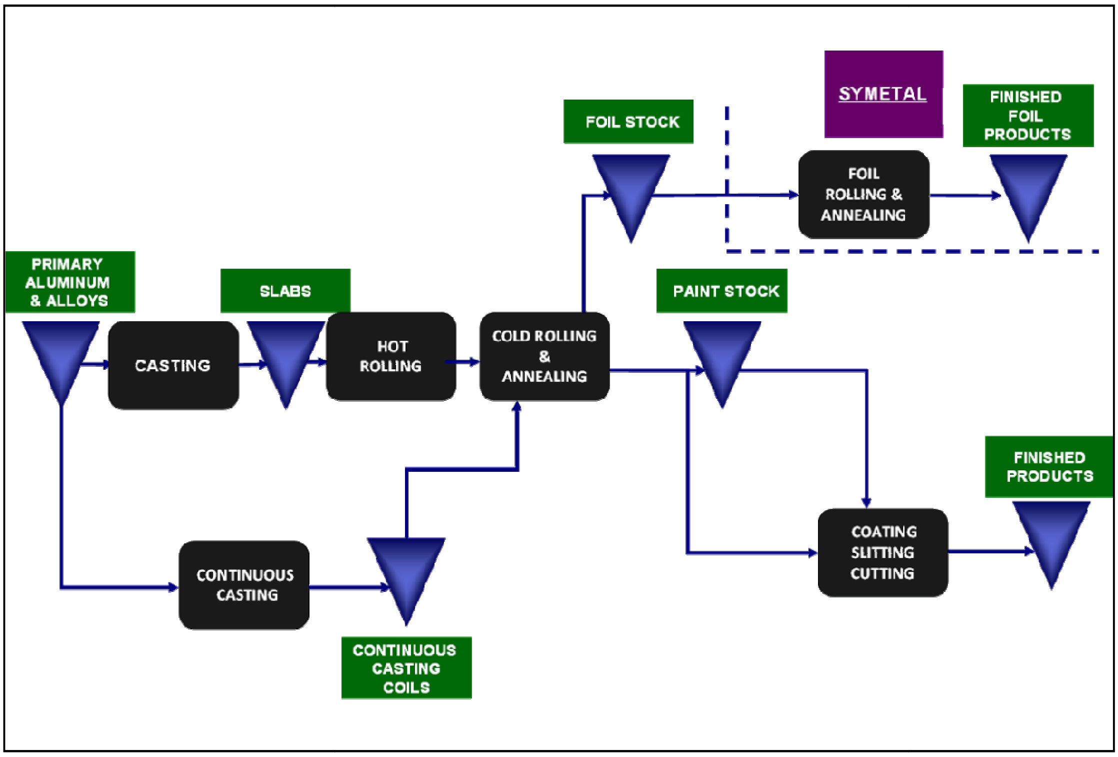


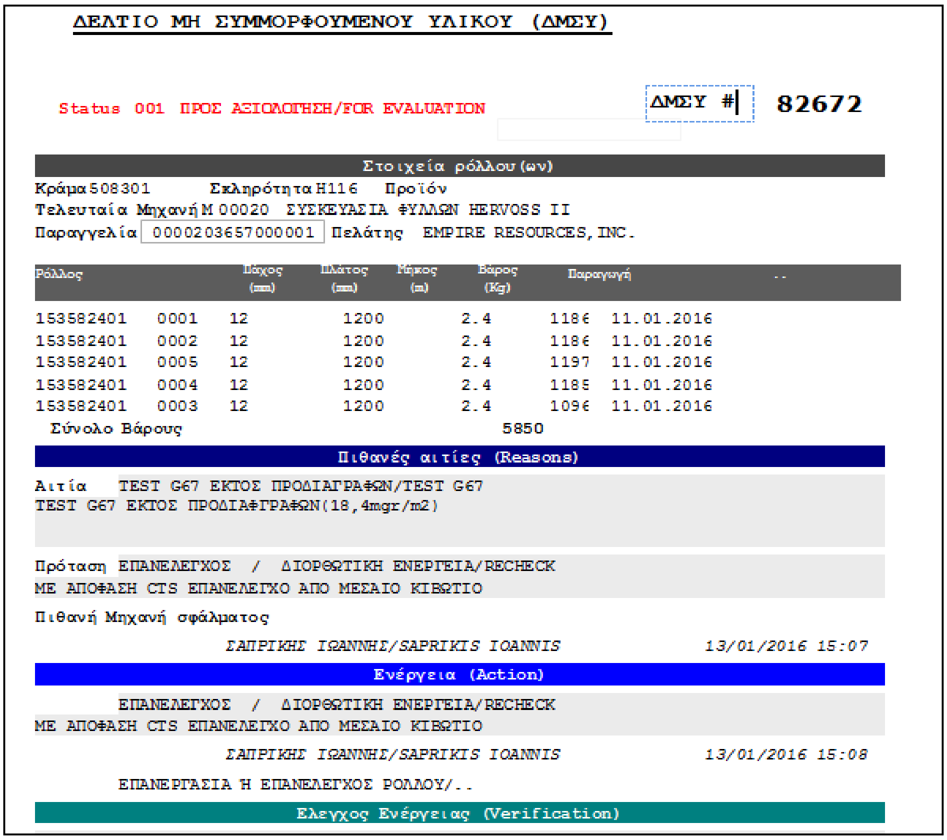
Figure 1.1.1: Production process flow chart

The production of aluminum coils and sheets as shown in the Figure 1.1.1 have the following steps.

- Aluminum alloying and casting into slabs, which are 0.6 m in thickness, up to 2.5 m in width and up to 8 m in length.
- Hot rolling, where a slab is shaped into a coil that is few millimeters in thickness.
- As an alternative to the two previous stages, continuous casting is used, where melted aluminum is cast directly into coils that are a few millimeters in thickness.
- Cold rolling, where a coil produced with the hot rolling or continuous casting method reaches the thickness specified for an end product.

- Intermediate or final thermal processing of the coils or end products, in furnaces, to acquire such required product properties as hardness, easy further processing, etc.
- Coating or other processing of the aluminum coil surface (e.g. tread plate).
- Cutting the basic coil into coils or foils, using special machinery.
- End product packaging in bundles or pallets, using advanced methods and materials that will protect the aluminum until its end use.

If an error exists in the production process, this will be mentioned at the DMSY report which has the following form:



ΔΕΛΤΙΟ ΜΗ ΣΥΜΜΟΡΦΟΥΜΕΝΟΥ ΥΛΙΚΟΥ (ΔΜΣΥ)

Status 001 ΠΡΟΣ ΑΞΙΟΛΟΓΗΣΗ/FOR EVALUATION

ΔΜΣΥ # **82672**

Στοιχεία ρόλλου(ων)

Κράμα 508301 Σκληρότητα H116 Προϊόν
Τελευταία Μηχανή M 00020 ΣΥΣΚΕΥΑΣΙΑ ΦΥΛΛΩΝ HERVOSS II
Παραγγελία 000020365700001 Πελάτης EMPIRE RESOURCES, INC.

Ρόλλος		Πάχος (mm)	Πλάτος (mm)	Μήκος (m)	Βάρος (Kg)	Παραγωγή	..
153582401	0001	12	1200	2.4	1186	11.01.2016	
153582401	0002	12	1200	2.4	1186	11.01.2016	
153582401	0005	12	1200	2.4	1197	11.01.2016	
153582401	0004	12	1200	2.4	1185	11.01.2016	
153582401	0003	12	1200	2.4	1096	11.01.2016	
Σύνολο Βάρους					5850		

Πιθανές αιτίες (Reasons)

Αιτία TEST G67 ΕΚΤΟΣ ΠΡΟΔΙΑΓΡΑΦΩΝ/TEST G67
TEST G67 ΕΚΤΟΣ ΠΡΟΔΙΑΦΓΡΑΦΩΝ(18,4mgr/m2)

Πρόταση ΕΠΑΝΕΛΕΓΧΟΣ / ΔΙΟΡΘΩΤΙΚΗ ΕΝΕΡΓΕΙΑ/RECHECK
ΜΕ ΑΠΟΦΑΣΗ CTS ΕΠΑΝΕΛΕΓΧΟ ΑΠΟ ΜΕΣΑΙΟ ΚΙΒΩΤΙΟ

Πιθανή Μηχανή σφάλματος

ΣΑΠΡΙΚΗΣ ΙΩΑΝΝΗΣ/SAPRIKIS IOANNIS *13/01/2016 15:07*

Ενέργεια (Action)

ΕΠΑΝΕΛΕΓΧΟΣ / ΔΙΟΡΘΩΤΙΚΗ ΕΝΕΡΓΕΙΑ/RECHECK
ΜΕ ΑΠΟΦΑΣΗ CTS ΕΠΑΝΕΛΕΓΧΟ ΑΠΟ ΜΕΣΑΙΟ ΚΙΒΩΤΙΟ

ΣΑΠΡΙΚΗΣ ΙΩΑΝΝΗΣ/SAPRIKIS IOANNIS *13/01/2016 15:08*

ΕΠΑΝΕΡΓΑΣΙΑ Ή ΕΠΑΝΕΛΕΓΧΟΣ ΡΟΛΛΟΥ/..

Ελεγχος Ενέργειας (Verification)

Figure 1.1.2: DMSY report

In DMSY report is written on the cause of the problem, the action which is performed after error's detection, the next department and the machine in which error appeared.

Furthermore, we have two very important criterions which characterize the quality of the process: The Return Index which describes the initial weight divided by the produced and the Recovery Index which is the produced weight divided by the initial. The first one will be used in the descriptive part of this thesis and the second one in the modeling because it takes values between zero and one. Moreover, the descriptive part of our research will be based on all data whereas modeling will be done separately for the two types of processes, the Hot Rolling and the Continuous Casting because there are variables that are not defined in both cases. Our main goal is to find where the most problems occurred during the production process, which factors affect more the final product and how can we predict the Recovery Index better having the minimum error. Firstly, let's see the available data and will be used for our analysis.

The available variables are the following:

1. RmatID: Initial material code
2. Year: The year in which the observation was mentioned.
3. Month: The month in which the observation was mentioned
4. General Category: General category of the final product.
5. Production: Aluminum weight which is exported by the production process.
6. Initial Weight: Aluminum weight which is imported in the production process.
7. DMSY: Error existence or not in the production process.
8. Painting Line 1: The line which used to paint the product.
9. Supplier: Supplier of the initial material.
10. Alloy: The aluminum alloy which is used.
11. Temper: The toughness of the alloy.
12. Utilization: Utilization or not of the defective product.
13. Coil Thickness: End thickness of the coil.
14. Coil Width: End width of the coil.
15. Initial Length: The initial length of the slab which imported in the production process.
16. Hot Mill: The type of the rolling (Continuous Casting, Tippins, Old Hot Mill).
17. WorkID: The process which applied in the Hot Rolling (R: Rolling without trimming, RT: rolling with trimming, R0: Hot+Bliss and RBR: Rolling for bright materials.
18. Slab Width: The width of the slab which imported in the Hot Rolling process.
19. Hot Mill Exit Width: Exit width from the Hot Rolling.
20. Hot Mill Thickness: Exit thickness from the Hot Rolling.
21. Trimmed Width: Slab Width – Coil Width.
22. Texit: Exit temperature from the Hot Rolling.
23. Reason_ID: Cause of the problem which appeared during the production process.
24. Final_Action_ID: Action which performed after a DMSY detection.
25. Destination_ID: The next department which follows after a DMSY detection.
26. Mach_Resid: The machine in which DMSY appeaed.
27. Thickness Reduction: : $\frac{\text{Hot Mill Thickness}}{\text{Coil Thickness}}$
28. Scalper Return Index: $\frac{\text{Scalper Imported Weight}}{\text{Scalper Exit Weight}}$
29. Hot Mill Return Index: $\frac{\text{Hot Mill Imported Weight}}{\text{Hot Mill Exit Weight}}$
30. Hot Rolling Return Index: $\frac{\text{Scalper Imported Weight}}{\text{Hot Mill Exit Weight}}$

31. Return Index: $\frac{\text{Initial Weight}}{\text{Produced Weight}}$

32. Recovery Index: $\frac{\text{Produced Weight}}{\text{Initial Weight}}$, belongs to interval [0,1).

The variables Slab Width, WorkID, Hot Mill Exit Width, Hot Mill Thickness, Texit, Trimmed Width, Thickness Reduction, Scalper Return Index, Hot Mill Return Index and Hot Rolling Return Index are defined only in the Hot Rolling and that is the reason why will do the modeling in different parts, one for the Hot Rolling and one for the Continuous Casting. Also the variables Reason_ID, Final_Action_ID, Destination_ID ans Mach_Resid get values only when there is a DMSY report.

Since we briefly described the steps of the production process and the nature of the problem, we should find statistical methods to deal it. The most important part of using a model or generally a statistical method with the right way is to understand completely the theoretical background. In Chapter 2 will be deal with the theoretical background of the models and statistical methods which contained in this thesis.

CHAPTER 2

THERORITCAL BACKGROUND

In this chapter we mention the theoretical background related to the methology used in this thesis. We will deal with Multiple Regression Model, BIC criterion, Lasso regression, CART models and Cross Validation Methods.

2.1 Multiple Regression Model

In several problems the response variable Y can be considered to be affected by more the one explanatory variables, such as $\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{p}$. We can use a linear model to investigate the dependence between Y and $\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{p}$. The model which is a generalization of the simple linear model $Y = \beta_0 + \beta_1 X + \varepsilon$ will have the form

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_{\mathrm{i}1} + \beta_2 \mathrm{X}_{\mathrm{i}2} + \cdots + \beta_\mathrm{p} \mathrm{X}_{\mathrm{ip}} + \varepsilon_i \text{ , i} = 1,2,\dots,\mathrm{n}$$

where errors $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are considered independent and identically random variables (i.i.d.) from $\mathrm{N}(0, \sigma^2)$ and the explanatory variables $\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{p}$ as in simple linear model are not considered random. The above model is written using matrices in a simpler form:

$$\mathrm{Y} = X\beta + \varepsilon$$

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathrm{X} = \begin{bmatrix} 1 & \mathrm{X}_{11} & \cdots & \mathrm{X}_{1\mathrm{p}} \\ 1 & \mathrm{X}_{21} & \cdots & \mathrm{X}_{2\mathrm{p}} \\ \vdots & \vdots & & \vdots \\ 1 & \mathrm{X}_{\mathrm{n}1} & \cdots & \mathrm{X}_{\mathrm{np}} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The parameters b of the model and the variance σ^2 of the errors are the uknown parameters of the model. Also, the matrix X called design matrix. It is very important to notice that in the multiple regression we consider that the points are not close to a straight line but near a p+1 dimensioned hyperplane (Freedman, 2009).

Parameter Estimation

As reference to the parameter of the model we assume that the random vector ε consists of n independent $\mathrm{N}(0, \sigma^2)$ random variables and therefore it will have joint pdf $N_n(0, \sigma^2 \mathrm{I}_\mathrm{n})$, that follows a multivariate normal distribution where I_n is the identity matrix with dimension n. Therefore, the random vector $\mathrm{Y} = [\mathrm{Y}_1, \mathrm{Y}_2, \dots, \mathrm{Y}_\mathrm{n}]^\mathrm{T}$ will also follow multivariate normal $N_n(\mathrm{X}\beta, \sigma^2 \mathrm{I}_\mathrm{n})$.

At this point is very important to give the definition of the multivariate normal distribution. In probability theory and statistics, the multivariate normal distribution or multivariate Gaussian distribution, is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions. One possible definition is that a random vector is said to be *k*-variate normally distributed if every linear combination of its *k* components has a univariate normal distribution. Its importance derives mainly from the multivariate central limit theorem. The multivariate normal distribution is often used to describe, at least approximately, any set of (possibly) correlated real-valued random variables each of which clusters around a mean value (Gut, 2009).

If $X \sim N_{\kappa}(\mu, \mathrm{K})$ then

$$f_x(x_1, \dots, x_k) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

where x is k-dimensional column vector, Σ a symmetric covariance matrix which is positive defined and $|\Sigma|$ is the determinant of Σ.

So the likelihood function will be

$$L(\beta, \sigma^2) = \mathrm{f}(\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_\mathrm{n}; \beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(\mathrm{y}-\mathrm{X}\beta)^\mathrm{T}(\mathrm{y}-\mathrm{X}\beta)}$$

To get the M.L.E. of β is enough to maximize the likelihood subject to β or minimize the $(\mathrm{y}-\mathrm{X}\beta)^T(\mathrm{y}-\mathrm{X}\beta) = \varepsilon^T \varepsilon = \sum_{i=1}^{n} \varepsilon_i{}^2$subject to b. The previous equation is written as

$$\left(\mathrm{Y}^\mathrm{T} - \beta^\mathrm{T}\mathrm{X}^\mathrm{T}\right)\ (\mathrm{Y}-\mathrm{X}\beta) = \mathrm{Y}^\mathrm{T}\mathrm{Y} - \mathrm{Y}^\mathrm{T}\mathrm{X}\beta - \beta^\mathrm{T}\mathrm{X}^\mathrm{T}\mathrm{Y} + \beta^\mathrm{T}\mathrm{X}^\mathrm{T}\mathrm{X}\beta$$

and derivative subject to β $\frac{\partial \mathrm{f}}{\partial \beta} = \left(\frac{\partial \mathrm{f}}{\partial \beta_0}, \dots, \frac{\partial \mathrm{f}}{\partial \beta_\mathrm{p}}\right)$ is equal to

$$\frac{d}{db}(\mathrm{Y}-\mathrm{X}\beta)^\mathrm{T}(\mathrm{Y}-\mathrm{X}\beta) = -2\mathrm{X}^\mathrm{T}\mathrm{Y} + 2\mathrm{X}^\mathrm{T}\mathrm{X}\beta$$

The above derivative (ie. the vector of partial derivatives) is equal to zero when

$$\mathrm{X}^\mathrm{T}\mathrm{X}\beta = \mathrm{X}^\mathrm{T}\mathrm{Y}$$

This p+1 equations with p+1 unknowns has unique solution when the inverse of $\mathrm{X}^\mathrm{T}\mathrm{X}$ exists and in this case the maximum likelihood estimators of $\beta = \left[\beta_0, \beta_1, \dots, \beta_\mathrm{p}\right]^\mathrm{T}$ will be

$$\hat{\beta} = (\mathrm{X}^\mathrm{T}\mathrm{X})^{-1}\mathrm{X}^\mathrm{T}\mathrm{Y}$$

Also, an unbiased estimator of σ^2 is

$$\widehat{\sigma^2} = \frac{1}{\mathrm{n}-\mathrm{p}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y}_\mathrm{i} - \hat{\mathrm{Y}}_\mathrm{i})^2 \equiv \mathrm{S}^2$$

Predictions of Y_i are given by

$$\hat{\mathrm{Y}}_\mathrm{i} = \widehat{\beta_0} + \widehat{\beta}_1\mathrm{X}_{\mathrm{i}1} + \dots + \widehat{\beta}_\mathrm{p}\mathrm{X}_{\mathrm{ip}}\ , \mathrm{i} = 1,2, \dots, \mathrm{n}$$

or in matrix form

$$\hat{\mathrm{Y}} = \mathrm{X}\hat{\beta} = \mathrm{X}(\mathrm{X}^\mathrm{T}\mathrm{X})^{-1}\mathrm{X}^\mathrm{T}\mathrm{Y} = \mathrm{PY}\ \text{ where } \mathrm{P} = \mathrm{X}(\mathrm{X}^\mathrm{T}\mathrm{X})^{-1}\mathrm{X}^\mathrm{T}$$

The differences $\varepsilon_\mathrm{i} = \mathrm{Y}_\mathrm{i} - \hat{\mathrm{Y}}_\mathrm{i}$ are called residuals and are calculated as

$$\hat{\varepsilon} = \mathrm{Y} - \hat{\mathrm{Y}} = \mathrm{Y} - \mathrm{X}\hat{\beta} = \mathrm{Y} - \mathrm{X}(\mathrm{X}^\mathrm{T}\mathrm{X})^{-1}\mathrm{X}^\mathrm{T}\mathrm{Y} = (\mathrm{I}_\mathrm{n} - \mathrm{P})\mathrm{Y}, \qquad \text{P is called hat matrix.}$$

Hypothesis testing and C.I. for the model parameters

Assuming that the error vector $\varepsilon \sim N_n(0, \sigma^2 \mathrm{I_n})$ we have that

$$\hat{\beta} = (X^T X)^{-1} X^T Y \sim N_n(\beta, \sigma^2 (X^T X)^{-1})$$

and

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \frac{1}{\sigma^2} \left(Y - X\hat{\beta}\right)^T \left(Y - X\hat{\beta}\right) = \frac{1}{\sigma^2} Y^T (I_n - P) Y \sim x^2_{n-p-1}$$

Therefore as unbiased estimator of σ^2 we can define the

$$\widehat{\sigma^2} = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \equiv S^2$$

Also, if we define with $c_{ii}, i = 0,1,\dots,p$ the diagonal elements of the matrix $(X^T X)^{-1}$ then obviously

$$\hat{\beta}_i \sim N(b_i, \sigma^2 c_{ii}), i = 0,1,\dots,p$$

which (and the fact that $\frac{(n-p-1)S^2}{\sigma^2} \sim x^2_{n-p-1}$ and S^2 are independent from $\hat{\beta}_i$) lead as to the conclusion that

$$\frac{\hat{\beta}_i - \beta_i}{S\sqrt{c_{ii}}} \sim t_{n-p-1}, i = 0,1,\dots,p$$

and therefore the C.I. for the $\beta_0, \beta_1, \dots, \beta_p$ with significance level 1-a is

$$\left(\hat{\beta}_i - S\sqrt{c_{ii}}\, t_{n-p-1,\frac{a}{2}}, \hat{\beta}_i + S\sqrt{c_{ii}}\, t_{n-p-1,\frac{a}{2}}\right), i = 0,1,\dots,p$$

Furthermore, for the hypothesis test $H_0\colon \beta_i = 0$ with significance level 1-a will have critical region

$$|T_i| > t_{n-p-1,\frac{a}{2}}, \qquad \text{where } T_i = \frac{\beta_i}{S\sqrt{c_{ii}}}, i = 0,1,\dots,p$$

If for some i reject the null hypothesis then we can say that the dependent variable Y depends on X_i (Freedman, 2009).

Assumptions

Multiple regression model have 5 assumptions.

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto correlation
- Homoscedasticity

Because these five assumptions are very difficult to be met in real data problems and our modeling analysis will be based on computational methods we will not give further information.

2.2 PRESS Statistic

In statistics, the predicted residual sum of squares (PRESS) statistic (Thaddeus, 2000) is a form of cross-validation used in regression analysis to provide a summary measure of the fit of a model to a sample of observations that were not themselves used to estimate the model. It is calculated as the sums of squares of the prediction residuals for those observations. A fitted model having been produced, each observation in turn is removed and the model is refitted using the remaining observations. The out-of-sample predicted value is calculated for the omitted observation in each case, and the PRESS statistic is calculated as the sum of the squares of all the resulting prediction errors

$$\mathrm{PRESS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)}\right)^2$$

Where y_i is the real value of the observation i and $\hat{y}_{(i)}$ is the out-of sample predicted value.

2.3 Bayesian Information Criterion (BIC)

When the number of explanatory variables is large, it can be computationally impossible to fit all 2^p models and choose the best. In these cases we can use optimazitaion algorithms to explore better the space of all the possible models. In addition they have been proposed various methods of selecting variables to find the best regression model separately. As the backward selection which we start with the model that has all the explanatory variables and with consecutives tests we clear every time the variable which is less important based on a criterion or the forward selection which we start from the simplest model and do the opposite work. Also, we can use stepwise selection which apply the previous techniques together whichever is more advantageous. Each algorithm stops when we can't add or remove variable depending on the criterion which has been used. One such criterion is the BIC (Schwarz, 1978), which is often used in Bayesian Statistics to reach the logarithm of the Bayes factor. Penalizes more longest models with more explanatory variables. The price of the BIC criterion for a model (m) is given by the formula

$$\mathrm{BIC(m)} = -2\log L(m) + d_m \log(n)$$

Where L(m) is the value of maximized (with respect to b and σ^2 of the model m) likelihood of the model (m), d_m is the number of the of unknown parameters of the model m and n is the sample size.

2.4 Lasso regression

Multicollinearity in general linear models

A very frequent reason which can create problems at the proper estimation of the model is the phenomenon of multicollinearity. Several times in general linear model is probably one or more independent variables are linearly dependent or there is a strong correlation between two or more explanatory variables. The presence of this phenomenon leads to increased standard errors which consequently makes difficult to estimate the effect of each explanatory variable in dependent variable. In other words its more difficult to identify statistically significant variables. In such cases, where multicollinearity dues to strong correlation between explanatory variables, regression analysis can be performed by subtracting a variable from linearly dependents. Also, sometimes the correlation between variables may not be linear. So we have to remove statistically non significant variables or to select best combination of variables to find the most appropriate model. In short begins the shrinkage of the model. Many techniques have been proposed to address the multicollinearity problem. The general approaches are collecting further data, redefinition of the model and use different estimators except least square estimators. A well known technique to reduce the number of parameters of the model is called LASSO (Tibshirani, 1996) and has a great advantage that we can use it to shrinkage the model and to select variables at the same time.

Lasso regression

The Lasso regression is an approximation of the normalized estimation for regression models which restrict the L_1-norm of the regression coefficients. Denote the data as (x_i, y_i) where $x_i = \left(x_{i1}, \ldots, x_{ip}\right)^T$ and y_i with $i = 1, \ldots, n$ are the values which correspond to the i observation under the general linear model. For the estimation of $\beta = \left(\beta_1, \ldots, \beta_p\right)$ the estimator with the Lasso method is defined as:

$$\left(\hat{\alpha}, \hat{\beta}\right) = argmin\left\{\left(y_i - a - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2\right\}$$

$$\text{subject to } \sum_{\mathrm{j}=1}^{\mathrm{p}} \left|\beta_j\right| \leq \mathrm{t}$$

At the above definition t is a tuning parameter.
Furthermore, for every value of t the solution of the above definition to α is $\hat{\alpha} = \bar{\mathrm{y}}$. We can assume without loss of generality that $\bar{\mathrm{y}} = 0$ and therefore to omit it. Then we have that the estimations of the coefficients for linear model with Lasso method are given by the solution of the system:

$$\hat{\beta} = argmin\left\{\left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2\right\}$$

$$\text{subject to } \sum_{j=1}^{p} \left|\beta_j\right| \leq t$$

Where the tuning parameter $t \geq 0$ controls the shrinkage of the coefficients.

Let say that, $\hat{\beta}_j^0$ are the complete least square estimators and $t_0 = \sum \left|\hat{\beta}_j^0\right|$. The values of s where $t < t_0$ will cause shrinking in solutions near to zero and the values of some coefficients can be made exactly equal to zero. For example if $t = t_0/2$ the result will be the same as the optimal set of variables with size p/2.

Let that the design matrix is orthonormal, namely that $X^T X = I$ where I is the identity matrix. In this case the solutions of the Lasso method will be

$$\hat{\beta}_j^{Lasso} = sign\left(\hat{\beta}_j^0\right)\left(\left|\hat{\beta}_j^0 - \gamma\right|\right)^+$$

where γ is defined by the condition $\sum_{j=1}^{p}\left|\beta_j\right| \leq t$.

Lasso in generalized linear models

Also we can use Lasso regression in the models which is based in likelihood such as generalized linear models (Fan and Li, 2001). Let consider that we have the simple linear model. Primarily we assume that the design matrix X is orthornormal. We will get the least square estimators if we minimize the $\|y - X\beta\|^2$ or equivalent $\left\|\hat{\beta} - \beta\right\|^2$, where $\hat{\beta}$ the usual least square estimators. Let that $z = X^T y$ and $\hat{y} = XX^T y = Xz$. So if want to to find the Lasso estimators in the simple linear model

$$minimize_\lambda \left\{\frac{1}{2}\|y - X\beta\|^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right|\right\}$$

or

$$minimize_\lambda \left\{\frac{1}{2}\|y - \hat{y}\|^2 + \frac{1}{2}\sum_{j=1}^{p}(z_j - \beta_j)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right|\right\}$$

Where ||.|| is the L^2 norm and λ is a tuning parameter.

We can find the solution of the above system using quadratic programming methods but this is quite complex and time consuming. For this reason we use cross-validation methods (Tibshirani, 1996). So we fit Lasso model for different values of λ and choose the value of λ which give us the lowest value of $MSE_{CV(1)}$ or the λ which is one standard deviation away from the λ with the minimum $MSE_{CV(1)}$ (gives more parsimonious models). The $MSE_{CV(1)}$ is defined as:

$$MSE_{CV(1)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2$$

Where y_i is the observed value of the observation i and $\hat{y}_{(i)}$ is the out-of sample predicted value.

2.5 Classification and Regression Trees (CART)

When our data is quite complex or have many categorical variables with a large number of levels or there is no evidence of correlation between the response variable and the explanatory variables, to fit a multiple regression model, it will not give us a model with high predictive power. So it would be useful to separate our data in more homogeneity fields so as to draw as much information as we can from them. This separation is done with the help of a Classification and Regression Trees (CART) algorithm (Breiman, 1984). In a few words a CART algorithm is an iterative process that splits the data into two subsets, so that data within each of the subsets to be more uniform than was the original set. Then the process is repeated in each of the subsets until a homogeneity criterion or another stop criterion is reached.

Algorithm Construction

The CART algorithm (Breiman, 1984) separates each node in such a way that the child node which occurs every time is more pure (based on a non-impurity measure) compared to the parent node. In a clean node all the entries have the same value in the target field. Starting from the initial node the steps are:

- For numeric fields order the values of the field in the node from the smallest to largest. Choose each item in turn, as a point of separation and calculate the non-impurity statistic for child nodes which are coming from the separation. The point which reduces most the non-impurity, compared to the previous node from which has occurred, is the best division point for this node
- For categorical fields examined every possible separation of values into two subsets. For each possible separation calculate the non-impurity statistic for child nodes that are coming from the separation. Choose as best separation point for this field the one which reduces most the non-impurity compared to the node which has been occurred.
- Find the best separation for the node. Define the field which separation gives the largest reduction in non-impurity for the node and choose this separation totally for the node.
- Check if a stop criterion is satisfied. If the initial node or the separation does not satisfy a stop criterion then repeat separation to create two child nodes and reiterate the algorithm to each of them.

Gaps-Missing Values

If the prediction field which is used for separation presents a gap or a missing value at a particular node, then another field which yields similar separation under the specific node is used in place of the field in the gap or missing value. The value of the replacement field is used to assign the entry to one of the child nodes. For speed reasons and saving memory, only a limited number of replacements fields is defined for each separation which takes place on the tree. If an entry has missing values in the separation field and at all replacements fields, then is used the child node with the highest weighted probability (Breiman, 1984). The weighted probability calculated by the formula:

$$\frac{N_{f,j}(t)}{\mathrm{N_f(t)}}$$

where $N_{f,j}(t)$ is the sum of the frequency weights for the entries of category j for the node t and $\mathrm{N_f(t)}$ is the sum of the frequency weights for all entries in the node t .

Process Stopping Rules

The separation process of the nodes in the tree stops when satisfied one of the following stopping rules (Breiman, 1984):

- All entries have the same value for the target field (the node is pure).
- All entries in the node have the same value for all the predictor fields which are used in the model.
- The depth of the tree for the current node is the maximum depth which has predetermined. The current node is determined by the number of consecutive separated nodes.
- The number of entries in the node is less than the minimum size of the parent node, as has fixed in advance.
- The number of entries in any of the child nodes, which resulting from the best node split is less than the minimum size of the child node which has predetermined.
- The best separation for the node yields a reduction in non-impurity which is less than the minimum non-impurity as has fixed in advance.

Non-impurity Measures

Depending on the type of the target field, we can use three different measures for the non-impurity at CART models (Breiman, 1984). If fields have symbolic target we can use Gini or Twoing index. For continuous fields target we use Least Square Deviation method.

The Gini Index

The Gini index g(t) to a node t of a CART tree, is defined as

$$g(t) = \sum_{i \neq j} p(j|t)p(i|t) \quad \text{or} \quad g(t) = 1 - \sum_{j} p^2(j|t)$$

Where i and j are categories in the target field and

$$p(j|t) = \frac{p(j,t)}{p(t)}\,, \qquad p(j,t) = \frac{\pi(j)N_j(t)}{N_j}, \quad p(t) = \sum_{j} p(j,t)$$

$\pi(j)$ is the prior probability for the category j, $N_j(t)$ is the number of entries in the category j of the t node and N_j is the number of category j entries in the initial root node. If we use the Gini index to find the improvement of the separation in the development of a tree we use only the entries in the node t and in the initial root node, which have valid values for the partition field to calculate $N_j(t)$ and N_j respectively. When the entries in a node evenly distributed across the categories, the Gini index takes the largest value of $1-\frac{1}{k}$ with k being the number of categories for the field target. If all entries in a node belongs in the same category, then the Gini index is g(t)=0.

The function of the Gini criterion for the separation s to node t, defined as:

$$\Phi(\mathrm{s},\mathrm{t}) = \mathrm{g}(\mathrm{t}) - \mathrm{p_L}\mathrm{g}(\mathrm{t_L}) - \mathrm{p_R}\mathrm{g}(\mathrm{t_R})$$

Where $\mathrm{p_L}$ is the total entries in node t which are sent to the left child node and respective $\mathrm{p_R}$ total entries which are sent to the right child node. We define as:

$$\mathrm{p_L} = \frac{\mathrm{p}(\mathrm{t_L})}{\mathrm{p}(\mathrm{t})} \quad \kappa\alpha\iota \quad \mathrm{p_R} = \frac{\mathrm{p}(\mathrm{t_R})}{\mathrm{p}(\mathrm{t})}$$

And choose the s which maximizes the function Φ(s,t).

The Twoing Index

The Twoing Index is based on the separation of the categories in two hyperclasses and finding the best separation in prediction field from these hyperclasses. Define the hyperclass C_1, C_2 as:

$$C_1 = \{j: p(j|t_L) \geq p(j|t_R)\}$$

$$C_2 = C - C_1$$

with C are all the categories of the target field and $p(j|t_L), p(j|t_R)$ is identical to p(j|t) which is deifined fot the Gini index for the left and right child nodes respectively. Defined as a function of Twoing criterion for the separation s to node t as follows:

$$\Phi(s,t) = p_L p_R \left[\sum_j |p(j|t_L) - p(j|t_R)|\right]^2$$

with t_L, t_R be the nodes which are created by the separation s which is selected to maximize the above criterion.

Least Squared Deviation Measure

The LSD measure for non-impurity is used when we have continuous target fields. Denoted by R(t) and is the weighted variation in node t

$$\mathrm{R}(\mathrm{t}) = \frac{1}{\mathrm{N_w}(\mathrm{t})}\sum_{\mathrm{i}\in\mathrm{t}} \mathrm{w_i}\mathrm{f_i}\left(\mathrm{y_i} - \overline{\mathrm{y}(\mathrm{t})}\right)^2$$

where $N_w(t)$ is the weighted number of entries in the node t, w_i is the value of the weighted field for any entry i, f_i is the frequency field value, y_i is the value of the target field and $\overline{y(t)}$ is the weighted average for node t. The function of LSD criterion for separating the node t is:

$$\Phi(s,t) = R(t) - p_L R(t_L) - p_R R(t_R)$$

where p_L all the entries in node t which are sent to the left child node and respective p_R all the entries which are sent to the right child node. Separation s is chosen such as to maximize the function Φ(s,t).

Pruning Tree

There are many ways to choose the right size of a tree but the simplest is to use cross-validation methods (Timofeev, 2004). The procedure of cross validation is based on optimal proportion between the complexity of the tree and misclassification error. With the increase in size of the tree, misclassification error is decreasing and in case of maximum tree, misclassification error can be equal to 0. But on the other hand, complex decision trees poorly perform on independent data. Performance of decision tree on independent data is called true predictive power of the tree. Therefore, the main task is to find the optimal proportion between the tree complexity and misclassification error. This task is achieved through cost-complexity function:

$$R_a(T) = R(T) + a(\overline{T}) \rightarrow \mathrm{min}_T$$

where R(T) is misclassification error of the tree T and $\alpha(\overline{T})$ is complexity measure which depends on $\overline{T}$ which is the total sum of terminal nodes in the tree and α is the parameter is found through the sequence of in-sample testing when a part of learning sample is used to build the tree, the other part of the data is taken as a testing sample. The process repeated several times for randomly selected learning and testing samples. Although cross-validation does not require adjustment of any parameters, this process is time consuming since the sequence of trees is constructed. Because the testing and learning sample are chosen randomly, the final tree may be deferent from time to time.

Variable Importance in CART

The variable importance in regression trees (Ishwaran, 2007) is defined as the sum of changes in mean square error (MSE) due to the separation of each explanatory variable, divide it with the sum of number of branch nodes.
Where MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2$$

In contrast, in the classification trees is defined as the sum of the risk due to the separation of each explanatory variable by dividing it with the sum of the number of branch nodes.
Where risk r(t) is:

$$r(t) = \frac{1}{N_f}\sum_{j} N_{f,j}(t)C(j^*(t)|j)$$

Where $C(j^*(t)|j)$ is the cost to be classified an entry which has target field j as $j^*(t)$, $N_{f,j}(t)$ is the sum of the frequency weights for the entries of category j for the node t and $N_f(t)$ is the sum of the frequency weights for all entries in the node t.

Advantages and disadvantages of CART

CART models have many advantages such as:

- CART is nonparametric
- CART does not require variables to be selected in advance
- CART results are invariant to monotone transformations of its independent variables
- CART can easily handle outliers.

but have also a great disadvantage that splits only by one variable.

2.6 Cross-Validation

Cross validation (Kohavi, 1995) is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. The *holdout method* is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The model fitted using the training set only. Then the model is asked to predict the output values for the data in the testing set (it has never seen these output values before). The errors it makes are accumulated to give the mean absolute test set error (or other criterion), which is used to evaluate the model. The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute. However, its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the test set, and thus the evaluation may be significantly different depending on how the division is made.

K-fold cross validation is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Then the average error (or other criterion) across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over. *Leave-one-out cross validation* is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. That means that N separate times, the model is trained on all the data except for one point and a prediction is made for that point. As before the average error (or other criterion) is computed and used to evaluate the model.

CHAPTER 3

DESCRIPTIVE ANALYSIS

In this chapter we will use descriptive statistics to describe the categorical and quantitative variables of our data.

3.1 Descriptive Analysis of the Categorical Variables

First of all, it would be interesting to see the type of products which are more and less produced in the company. From Figure 3.1.1 we notice that the products with the highest percentage of production are the COILS FOILSTOCK, COILS PAINTED R/SHUT and HI-MG whereas these with the lowest percentage SHEETS 5XXX, AUTOMOTIVE, FOOD SCROLL and AEROSOL.

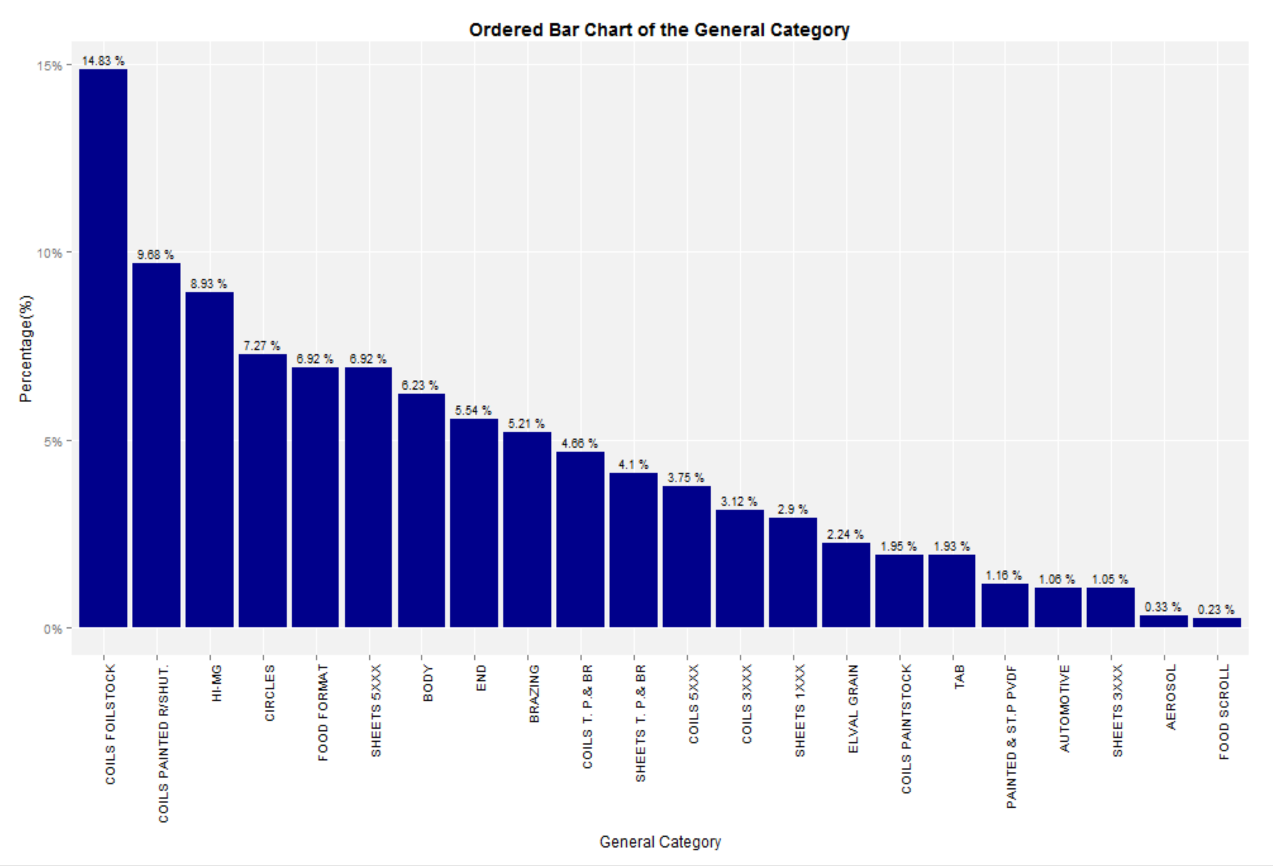


Figure 3.1.1: Ordered Bar Chart of the General Category

However, it would be more important to see figures which are referred to the analysis of categorical variables when there is a DMSY report, in order to draw a conclusion about the problems which are appeared in the production process. From Figure 3.1.2 we notice that 25.57% of the available data appear DMSY report whereas 74.73% does not have any production error.

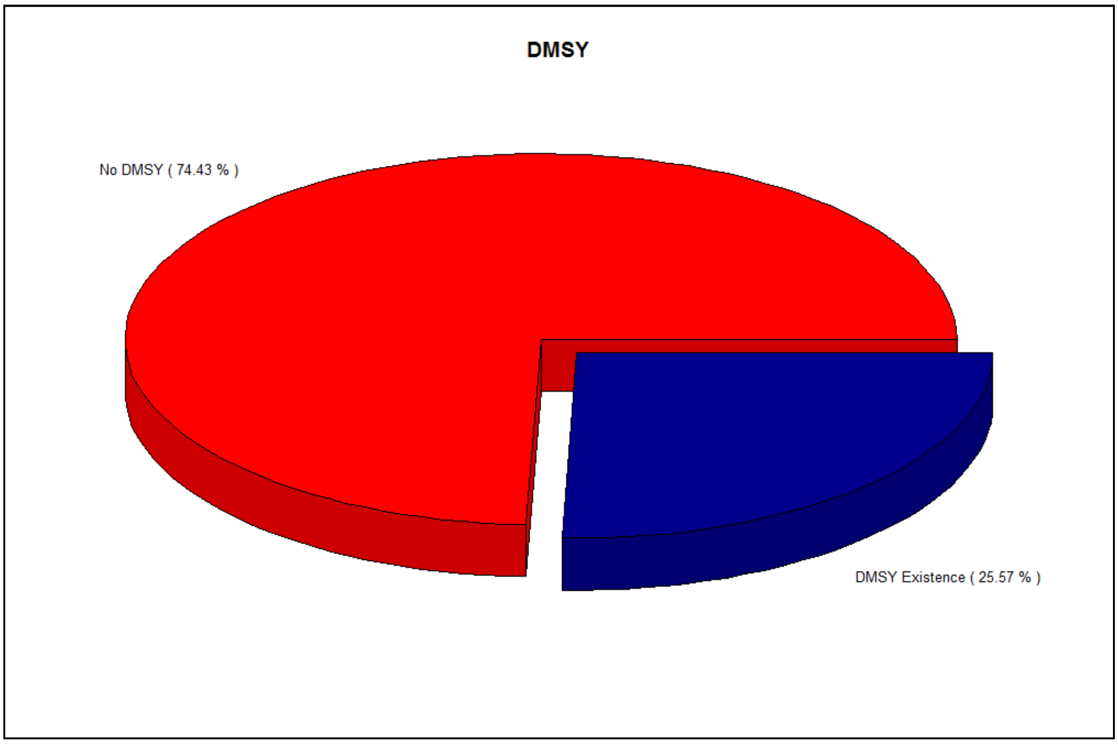


Figure 3.1.2: Pie Chart of the variable DMSY

For the purpose of getting through where we have the largest frequency existence of DMSY in each categorical variables and take the right decisions we will use Pareto charts. A Pareto chart is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line. The left vertical axis is the frequency of occurrence and the right vertical axis is the cumulative percentage of the total number of occurrences. The goal of the Pareto chart is to highlight the most important among a (typically large) set of factors.

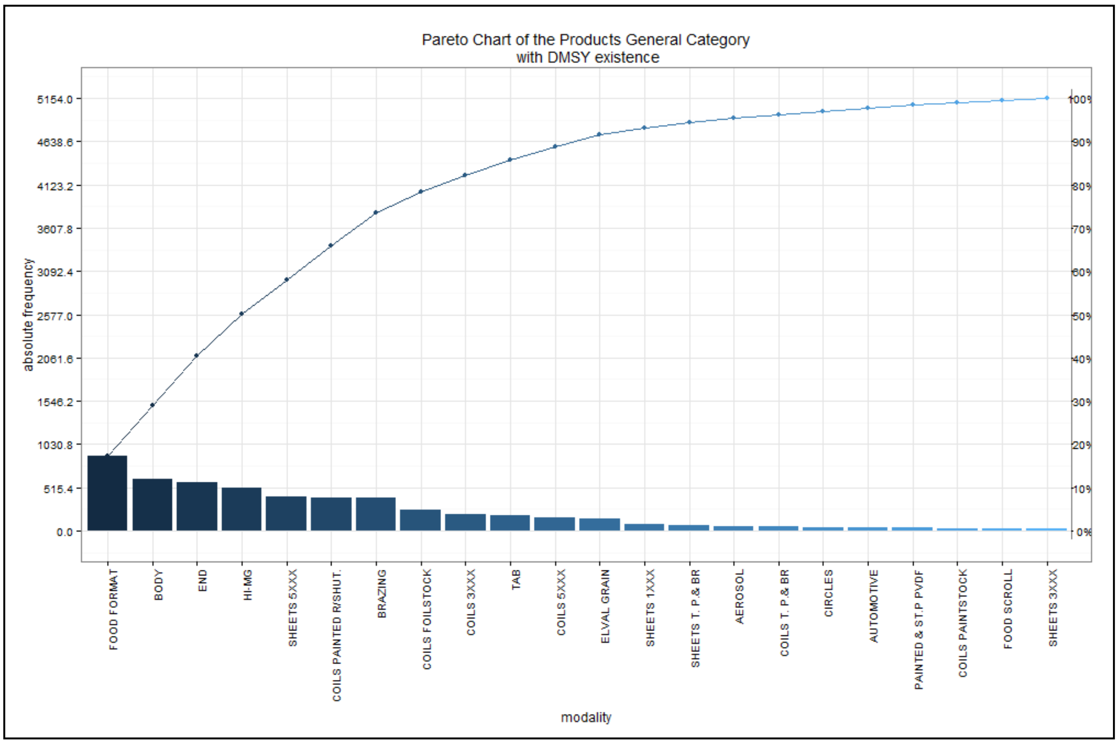


Figure 3.1.3: Pareto Chart of the Products General Category with DMSY existence

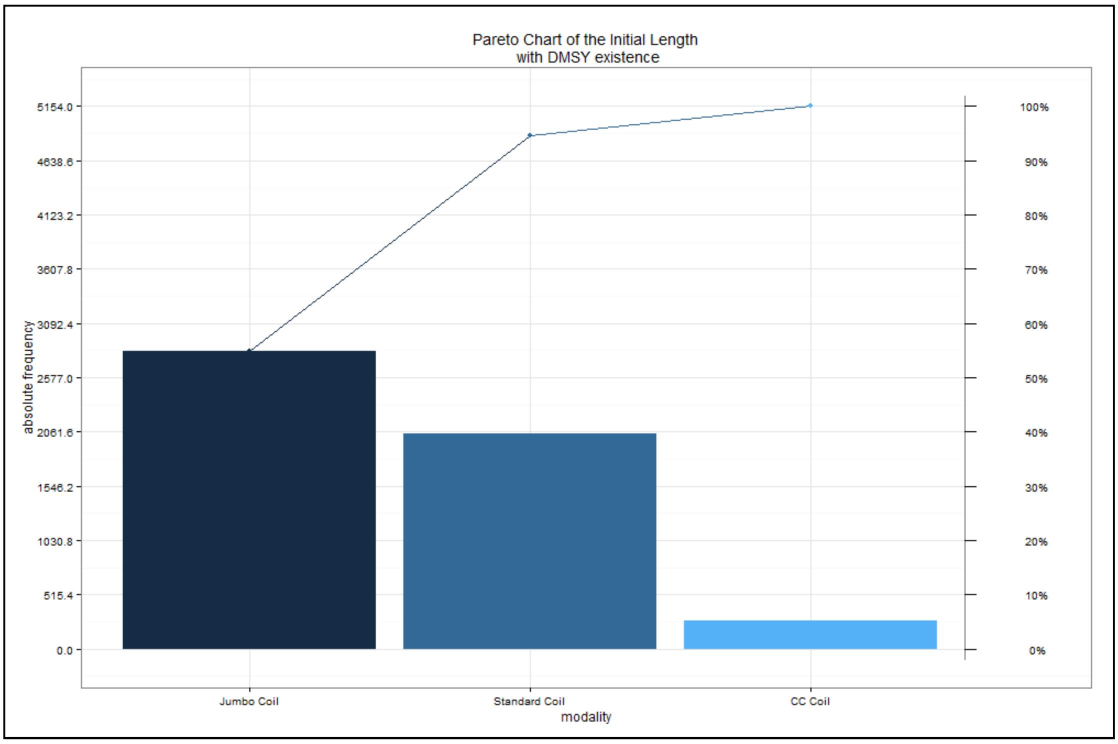


Figure 3.1.4: Pareto Chart of the Initial Length with DMSY existence

From Figure 3.1.3 which is illustrated a pareto chart for the products General Category with DMSY existence we notice that the products FOOD FORMAT (17.23%), BODY (11.83%), END (11.31%), HI-MG (9.82%) and SHEETS 5XXX (7.92%) have the highest frequency levels of DMSY existence in contrast with the AUTOMOTIVE (0.7%), PAINTED & ST.P PVDF (0.66%), COILS PAINSTOCK (0.6%), FOOD SCROLL (0.49%) and SHEETS 3XXX (0.48%). Furthermore, from Figure 3.1.4 we can understand that the slabs which end up to Jumpo Coils have higher DMSY percentage (54.9%) than these which end up to Standard Coils (39.8%) and CC Coils (5.3%).

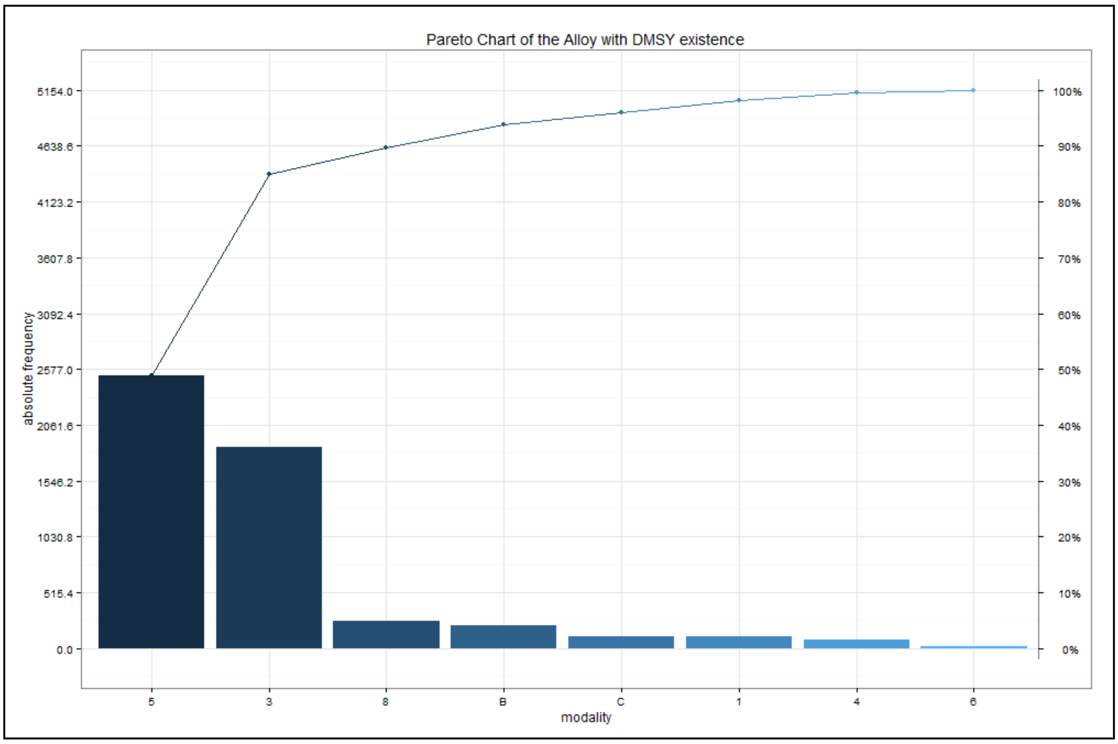


Figure 3.1.5: Pareto Chart of the Alloy with DMSY existence

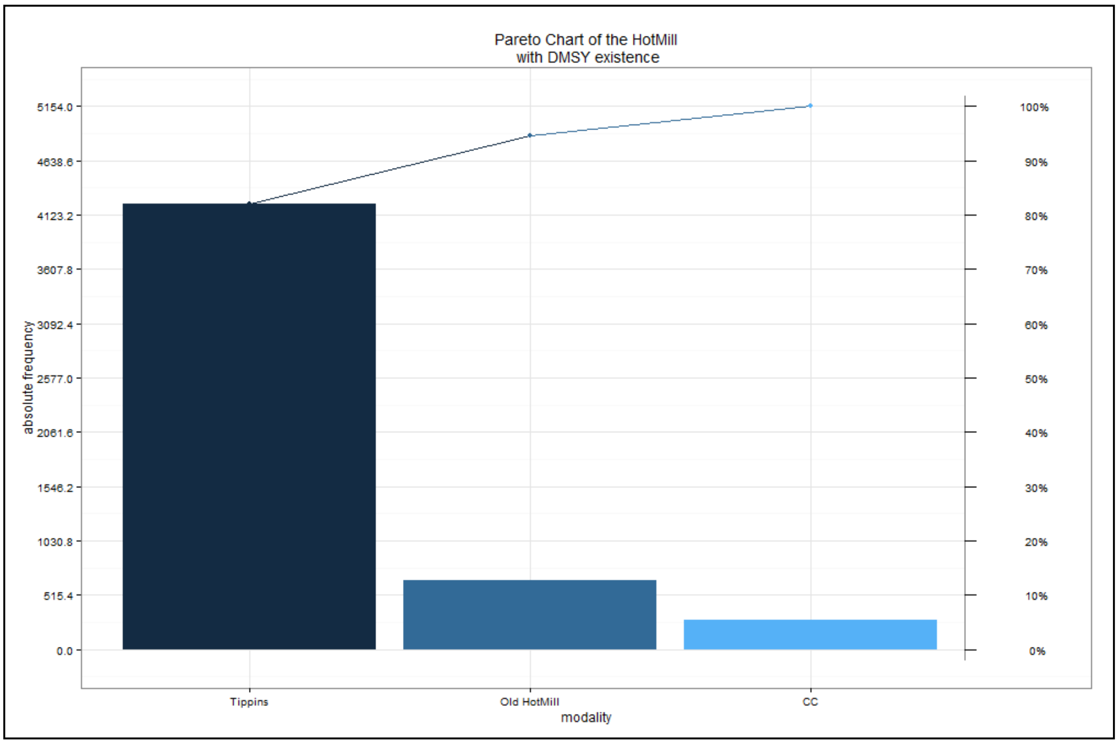


Figure 3.1.6: Pareto Chart of the HotMill with DMSY existence

From Figure 3.1.5 we can draw the conclusion that the Alloys which belong to the series 5xxx (48.91%) and 3xxx (36%) have blatantly higher percentages of DMSY existence relative to these which belong to series 1xxx (2.08%), 4xxx (1.51%) and 6xxx (0.35%). Also the Continuous Casting (5.4%) has lower percentage of DMSY existence relative to the Tippins (81.9%) and Old HotMill (12.7%) as it is presented in the Figure 3.1.6.

Furthermore, is very important to understand the reasons why each product category has greater DMSY existence. Because of the Reason_ID is a factor variable with 117 levels, only the first 15 will be selected from a Pareto Chart to have a more readable figure. So, from Figure 3.1.7 we notice that the reasons with code 154 and 24 appear with higher percentages in the product SHEETS 5XXX, the reason with code 25 in the FOOD FORMAT, HI-MG and BODY and the reason with code 403 in the END and BODY. Also, the reason with code 73 appears more in the product BRAZING and the reason 64 in the FOOD FORMAT and END. A very important conclusion is that in the product FOOD FORMAT there are the most appearances of DMSY with the most reasons.

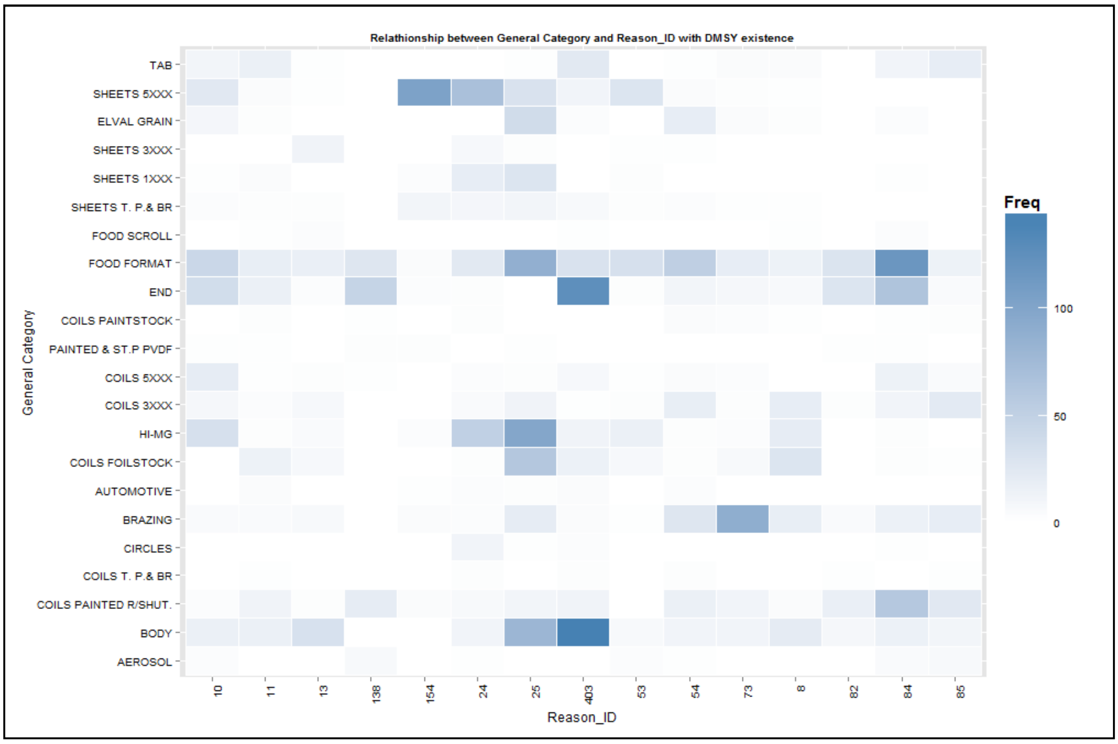


Figure 3.1.7: Relationship between General Category and Reason_ID with DMSY existence

Finally, it would be useful to understand where more the DMSY reports are appeared during the Hot Rolling process relative to the variable General Category. The figure below was made only from Hot Rolling data because the process of Hot Rolling has different steps from Continuous Casting. From Figure 3.1.8 we notice that we have higher percentages of DMSY existence at the step of HOT ROLLING for the product BODY and at the step of PAINTING LINE1 for the FOOD FORMAT. Also, at the step of SLLITERING more DMSY reports appear in the FOOD FORMAT, BODY, END, COILS PAINTED R/SHUT and at the CUTTING TO LENGTH MACHINE the products SHEETS 5XXX and HI-MG have the highest percentages of DMSY existence.

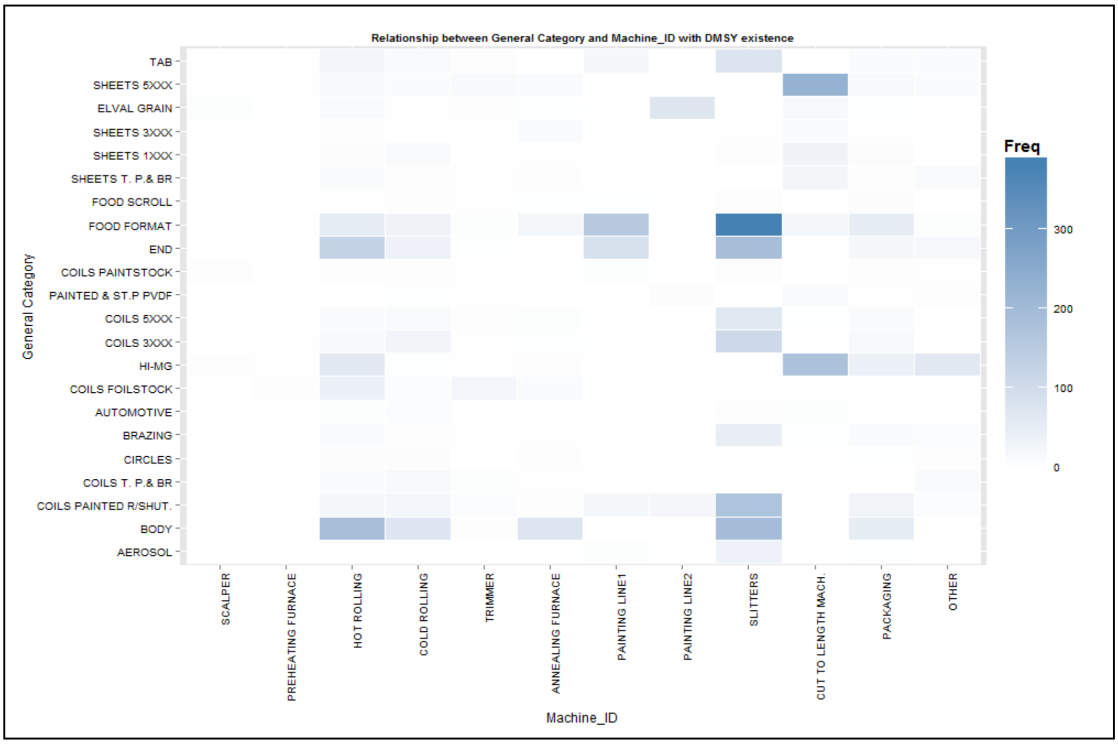


Figure 3.1.8: Relationship between General Category and Machine_ID with DMSY existence in Hot Rolling

3.2 Descriptive Analysis of the Quantitative Variables

The quantitative variable which is more important in our analysis is the Return Index which describes the initial weight divided by the produced. Firstly, let's take a look at some characteristics of the Return's Index distributions for each product category.

Skewness and Kurtosis are very important characteristics of distributions. In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real valued random variable about its mean. The skewness value can be positive or negative, or even undefined when

$$(E[(X-\mu)^2])^{3/2} = 0$$

Negative skew indicates that the tail on the left side of the probability density function is longer or fatter than the right side, positive skew indicates that the tail on the right side is longer or fatter than the left side and a zero value indicates that the tails on both sides of the mean balance out, which is the case for a symmetric distribution. Kurtosis is a measure of the tailedness of the probability distribution of a real valued random variable. The kurtosis of any univariate normal distribution is 3 and it is common to compare the kurtosis of a distribution to this value. Distribution with kurtosis less than 3 are said to be platykurtic and with kurtosis greater than 3 are said to be leptokurtic.

From Figure 3.2.1 we notice that the products Return Index with the highest Standard Deviation are the BRAZING, SHEETS 5XX and AEROSOL whereas the BODY, COILS FOILSTOCK and COILS T.P & BR have the lowest values. Also, from figure 3.2.2 which represents the Return Index kurtosis for each product category we gather that all products have positive skewness and especially the COILS T.P&BR, BODY and ELVAL GRAIN have the highest values which means that is more likely to have higher outlier values(right fat tails). Furthermore, from Figure 3.2.3 we understand that most products Return Index distribution are leptokurtic because have kurtosis greater than 3. The Return's Index distributions of AEROSOL and CIRCLES look like to approach the Normal because have kurtosis near to 3 whereas AUTOMOTIVE, BRAZING and FOOD SCROLL with kurtosis less than 3 are platykurtic.

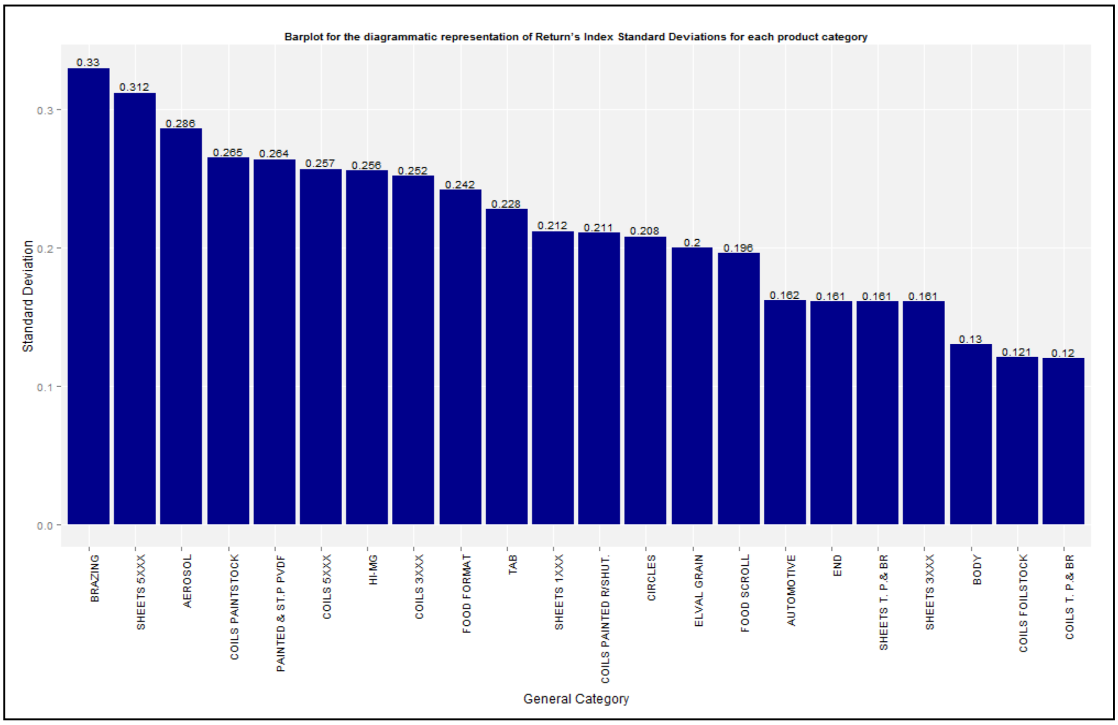


Figure 3.2.1: Barplot for the diagrammatic representation of Return's Index Standard Deviations for each product category

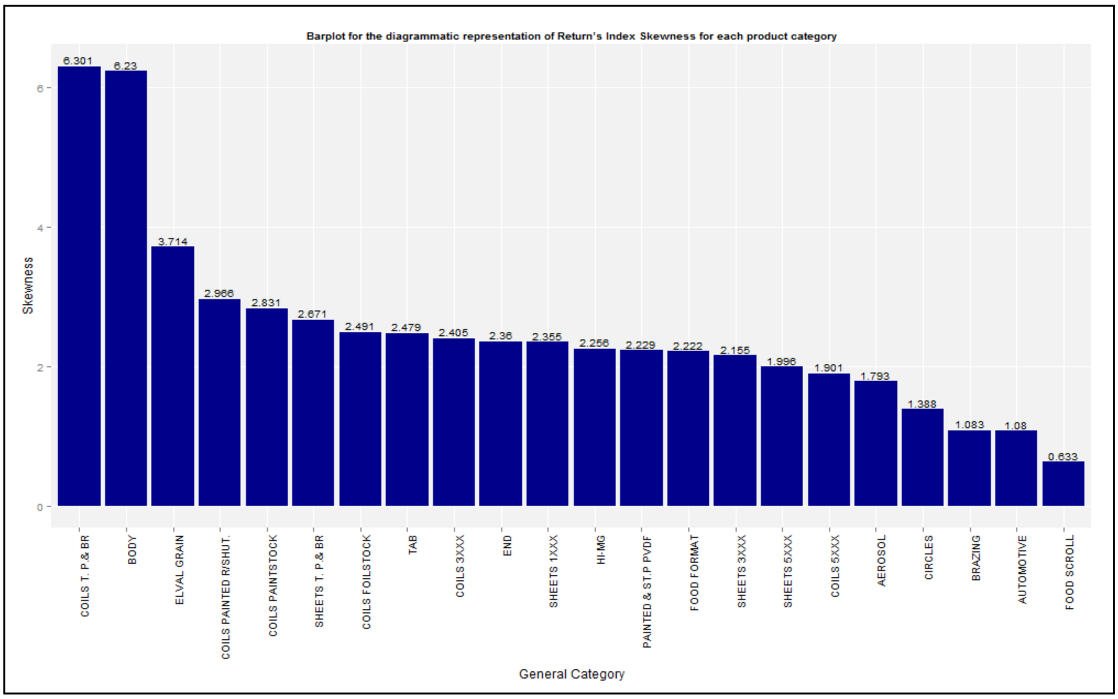


Figure 3.2.2: Barplot for the diagrammatic representation of Return's Index Skewness for each product category

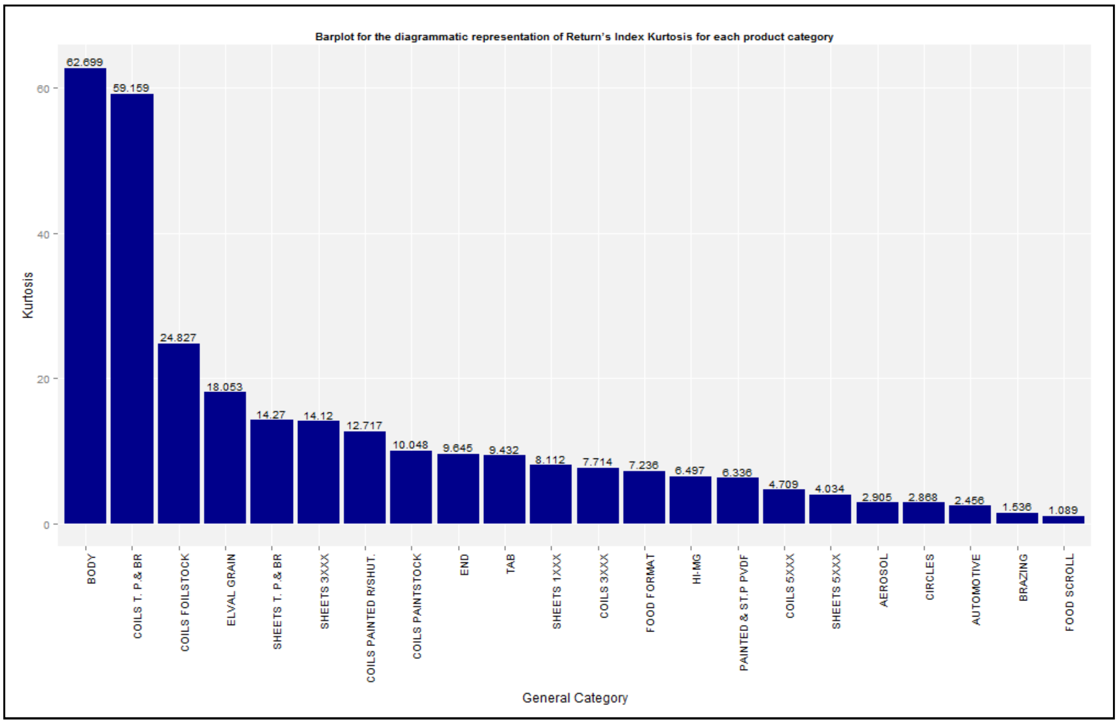


Figure 3.2.3: Barplot for the diagrammatic representation of Return's Index Kurtosis for each product category

From Figure 3.2.4 in which is represented with Box Plots the Return Index for each product category from the one with the lowest median to the one with the higher median we notice that the products COILS FOILSTOCK, COILS T.P&BR, COILS PAINSTOCK and BODY take the lowest values of the Return Index and the FOOD SCROLL, BRAZING and CIRCLES the highest. Also, is obvious that the product with DMSY existence take higher values of the Return Index relative to these which doesn't exist. The most significant differences were located in the products SHEETS T.P&BR, COILS 3XXX, PAINTED & ST.P PVDF, SHEETS 5XXX, COILS 5XXX and CIRCLES as we can understand from the Figure 3.2.5. Furthermore, from Figure 1.3.6 we can draw the conclusion that, for the initial length which ends up to CC Coils we have lower values of Return Index relative to these which end up to Jumbo and Standard Coils.

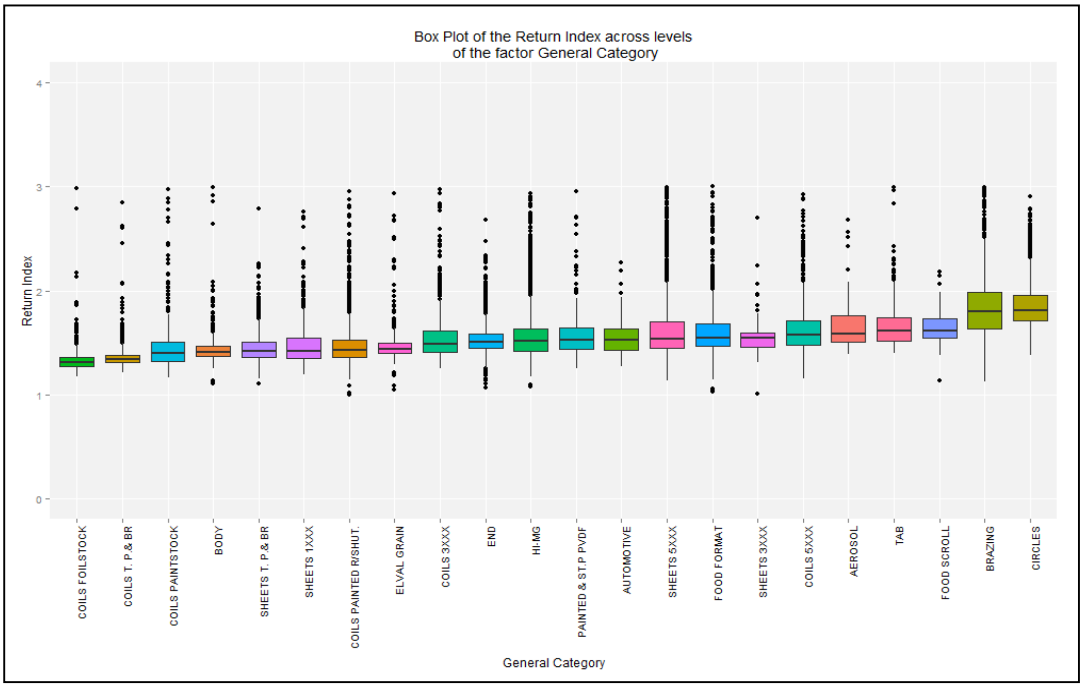


Figure 3.2.4: Figure of the Return Index across levels of the factor General Category

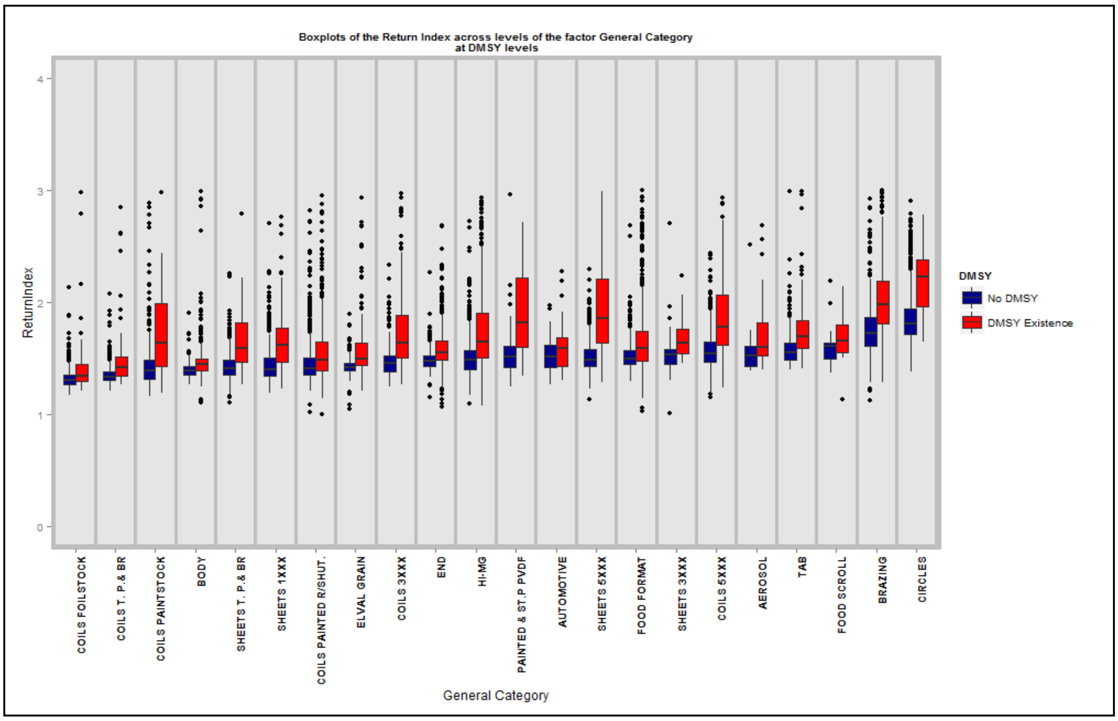


Figure 3.2.5: Boxplots of the Return Index across levels of the factor General Category at DMSY's levels.

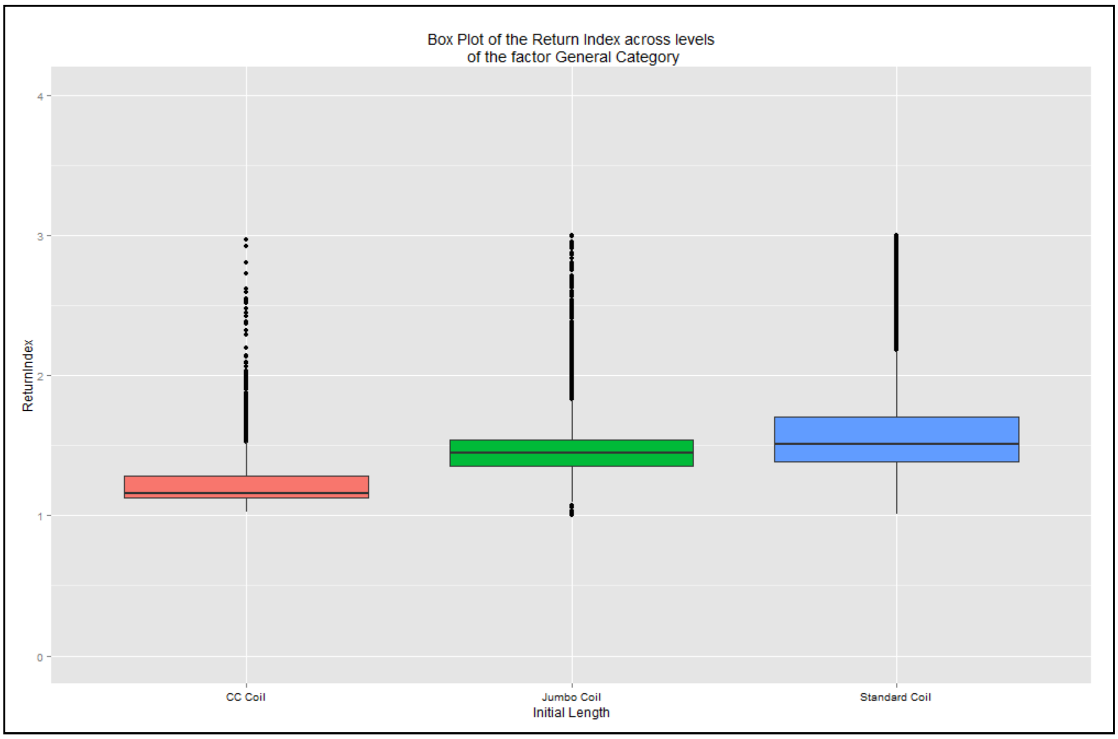


Figure 1.3.6: Boxplots of the Return Index across levels of the factor Initial Length

Finally, is very important to analyze in the process of Hot Rolling if there is relationship firstly between the Return Index and Scalper Return Index and secondly between the Return Index and Hot Mill Return Index, as strong positive correlations can lead us to expect higher Return Index while strong negative the opposite conclusion.

The coefficients used were the Pearson and Spearman. The Pearson correlation coefficient ($r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}}$) measures the linear relationship between two variables and takes values between -1 and 1. If it gets a value equal to 1 we have a perfect positive linear relationship, -1 a perfect negative linear relationship and 0 there is no linear relationship between those variables. The Spearman correlation coefficient $\left(\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, d_i = x_i - y_i\right)$ is a non-parametric measure of statistical dependence between two variables using a monotonic function even if it is not linear. If there are no repeated data values a perfect Spearman correlation by 1 or -1 is the case where each of the variables is a perfectly monotonic function of the other.

The highest positive Pearson correlation coefficients between Return Index and Scalper Return Index appear in the products CIRCLES and SHEETS 5XXX and PAINTED &ST.P PVDF which are statistically significant whereas this with the highest negative Pearson coefficient is the FOOD SCROLL which is statistically significant as we can understand from the Figure 3.2.7. From Figure 3.2.8 we notice that for the Spearman correlation coefficients between Return Index and Scalper Return Index we have the same results as in Pearson's. Also, from Figure 3.2.9 which is represented the Pearson correlation coefficient between Return Index and Hot Mill Return Index we draw the conclusion that the products with the highest positive correlation coefficients are CIRCLES, AUTOMOTIVE and AEROSOL which are statistically significant whereas don't exist products with strong negative correlations. Furthermore, the products with the highest positive Spearman correlation coefficients are the AUTOMOTIVE, BODY, PAINTED &ST.P PVDF, CIRCLES and COILS FOILSTOCK which are statistically significant and these with the highest negative Spearman coefficient is SHEETS 3XXX which is statistically significant as we can understand from Figure 3.2.10. Moreover, it's very important to represent some relationships firstly between the Return Index and Scalper Return Index and secondly between the Return Index and Hot Mill Return Index for specific product categories (which have strong correlations) from scatter plots with smooth fitted line. In the first case will refer to the products CIRCLES (Figure 3.2.11), SHEETS 5XXX (Figure 3.2.12) and PAINTED &ST.P PVDF (Figure 3.2.13) and in the second case to the products CIRCLES (Figure 3.2.14), PAINTED &ST.P PVDF (Figure 3.2.15) and COILS FOILSTOCK (Figure 3.2.16).

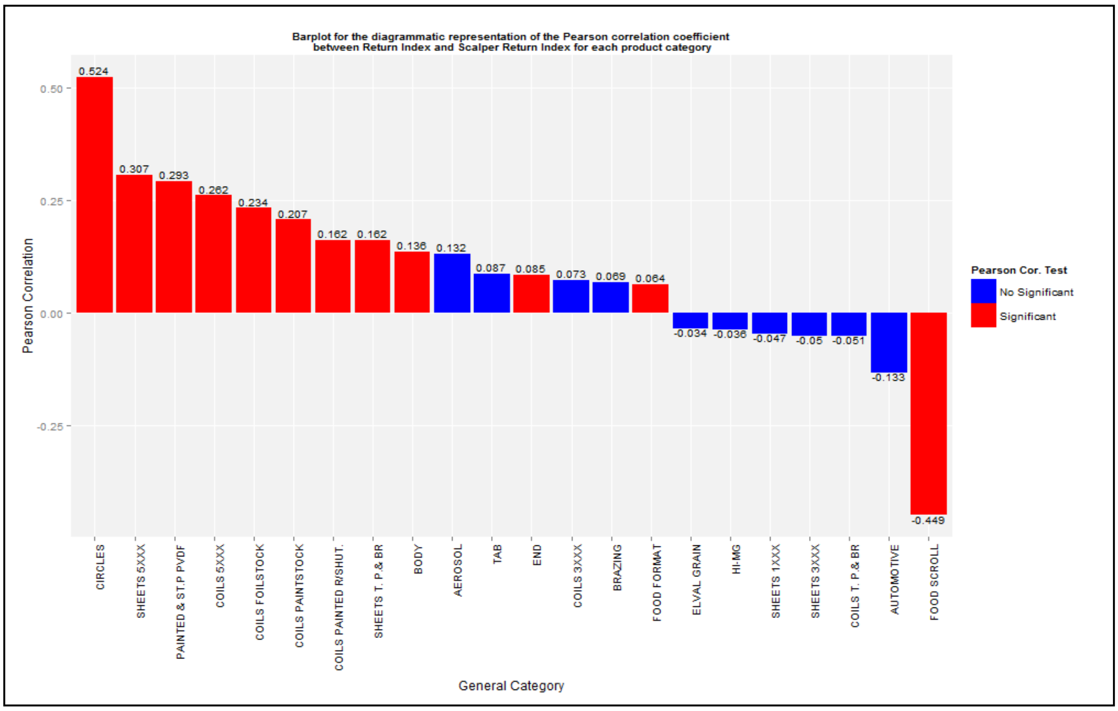


Figure 3.2.7: Barplot for the diagrammatic representation of the Pearson correlation coefficient between Return Index and Scalper Return Index for each product category.

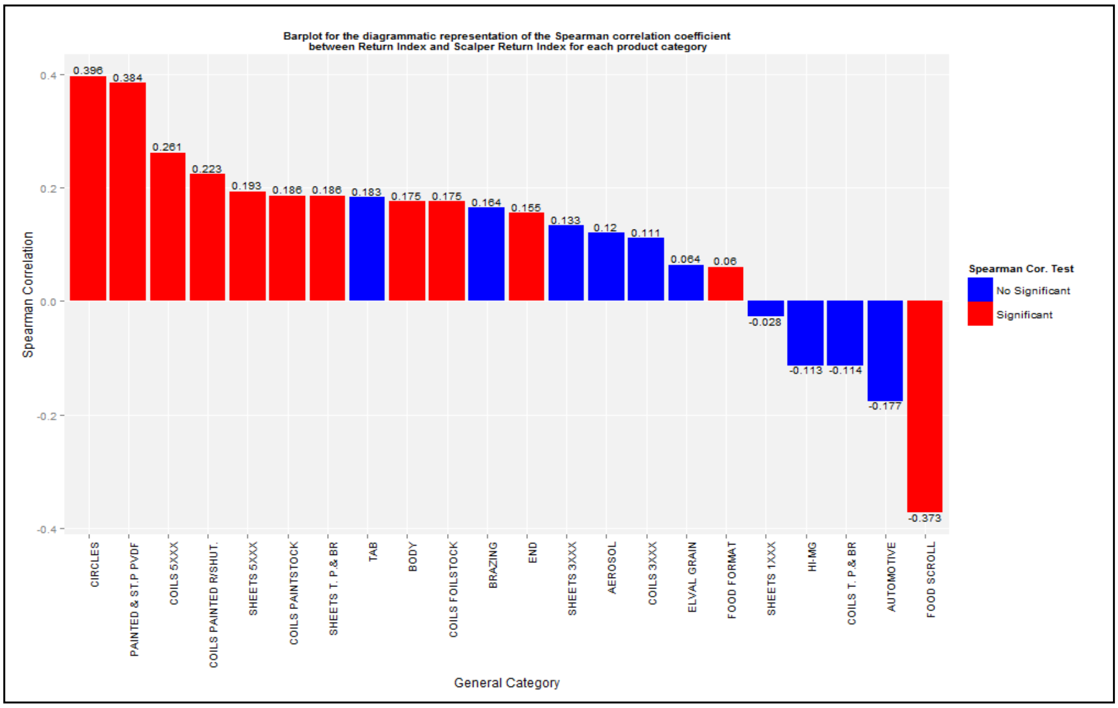


Figure 3.2.8: Barplot for the diagrammatic representation of the Spearman correlation coefficient between Return Index and Scalper Return Index for each product category.

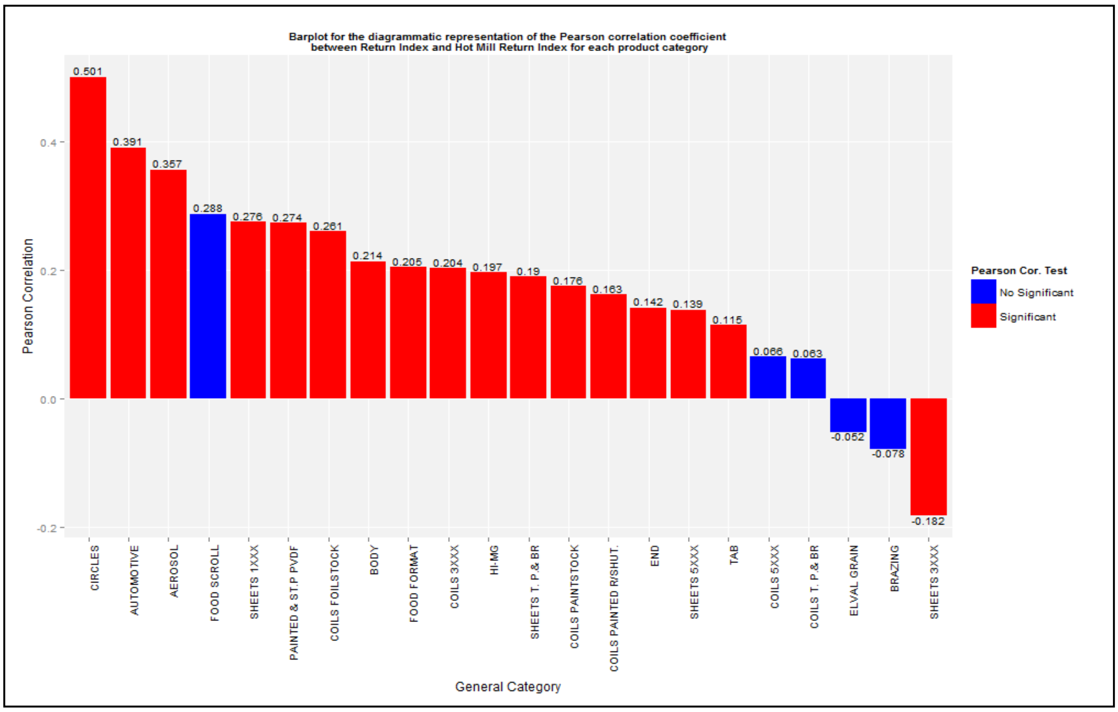


Figure 3.2.9: Barplot for the diagrammatic representation of the Pearson correlation coefficient between Return Index and Hot Mill Return Index for each product category.

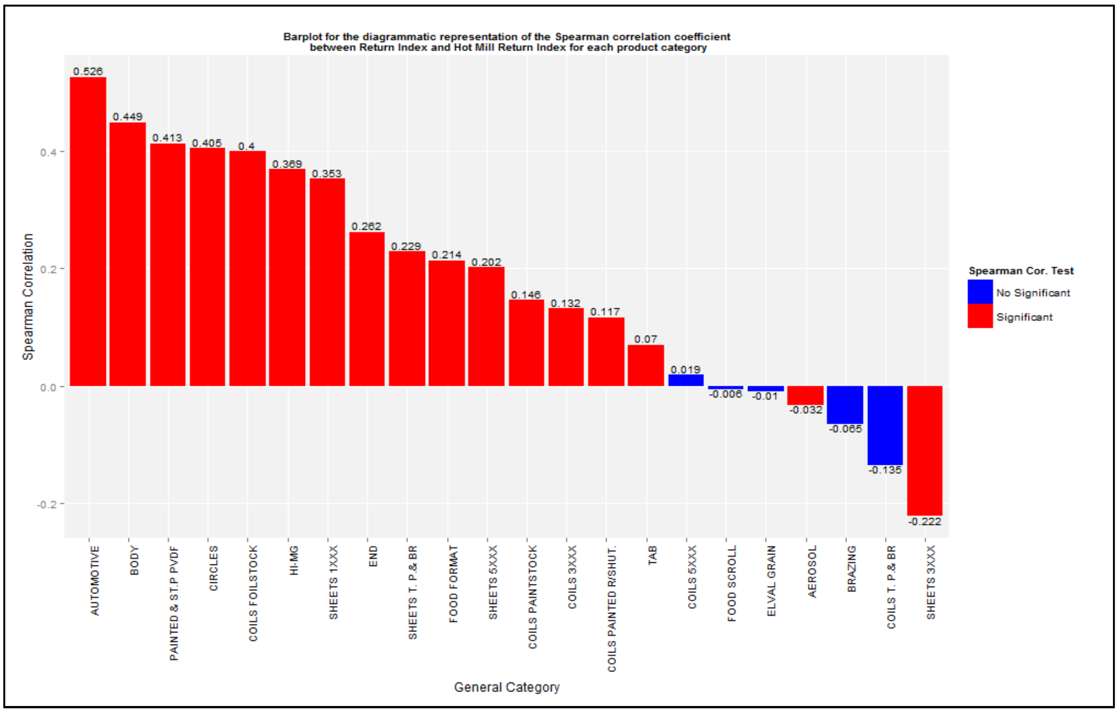


Figure 3.2.10: Barplot for the diagrammatic representation of the Spearman correlation coefficient between Return Index and Hot Mill Return Index for each product category.

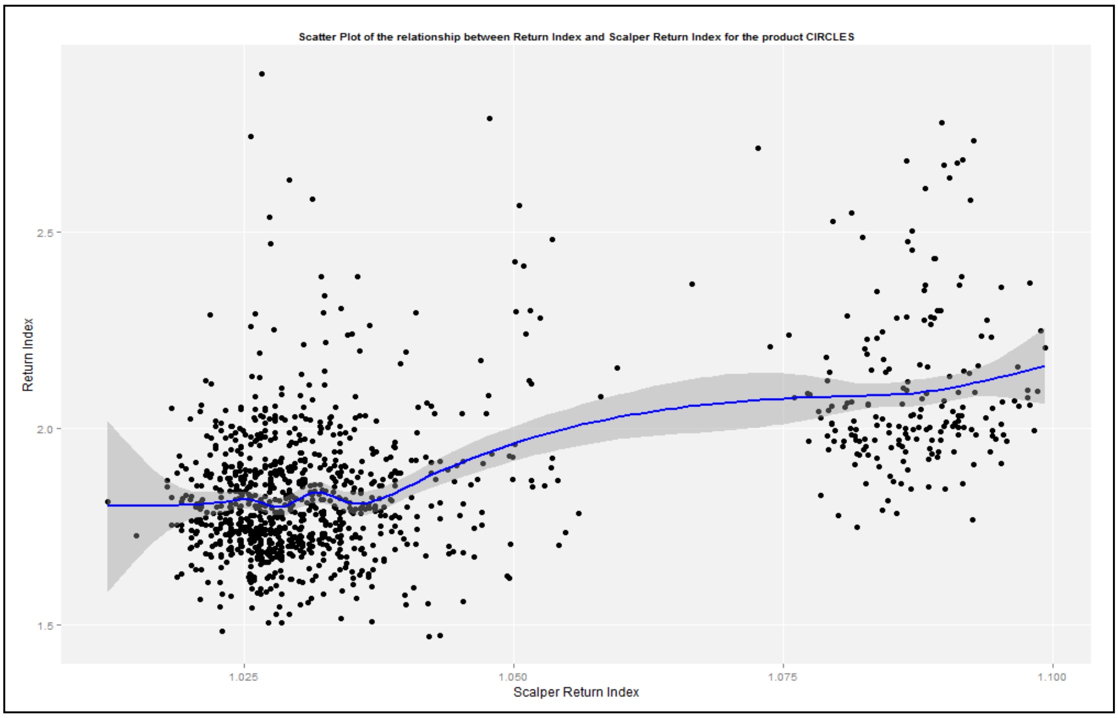


Figure 3.2.11: Scatter Plot with smooth line and 95% pointwise confidence interval of the relationship between Return Index and Scalper Return Index for the product CIRCLES

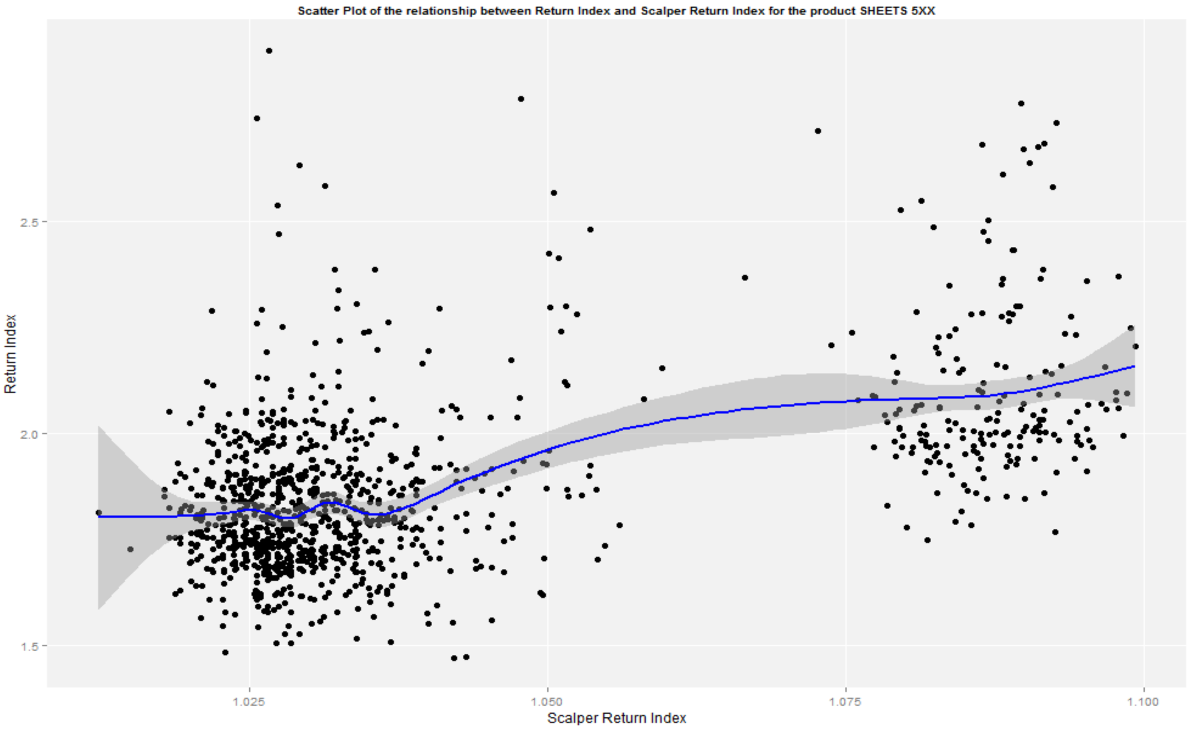


Figure 3.2.12: Scatter Plot with smooth line and 95% pointwise confidence interval of the relationship between Return Index and Scalper Return Index for the product SHEETS 5XX

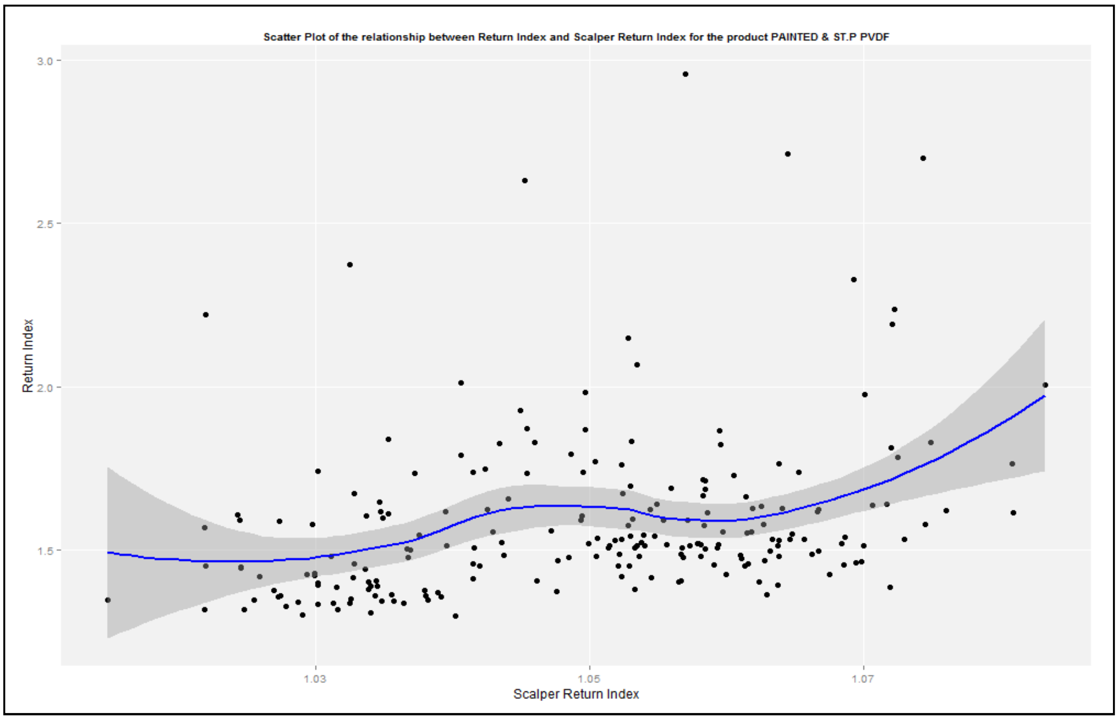


Figure 3.2.13: Scatter Plot with smooth line and 95% pointwise confidence interval of the relationship between Return Index and Scalper Return Index for the product PAINTED &ST.P PVDF

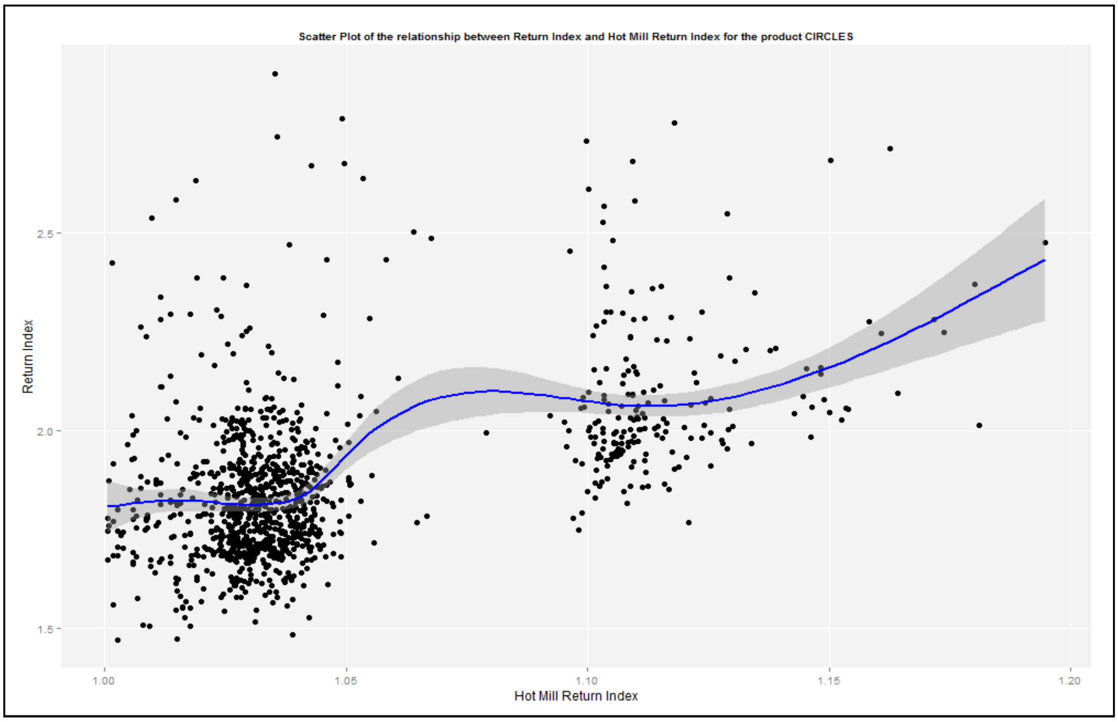


Figure 3.2.14: Scatter Plot with smooth line of the relationship and 95% pointwise confidence interval between Return Index and Hot Mill Return Index for the product FOOD SCROLL

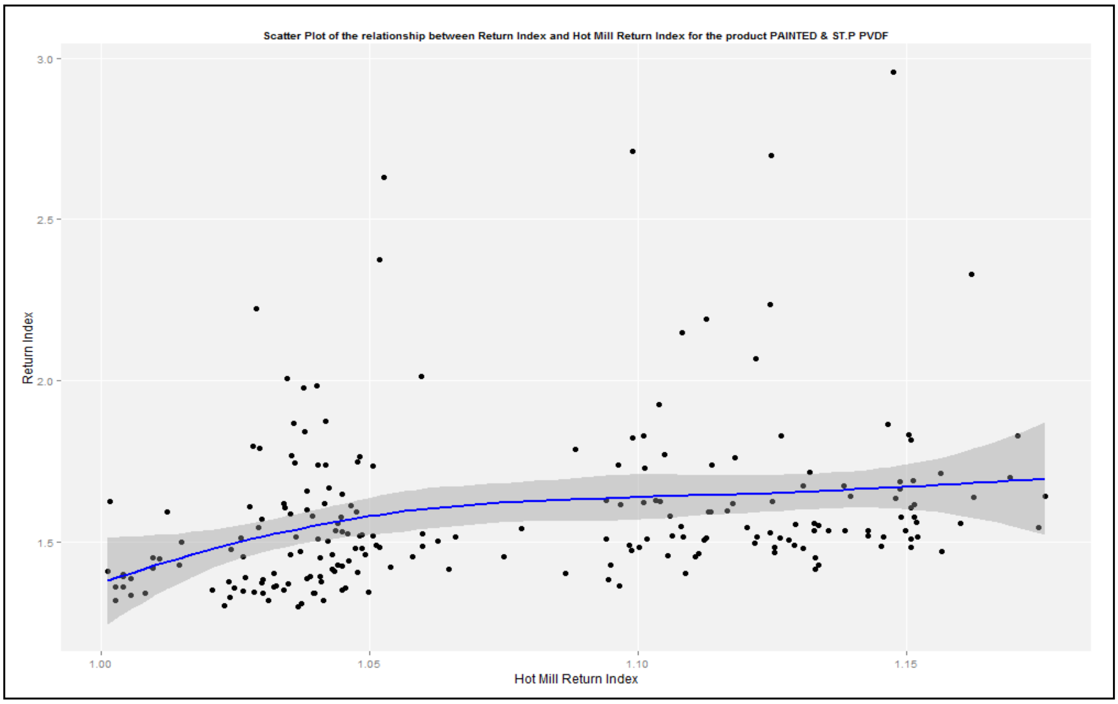


Figure 3.2.15: Scatter Plot with smooth line and 95% pointwise confidence interval of the relationship between Return Index and Hot Mill Return Index for the product PAINTED &ST.P PVDF

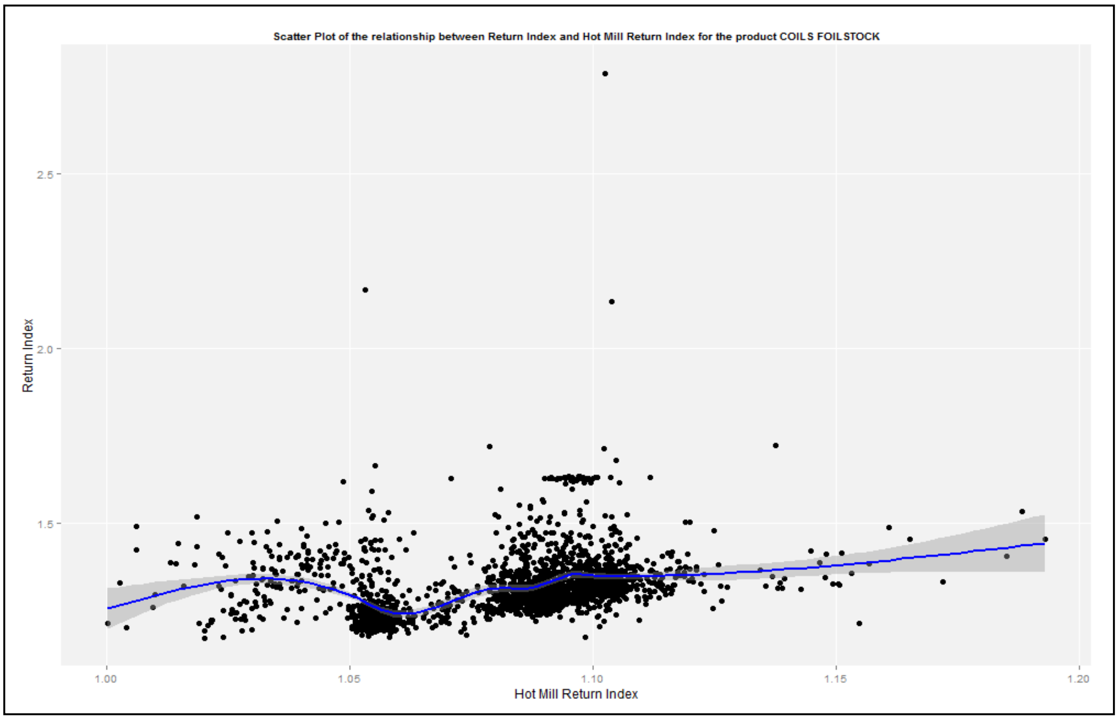


Figure 3.2.16: Scatter Plot with smooth line and 95% pointwise confidence interval of the relationship between Return Index and Hot Mill Return Index for the product CIRCLES FOILSTOCK

To sum up, based on the descriptive analysis of categorical variables which have been done done we have seen that the products FOOD FORMAT (17.23%), BODY (11.83%), END (11.31%), HI-MG (9.82%) and SHEETS 5XXX (7.92%) have the highest frequency levels of DMSY existence in contrast with the AUTOMOTIVE (0.7%), PAINTED & ST.P PVDF (0.66%), COILS PAINSTOCK (0.6%), FOOD SCROLL (0.49%) and SHEETS 3XXX (0.48%). Also, the slabs which end up to Jumpo Coils have higher DMSY percentage (54.9%) than these which end up to Standard Coils (39.8%) and CC Coils (5.3%). Furthermore, the Alloys which belong to the series 5xxx (48.91%) and 3xxx (36%) have blatantly higher percentages of DMSY existence relative to these which belong to series 1xxx (2.08%), 4xxx (1.51%) and 6xxx (0.35%). A very important result is that the reasons with code 154 and 24 appeared with higher percentages in the product SHEETS 5XXX, the reason with code 25 in the FOOD FORMAT, HI-MG and BODY and the reason with code 403 in the END and BODY. Moreover, the reason with code 73 appears more in the product BRAZING and the reason 64 in the FOOD FORMAT and END. Extremely interesting was the conclusion about the DMSY existence during the Hot Rolling process that we have higher percentages of DMSY existence in the stage of HOT ROLLING for the product BODY and in the stage of PAINTING LINE1 for the FOOD FORMAT. Also, in the stage of SLLITERING more DMSY reports appeared in the FOOD FORMAT, BODY, END, COILS PAINTED R/SHUT and in the CUTTING TO LENGTH MACHINE the products SHEETS 5XXX and HI-MG have the highest percentages of DMSY existence.

Based on the descriptive analysis of quantitative variables which have been done we notice that the products COILS FOILSTOCK, COILS T.P&BR, COILS PAINSTOCK and BODY take the lowest values of the Return Index and the FOOD SCROLL, BRAZING and CIRCLES the highest. Also, it is obvious that the products with DMSY existence take higher values of the Return Index relative to these which don’t exist. The most significant differences located in the products SHEETS T.P&BR, COILS 3XXX, PAINTED & ST.P PVDF, SHEETS 5XXX, COILS 5XXX and CIRCLES. Furthermore, we drew the conclusion that for the initial length which ends up to CC Coils we have lower values of the Return Index relative to these which end up to Jumbo and Standard Coils.

Also, apropos of the extra analysis about the correlation between Scalper Return Index, Return Index and Hot Mill Return Index, Return Index for each product category because strong positive correlations can lead us to expect higher Return Index while strong negative the opposite conclusion. The highest positive correlations coefficients between Return Index and Scalper Return Index appear in the products CIRCLES and SHEETS 5XXX and PAINTED &ST.P PVDF which are statistically significant whereas this with the highest negative coefficient is the FOOD SCROLL which is statistically significant. Furthermore, the products with the highest positive correlation coefficients between Hot Mill Return Index and Return Index are CIRCLES, AUTOMOTIVE, BODY, PAINTED &ST.P PVDF, COILS FOILSTOCK and AEROSOL which are statistically significant whereas with strong negative correlations is SHEETS 3XXX.

CHAPTER 4

MODELIING

In this chapter, will be done a model analysis based on multiple linear models, the BIC criterion, penalized regression (Lasso), CART algorithms (Regression Tree) and Cross Validation methods to check the accuracy of the above modeling methods. Modeling of our data will be done separately for each type of process, the Hot Rolling and the Continuous Casting because have different particularities and there are variables which they are not defined in both cases. As response variable we will use the Recovery Index because has a very good property that is located between zero and one.

4.1 Regression models for the Hot Rolling

Since the regression models have difficulties to handle categorical variables with many levels because for some of the levels have very little data to accurately estimate the coefficients, we run some pareto charts and for the variable Reason_ID kept 16 levels, for the Temper 16 levels, for the Final_Action_ID 5 levels, for the Destination_ID 6 levels, for the Machine_ID 11 levels and for the General Cat. 17 levels. As we can see from Figures 4.1.1 and 4.1.2 which are represented by heatmaps the Pearson and Spearman correlations between the quantitative variables in Hot Rolling, there is no strong correlation between them which could create a problem to fit a regression model or to draw us at some wrong conclusions.

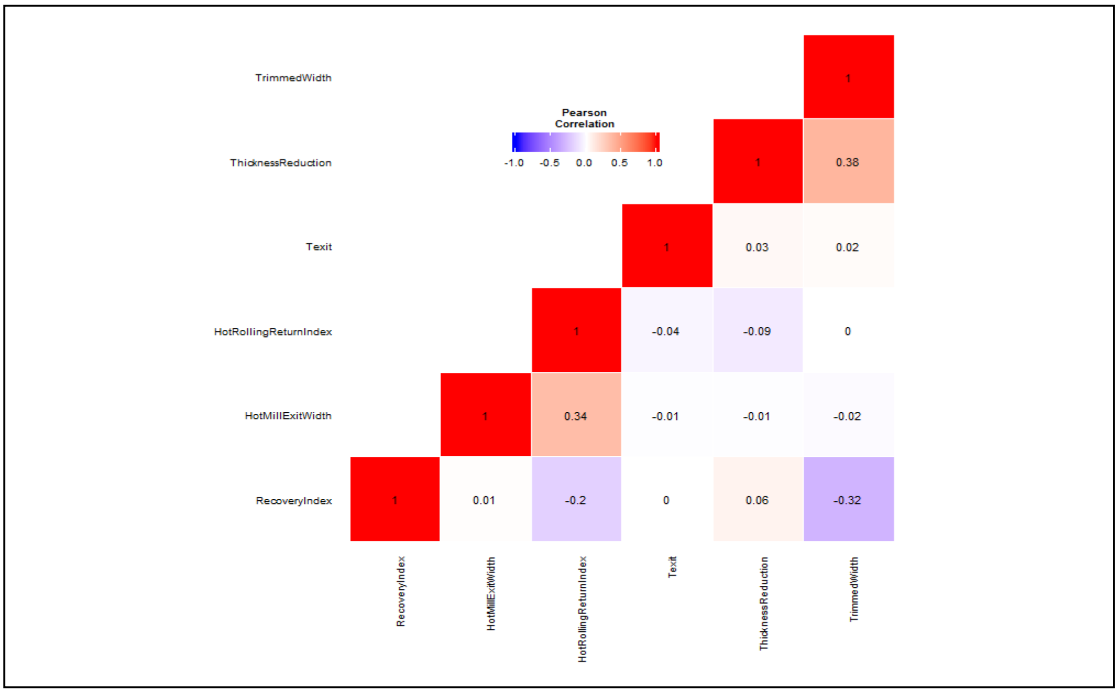


Figure 4.1.1: Pearson correlation coefficients between the quantitative variables in Hot Rolling.

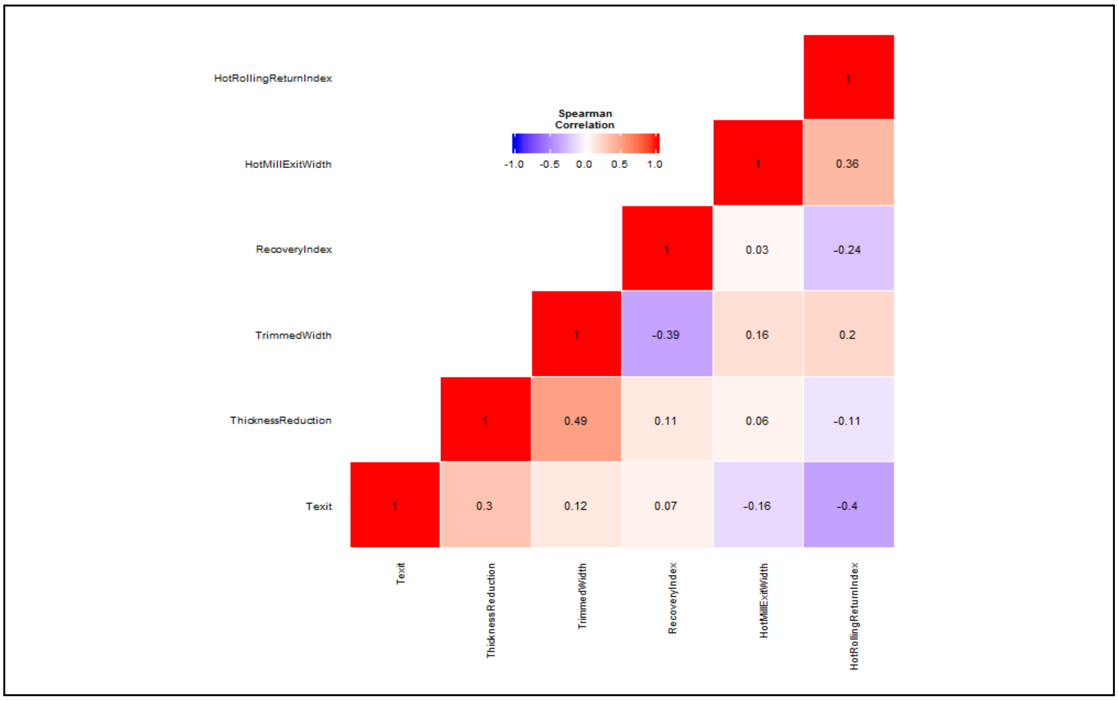


Figure 4.1.2: Spearman correlation coefficients between the quantitative variables in Hot Rolling.

Using the BIC criterion with forward direction in which basically fit the regression model by adding covariates one at a time based on BIC criterion and stop when the BIC is minimum, we draw the conclusion that the most important variables which selected are General Category, Final_Action_ID, Initial Length, HotRollingInex, Machine ID, Supplier, Utilization, WorkID, DestinationID and Temper. We stopped in the variable Temper because then the BIC criterion is not decreased anymore (Figure 4.1.3):

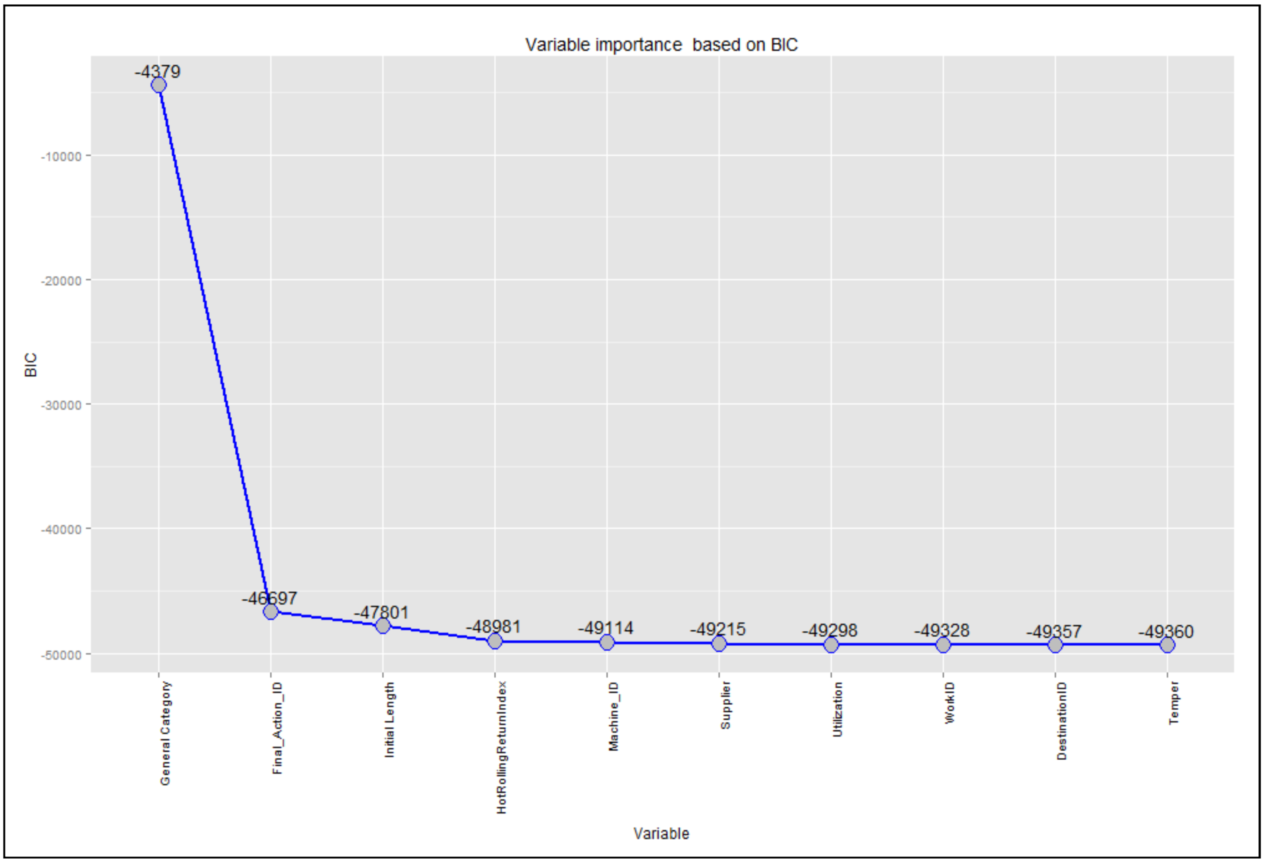


Figure 4.1.3: Variable importance based on BIC for the Hot Rolling

Through Lasso penalized regression, the variables which were selected and their levels (Table 4.1.1) based on tuning parameter lamda which is a standard deviation (more parsimonious models) from this which minimizes the $\mathrm{MSE}_{\mathrm{CV}(1)}$ (Figure 4.1.4) are:

Variable	**Levels**
1) General Category	
	BODY
	COILS PAINTED R/SHUT
	COILS T.P.&BR
	CIRCLES
	COILS FOILSTOCK
	HI-MG
	COILS 3XXX
	COILS 5XXX
	PAINSTOCK
	END
	SHEETS T.P.&BR
	ELVAL GRAIN
	SHEETS 5XXX
2) Final_Action_ID	
	DOWNGRADING COIL
	REMELT
	CONTINUE PROCESS
3) Hot Rolling Return Index	
4) Initial Length	
5) Machine_ID	
	OTHER
	STOCK
	P4
	SH1
	S6
	S2
	HM1
6) Temper	
	OTHER
	H116
	H14
	H16
	H18
	H26
	H42

	H46
	H48
	H49
7) Reason_ID	
	11
	154
	25
	403
	53
	54
	73
	85
8) Supplier	
	External
9) WorkID	
	RBR
	RT
10) Trimmed Width	
11) Destination_ID	
	HOT ROLLING
12) Alloy	
	4
	5

Table 4.1.1: Important variables based on Lasso regression for the Hot Rolling

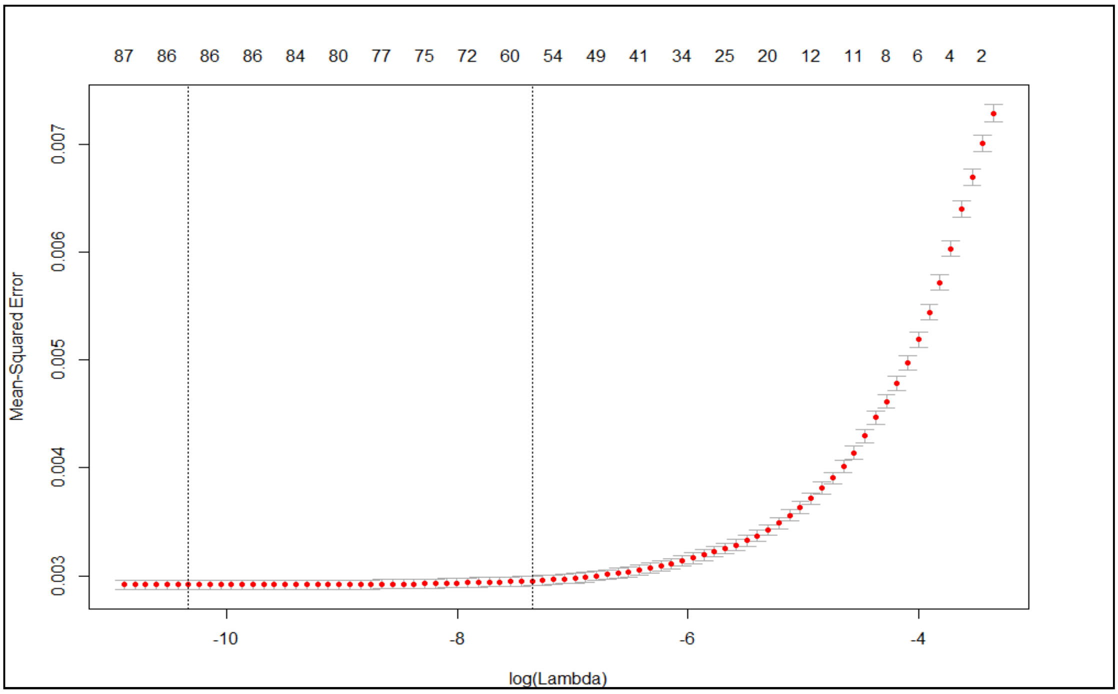


Figure 4.1.4: Relationship between Mean Squared Error and log(lamda) after 10-fold Cross Validation in Hot Rolling’s Lasso Regression

Also, as we can see from Figure 4.1.5 the variables which should be used as explanatory additively, in a multiple regression model to have the best possible prediction of the Return Index in the process of Hot Rolling are General Category, Final_Action_ID, Initial Length, Hot Rolling Return Index, Machine_ID, Temper, Reason_ID, Supplier, Utilization, Alloy and WorkID. We stopped in the variable WorkID because the PRESS is not decreased anymore. Our goal was to get interactions and square or higher degree terms in the model but because of a strong correlation between the categorical variables (if DMSY does not exist the categorical variables have a particular pattern) and that the generalized linear model cannot handle categorical variables with many levels, it was impossible.

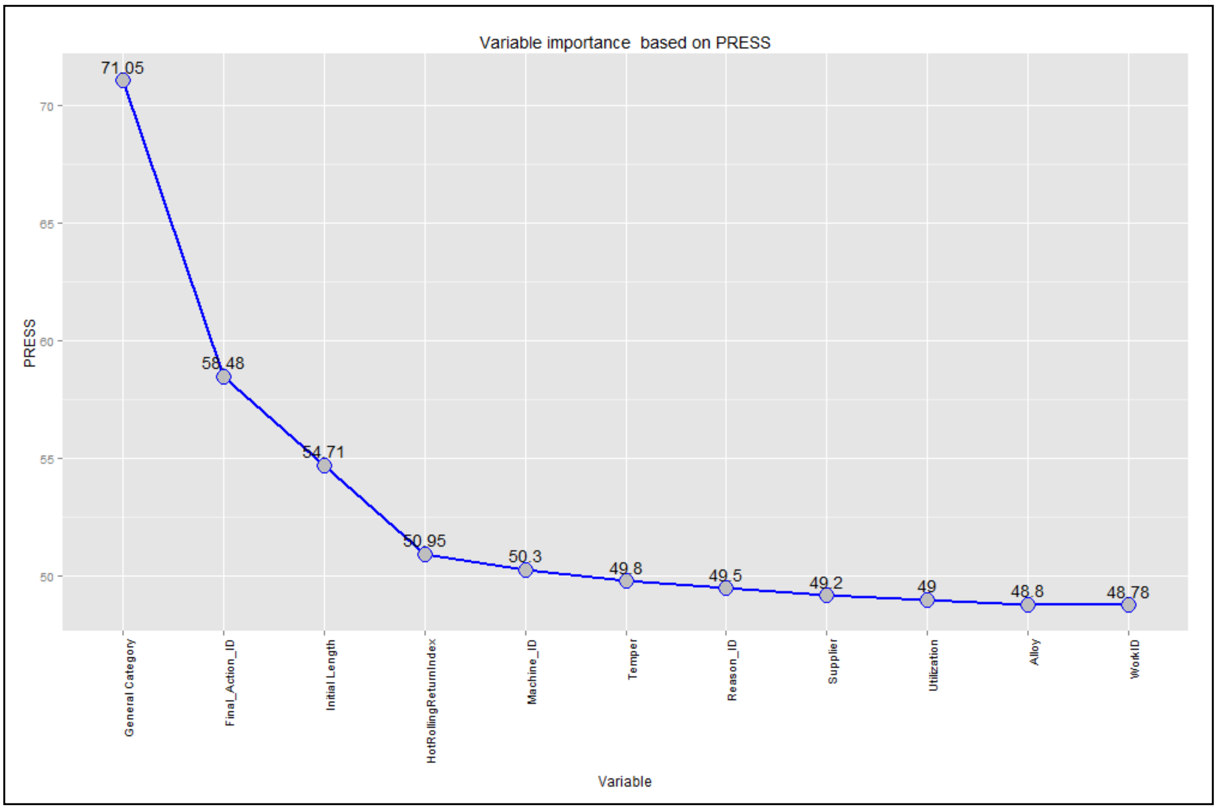


Figure 4.1.5: Variable Importance based on PRESS for the Hot Rolling

4.2 Regression Tree for the Hot Rolling

Figure 4.2.1 which shows the Hot Rolling's Regression Tree variable importance (page 16, Theoretical Background) we notice that the variables Final_Action_ID, General Category, Machine_ID, Destinatioin_ID and Reason_ID are the most important, whereas Painting Line1, Supplier, Utilization, HotMill, WorkID and Texit are unimportant. Furthermore, the regression tree which was created, will have the form as represented in the Figure 4.2.2 where the acronym G_C is General Category, FIN is Final Action ID, TrW is Trimmed Width, IL is Initial Length and REA is Reason ID.

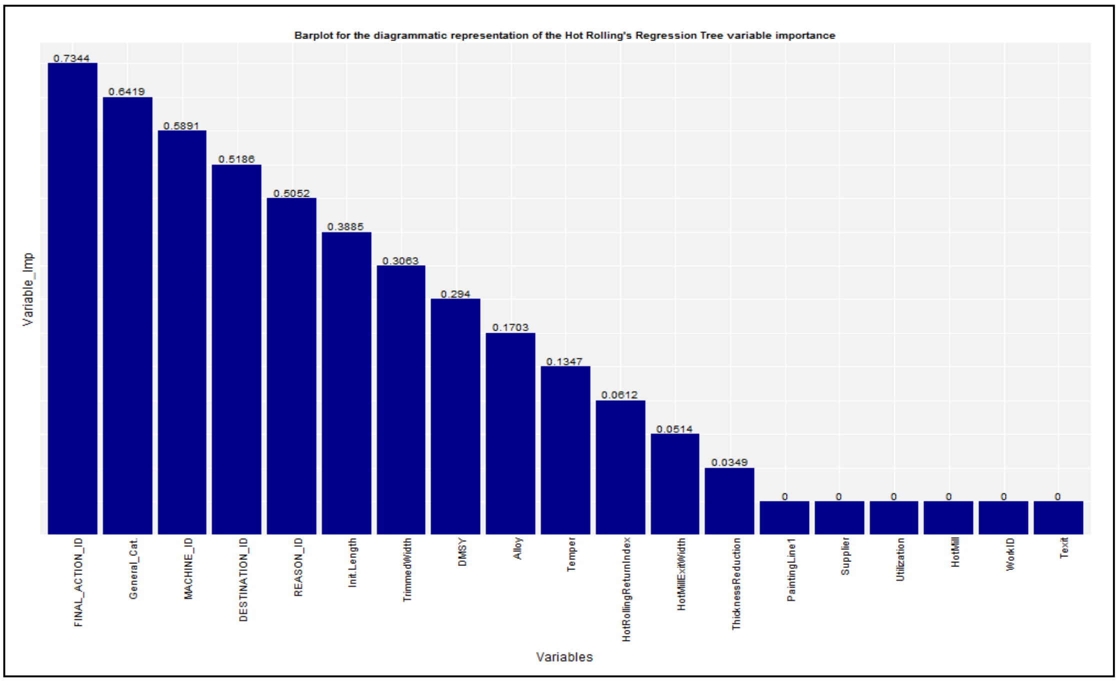


Figure 4.2.1: Barplot for the diagrammatic representation of the Hot Rolling's Regression Tree variable importance.

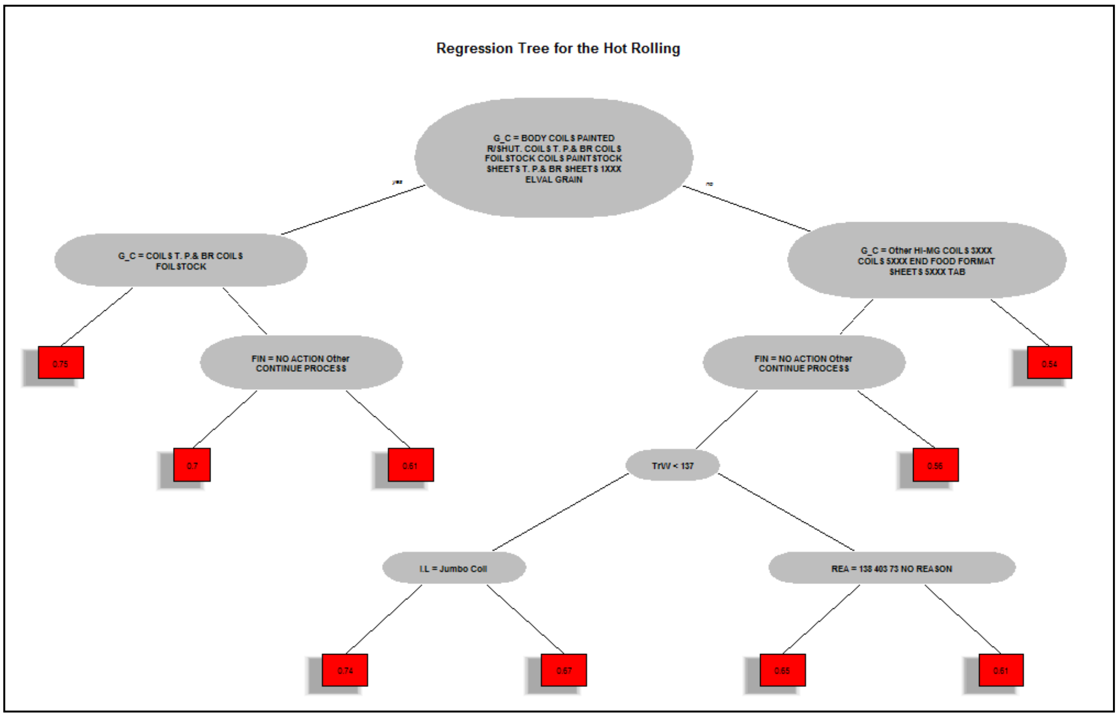

Regression Tree for the Hot Rolling
G_C = BODY COILS PAINTED R/SHUT. COILS T. P.& BR COILS FOILSTOCK COILS PAINTSTOCK SHEETS T. P.& BR SHEETS 1XXX ELVAL GRAIN
yes
no
G_C = COILS T. P.& BR COILS FOILSTOCK
G_C = Other HI-MG COILS 3XXX COILS 5XXX END FOOD FORMAT SHEETS 5XXX TAB
0.75
FIN = NO ACTION Other CONTINUE PROCESS
FIN = NO ACTION Other CONTINUE PROCESS
0.54
0.7
0.61
TrW < 137
0.56
I.L = Jumbo Coil
REA = 138 403 73 NO REASON
0.74
0.67
0.65
0.61

Figure 4.2.2: Regression Tree for the Hot Rolling

Finally, it's very important to check the predictive ability of the Hot Rolling's regression tree. For this reason, we would use 10-fold Cross Validation to calculate the prediction errors of the model. Company's first goal is to have a prediction error which is located between -0.05 and 0.05. From Figure 4.2.3 which is represented a histogram of the prediction errors we draw the conclusion that the majority of the errors are located in the specific interval. Also, we notice that its more likely the regression tree to overestimate the prediction than otherwise.

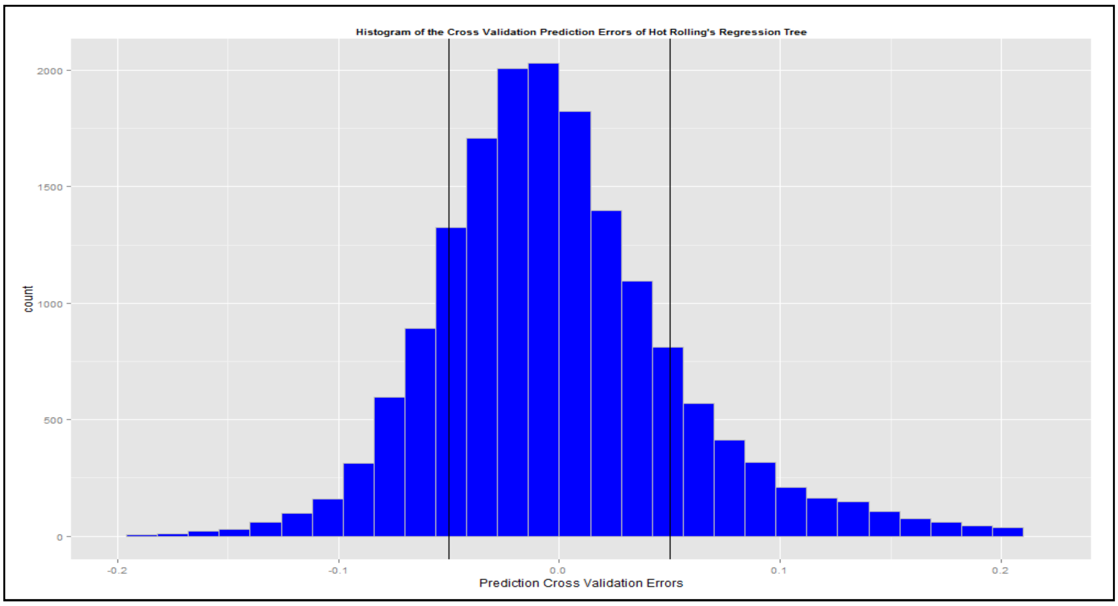


Figure 4.2.3: Histogram of the Cross Validation Prediction Errors for Hot Rolling's Regression Tree.

4.3 Regression Models for the Continuous Casting

As in the Hot Rolling so at Continuous Casting will try to find which variables influence most the Return Index and how we can achieve a better prediction of it, based on multiple regression models. Due to the regression models have difficulties to handle categorical variables with many levels because for some of the levels have very little data to accurately estimate the coefficients, we run some pareto charts and for the variable Reason_ID kept 11 levels, Temper 8 levels, Final_Action_ID 4 levels, Destination_ID 4 levels, Machine_ID 3 levels and General Cat. 4 levels. As we can see from Figures 4.3.1 and 4.3.2 which are represented by heatmaps the Pearson and Spearman correlations between the quantitative variables in Continuous Casting, whic there is positive correlation between Return Index, Coil Width and Coil Thickness and Coil Width.

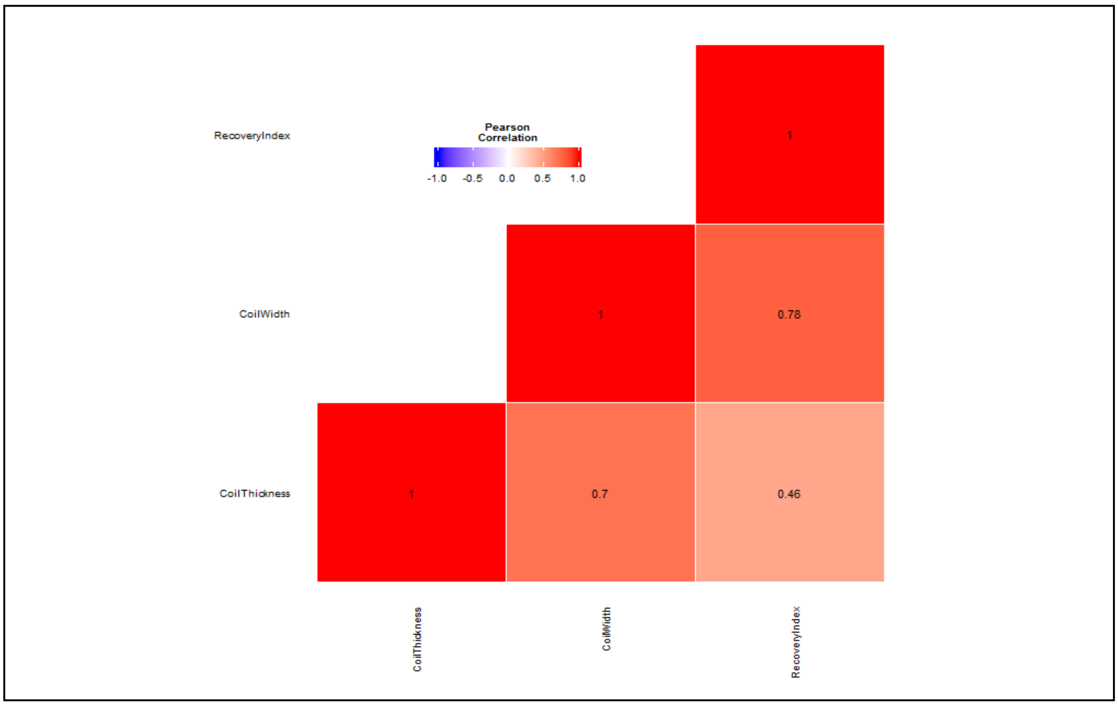


Figure 4.3.1: Pearson correlation coefficients between the quantitative variables in Continuous Casting

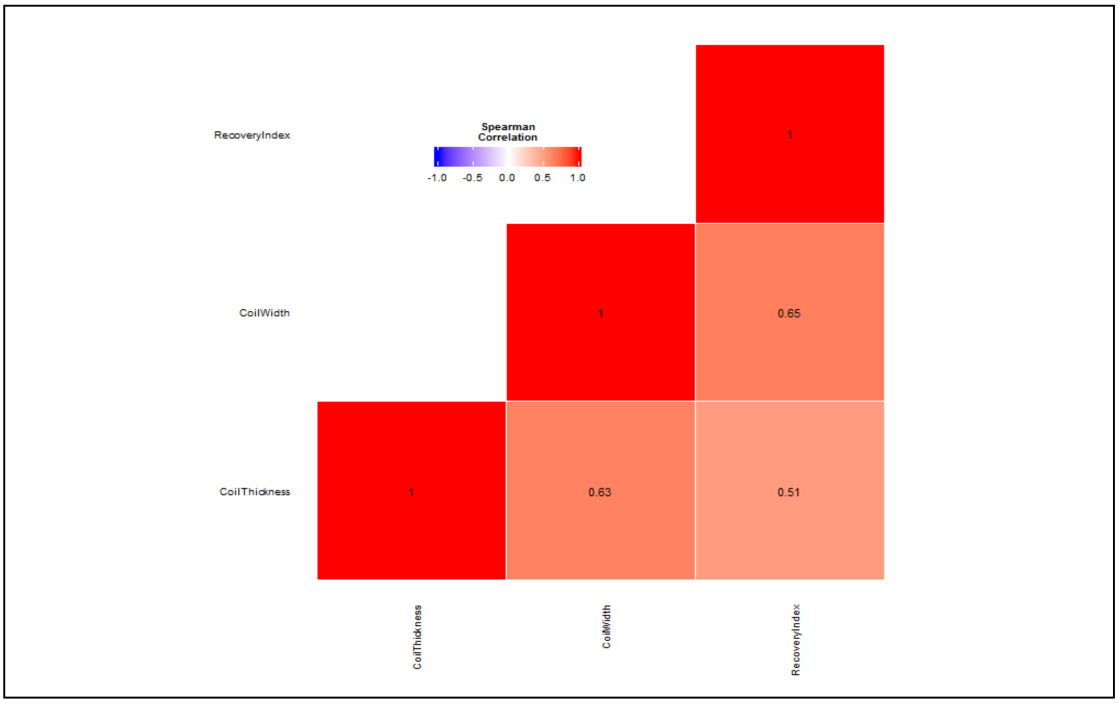


Figure 4.3.2: Spearman correlation coefficients between the quantitative variables in Continuous Casting

Using the BIC criterion with forward direction in which basically fit the regression model by adding covariates one at a time based on BIC criterion and stop when the BIC is minimum, we draw the conclusion that the most important variables which selected are Alloy, Final_Action_ID, General Category, PaintingLine1, Supplier, MachineID and Utilization. We stopped in the variable Utilization because then the BIC criterion is not decreased anymore (Figure 4.3.3):

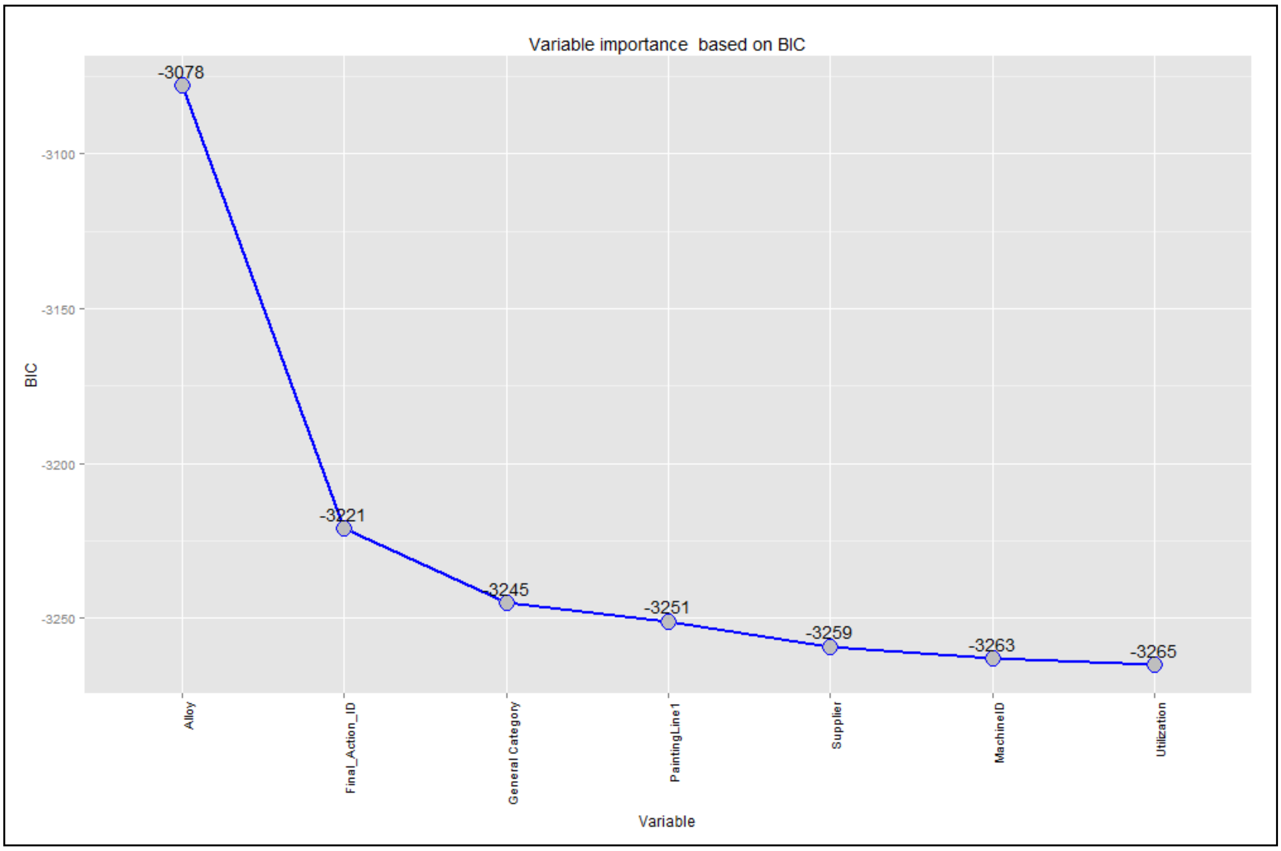


Figure 4.3.3: Variable importance based on BIC for the Continuous Casting

Through Lasso penalized regression, the variables which were selected and their levels (Table 4.3.1) based on tuning parameter lamda which is a standard deviation (more parsimonious models) from this which minimizes the $MSE_{CV(1)}$ (Figure 4.3.4) are:

Variable	**Levels**
1) Alloy	
	3
	4
2) Temper	
	OTHER
3) Utilization	
4) Coil Width	
5) Reason_ID	
6) Final_Action_ID	
	OTHER
	REMELT
7) General Category	
	COILS FOILSTOCK

Table 4.3.1: Important variables based on Lasso regression for the Continuous Casting

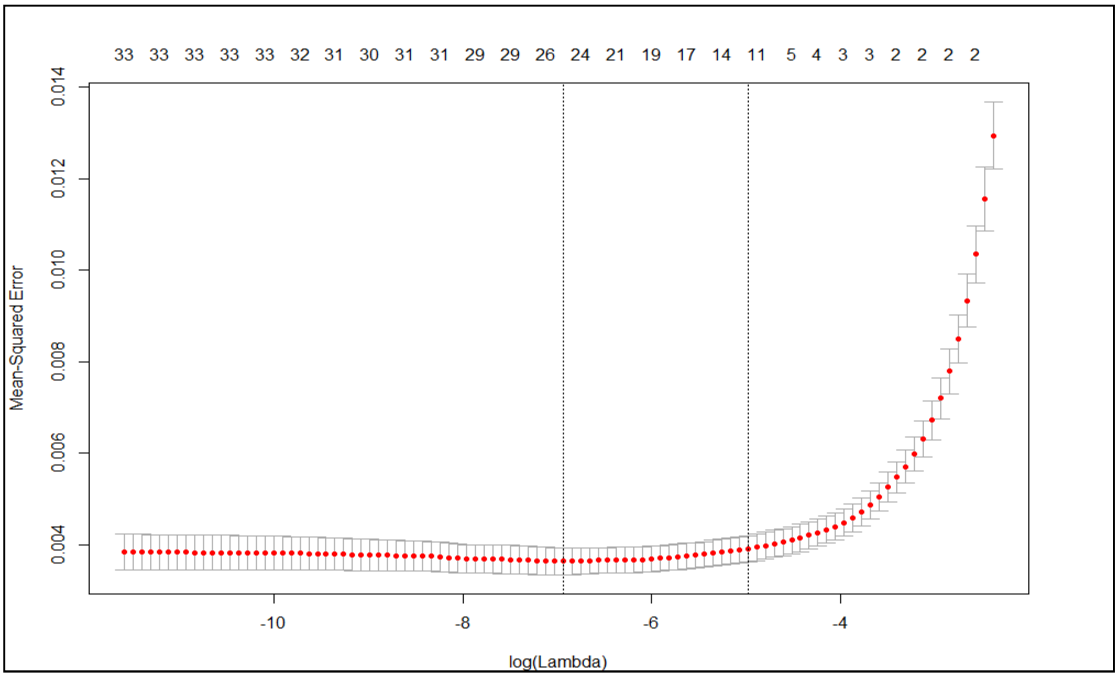


Figure 4.3.4: Relationship between Mean Squared Error and log(lamda) after 10-fold Cross Validation in Continuous Casting's Lasso Regression

Also, as we can see from Figure 4.3.5 the variables which should be used as explanatory additively, in a multiple regression model to have the best possible prediction of the Return Index in the process of Continuous Casting are Alloy, Reason_ID, General Category, Final_Action_ID, Temper, Machine_ID and Utilization. We stopped in the variable Utilization because then the PRESS is not decreased anymore. As in Hot Rolling so at Continuous Casting our goal was to get interactions and square or higher degree terms in the model but because of a strong correlation between the categorical variables (if a DMSY does not exist the categorical variables have a particular pattern) and that the generalized linear model cannot handle categorical variables with many levels, it was impossible.

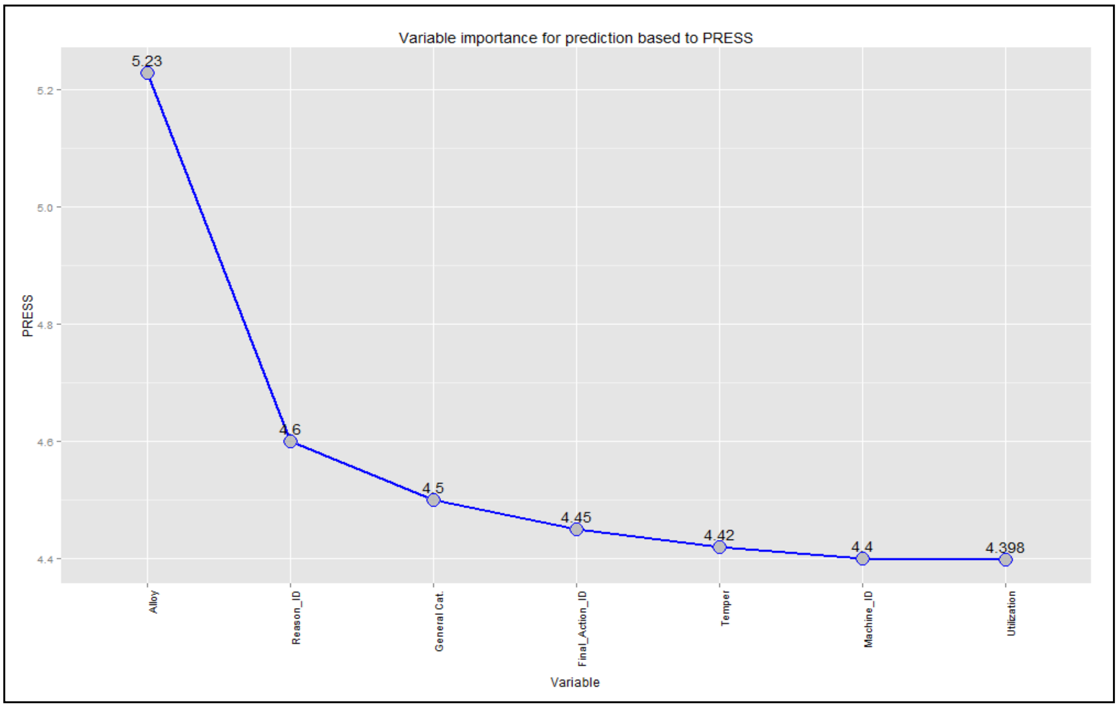


Figure 4.3.5: Variable Importance for prediction based on PRESS for the Continuous Casting

4.4 Regression Tree for the Continuous Casting

In this chapter, our goal is to be able to draw the conclusion about the modeling of the Continuous Casting's Return Index based on Regression Tree. From Figure 4.4.1 which represented diagrammatic the Continuous Casting's Regression Tree variable importance (page 16, Theoretical Background) we notice that the variables Coil Thickness, Coil Width, General Category and Temper are the most important, whereas Supplier, Utilization and DMSY are unimportant. Furthermore, the regression tree which was created will have the form as represented in the Figure 4.4.2 where the acronym All is Alloy, C/L is Coil Thickness and REA is Reason ID.

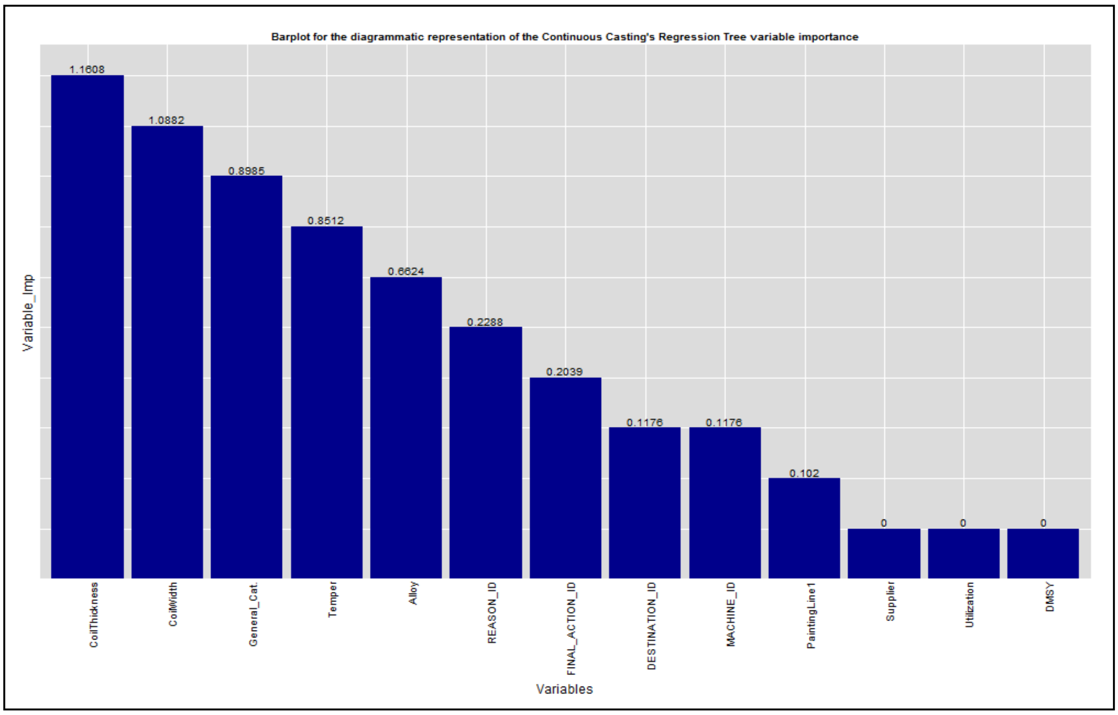


Figure 4.4.1: Barplot for the diagrammatic representation of the Continuous Casting's Regression Tree variable importance.

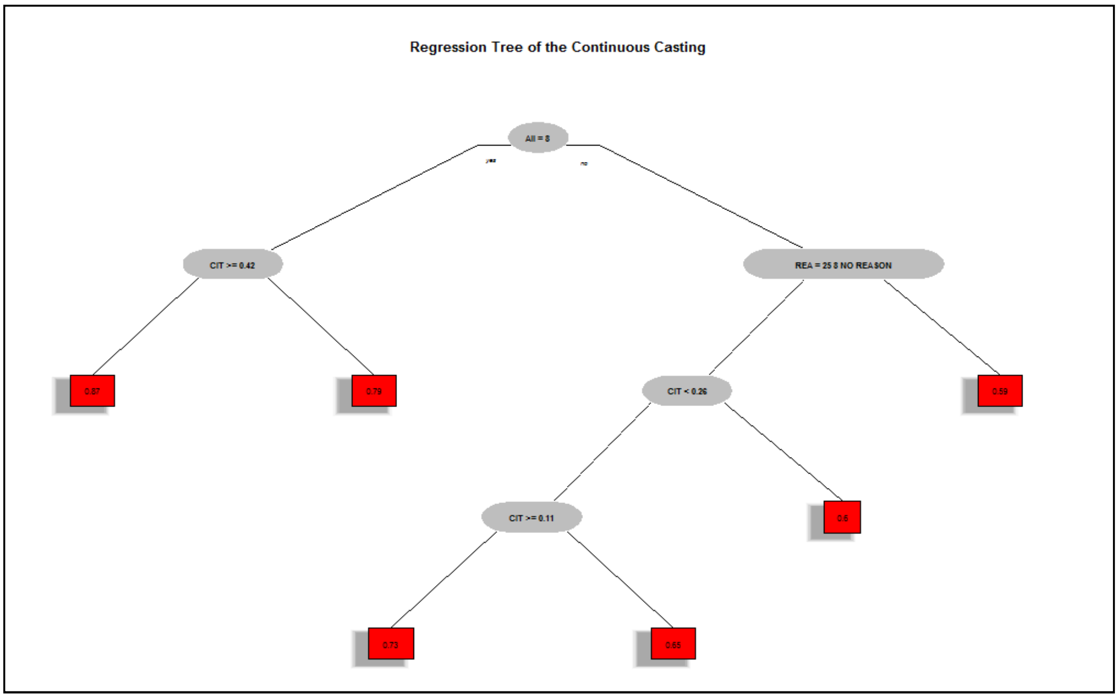


Figure 4.4.2: Regression Tree for the Continuous Casting

Finally, to check the predictive ability of the Continuous Casting's regression tree we would use again 10-fold Cross Validation to calculate the prediction errors of the model. Our goal is to have a prediction error which is located between -0.05 and 0.05. From Figure 4.4.3 which is represented a histogram of the prediction errors we draw the conclusion that the majority of the errors are located in the specific interval. Also, we notice that it's more likely the regression tree to overestimate the prediction than otherwise.

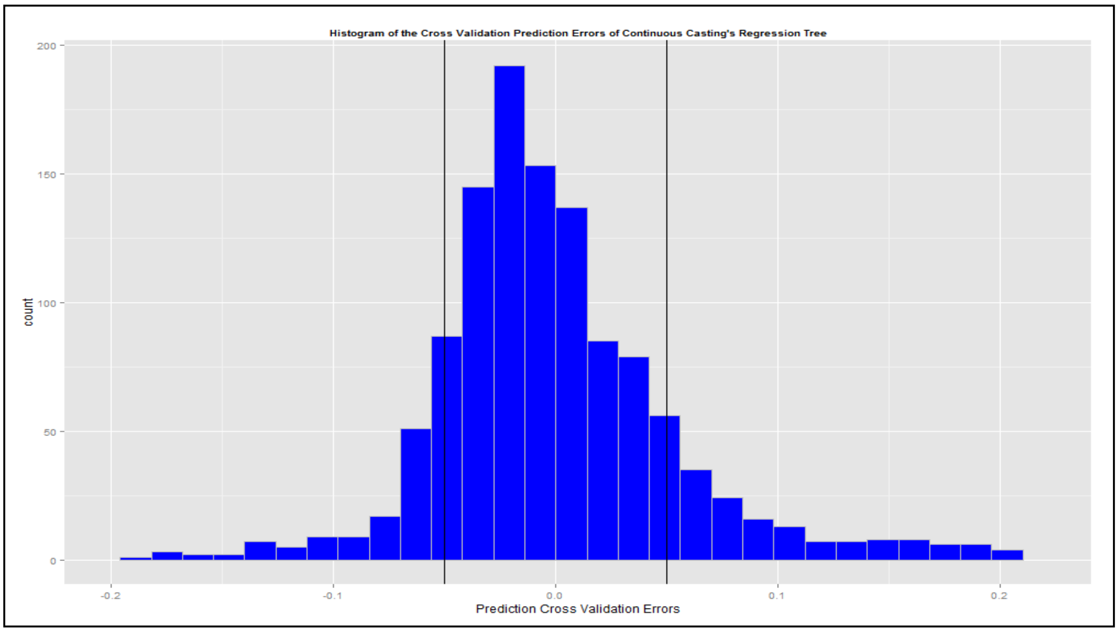


Figure 3.3.2.3: Histogram of the Cross Validation Prediction Errors for Continuous Casting's Regression Tree.

CHAPTER 5

CONCLUSIONS

In this chapter we will be presented the findings of our analysis, problems and further issues. We tried to to find where the most problems occurred during the production process, which factors affect more the final product and how can we predict the Recovery Index better having the minimum error using simple statistics measures such as Pearson Correlation Coefficient, Spearman Correlation Coefficient, Skewness, Kurtosis etc. and advanced figures such as Barplots, Pareto Charts, Box Plots etc. Also, for the modeling parted of this thesis used included multiple linear models, BIC criterion, penalized regression (Lasso), CART algorithms (Regression Tree) and Cross Validation methods.

5.1 Findings

Firstly, it is very important to understand where the most errors of the process exists, in our analysis means where the most DMSY reports appear. Based on the descriptive analysis of categorical variables which have been done we draw the conclusion that the products FOOD FORMAT (17.23%), BODY (11.83%), END (11.31%), HI-MG (9.82%) and SHEETS 5XXX (7.92%) have the highest frequency levels of DMSY existence in contrast with the AUTOMOTIVE (0.7%), PAINTED & ST.P PVDF (0.66%), COILS PAINSTOCK (0.6%), FOOD SCROLL (0.49%) and SHEETS 3XXX (0.48%). Also, the slabs which end up to Jumpo Coils have higher DMSY percentage (54.9%) than these which end up to Standard Coils (39.8%) and CC Coils (5.3%). Furthermore, the Alloys which belong to the series 5xxx (48.91%) and 3xxx (36%) have blatantly higher percentages of DMSY existence relative to these which belong to series 1xxx (2.08%), 4xxx (1.51%) and 6xxx (0.35%). A very interesting result is that the reasons with code 154 and 24 appeared with higher percentage in the product SHEETS 5XXX, the reason with code 25 in the FOOD FORMAT, HI-MG and BODY and the reason with code 403 in the END and BODY. Moreover, the reason with code 73 appears more in the product BRAZING and the reason 64 in the FOOD FORMAT and END. Extremely interesting was the conclusion about the DMSY existence during the Hot Rolling process that we have higher percentages of DMSY existence in the stage of HOT ROLLING for the product BODY and in the stage of PAINTING LINE1 for the FOOD FORMAT. Also, in the step of SLLITERING more DMSY reports appeared in the FOOD FORMAT, BODY, END, COILS PAINTED R/SHUT and in the CUTTING TO LENGTH MACHINE the products SHEETS 5XXX and HI-MG have the highest percentages of DMSY existence.

Based on the descriptive analysis of quantitative variables which have been done we notice that the products COILS FOILSTOCK, COILS T.P&BR, COILS PAINSTOCK and BODY take the lowest values of the Return Index and the FOOD SCROLL, BRAZING and CIRCLES the highest. Also, it is obvious that the products with DMSY existence take higher values of the Return Index relative to these which don’t exist. The most significant differences located in the products SHEETS

T.P&BR, COILS 3XXX, PAINTED & ST.P PVDF, SHEETS 5XXX, COILS 5XXX and CIRCLES. Furthermore, we drew the conclusion that for the initial length which ends up to CC Coils we have lower values of the Return Index relative to these which end up to Jumbo and Standard Coils.

Also, apropos of the extra analysis about the correlation between Scalper Return Index, Return Index and Hot Mill Return Index, Return Index for each product category because strong positive correlations can lead us to expect higher Return Index while strong negative the opposite conclusion. The highest positive correlations coefficients between Return Index and Scalper Return Index appear in the products CIRCLES and SHEETS 5XXX and PAINTED &ST.P PVDF which are statistically significant whereas this with the highest negative coefficient is the FOOD SCROLL which is statistically significant. Furthermore, the products with the highest positive correlation coefficients between Hot Mill Return Index and Return Index are CIRCLES, AUTOMOTIVE, BODY, PAINTED &ST.P PVDF, COILS FOILSTOCK and AEROSOL which are statistically significant whereas with strong negative correlations is SHEETS 3XXX.

Furthermore we worked in the modeling of Hot Rolling and Continuous Casting which was used as response variable the Recovery Index because it is located between zero and one. The variables which affect most the final Recovery Index of the Hot Rolling process are Final_Action_ID, General Category, Machine_ID, Destinatioin_ID and Reason_ID, whereas Painting Line1, Supplier, Utilization, HotMill, WorkID and Texit don't have any affection. In the process of Continues Casting the most important variables are Coil Thickness, Coil Width, General Category and Temper, whereas Supplier, Utilization and DMSY are unimportant. As we can understand the variables which describe where exist an error in the process (DMSY) such as Final_Action_ID, Machine_ID, Destinatioin_ID and Reason_ID are very important for the final value of the Recovery Index. Also, the most important variable seems to be the General Category. This may happen because of every product is a result of a different process with different characteristics. In respect of the prediction abilitry of our models especially Regression Trees for each process we noticed that was fairly well, since the prediction errors were located between -0.05 and 0.05. Definitely, this is not the best prediction error interval but a good approximation.

5.2 Further Issues

Furthermore, it is very important to accent the weakness which has the regression models to handle a large number of categorical variables with many levels as we have in our dataset. Despite we grouped some levels of our categorical variables to solve the previous problem, we didn't make it. In contrast, we lost important information of our data. For this reason, if we want better results it would be ideal in the next analysis relating to the specific data to use data mining algorithms such as Random Forests, Boosted Regression etc. because are more efficient to deal with large datasets with many multilevel categorical variables. Finally, it's very important to have correct data entries because even if there is a very good model, with wrong data entries it is more likely to lead us in wrong results.

REFERENCES

1. ***Allan Gut (2007).*** *An Intermediate Course in Probability,* Springer.
2. ***David Freedman (2009).*** *Statistical Models Theory and Practice,* University of California
3. ***Dimitris Fouskakis (2013).*** *Ανάλυση Δεδομένων με χρήση της R,* ΤΣΟΤΡΑΣ.
4. ***Gideon Schwarz (1978).*** *Estimating the Dimension of a Model,* The Annals of Statistics.
5. ***Hemant Ishwaran (2007).*** *Variable Importance in binary regression trees and forests,* Electronic Journal of Statistics.
6. ***Ioannis Ntzoufras (2014).*** *Advanced Data Analysis ,* Course Notes, AUEB.
7. ***John Fox, Sanford Weisberg (2011).*** *An R Companion to Applied Regresiion, 2nd Edition,* SAGE.
8. ***Leo Breiman, Jerome Friedman, Charles J. Stone, R.A. Olshen (1984).*** *Classification and Regression Trees,* Taylor & Francis.
9. ***Robert Tibshirani (1996).*** *Regression Shrinkage and Selection via the Lasso,* Journal of the Royal Statistic Society.
10. ***Roman Timofeev (2004).*** *Classification and Regression Trees (CART) Theory and Application,* Humboldt University, Berlin.
11. ***Ron Kohavi (1995).*** *A study of Cross Validation and Bootstrap for accuracy estimation and model selection,* Morgan Kaufmann Publishers Inc. San Francisco.
12. ***Tarpey Thaddeus (2000).*** *A note on the prediction sum of squares statistic for restricted least squares,* The American Statistician.