DETERMINING THE FAULTY AND REFUND PRODUCTS IN MANUFACTURING SYSTEM: APPLICATION ON A TEXTILE FIRM

Ezgi Demir[1], Sait Erdal Dincer[2]

[1]Piri Reis University, Department of Management Information Systems, Istanbul, Turkey.
 edemir@pirireis.edu.tr , ORCID: 0000-0002-7335-5094
[2]Marmara University, Department of Econometrics, Istanbul, Turkey.
 edincer@marmara.edu.tr , ORCID: 0000-0002-8310-1418

ABSTRACT

Purpose- In this study, it is aimed to improve the production and quality control processes of a company operating in the textile industry. For this purpose, predicting faulty and refund products by using simulation of oversampling and undersampling applications.
Methodology– In this study, there are 250 different variables and 72959 lines of data on the production line. These data have been taken from the last 1-year data of the firm. In this study, simulation has been done. New machine learning methods have been used by simulating. The reason for the simulation is that it was easy to detect the refund and faulty conditions made in a large lot group production in previous studies. However, the aim is to investigate whether the accuracy of the prediction algorithms will yield consistent results in terms of the increase in the number of refund and faulty products when production is made in a larger structure. In the simulation method, "oversampling" and "undersampling" methods have been used. While making simulation prediction, in the literature, boosting algorithms, which are used as ensemble machine learning techniques, have been used. In this study, simulation has been done as follows, while the number of production lots increased, refund and faulty products were increased within the same application. The reason for doing this is to investigate whether the prediction status in normal machine learning algorithms can be captured in a larger data stack. This process is called oversampling. Then, the "undersampling" method was applied. According to the "undersampling" method, it is aimed to determine the refund and defect situations in a smaller lot by taking samples of refund and defective products with less frequency. At the end of the study, the results were interpreted by applying boosting algorithms.
Findings- As a result of the study, it is concluded that "undersampling" and "oversampling" simulations predict better than usual machine learning methodology.
Conclusion- In this study, it has been observed that the ensemble machine learning algorithms (adaboost, xgboost, gradient boosting algorithms), which are one of the ensemble machine learning methods that emerged in 2016, were applied to the production data for the first time and showed success in the prediction of faulty and refund products.

Keywords: Machine learning, boosting algorithms, production planning, fault detection.
JEL Codes: C45, L23, C44

1. INTRODUCTION

Production planning is one of the most important issues in the production process. Production planning has become more important issue within factories due to the intensification of production systems and the improvement of capital activities. Factors that make production planning to be at the forefront in a business can be listed as follows:

- ✓ Difficulties in administrative process management of activities in businesses,
- ✓ Competition status and continuity of relationships among the existing businesses in the same market,
- ✓ Production processes are continuous and complex,
- ✓ Determining the factors in the demands and preferences of consumers,
- ✓ Efforts to have a place in the market in terms of service, quality and price with technological developments,
- ✓ Proliferation of supply and distribution activities with the increase of the customer network,

- ✓ Providing service at the enterprise's mass production level,
- ✓ Ensuring that material, raw material, machine hours and labor losses are minimized.

The main functions of the production planning process can be listed as follows: First of all, orders are transferred to the system according to their delivery dates and priorities. Business processes are initiated and assigned according to business processes. In order to complete the orders, priority order information of the transactions is provided in the system. It is ensured that the delay times in production and the speed of the processes are controlled. Priority status of the works is controlled with other report operations such as reporting summaries, faulty and warehouse products. According to machines, people, departments, workload and capacity are compared and input / output analyzes are made and reported.

The production planning process is extremely important in a business. The main purpose of production planning is to realize the production of a particular product in the required quantity, at the specified quality and time. In order to achieve this, the necessary production factors must be either produced by the enterprise or supplied on time from another location. In addition, the qualities of the product, namely its quality, quantity, required raw materials, materials, staff, labor force and capital costs are the direct factors affecting production (Tekin, 2012). Production planning fulfills the specified functions, respectively (Demiröğen & Güzel, 2009).

- ✓ Adjusting the inventory level of the finished goods or semi-finished goods.
- ✓ In this process, production times are shortened with effective production planning.
- ✓ Provision of supply to main suppliers and distribution centers in a short time,
- ✓ If more than one product will be produced in the production plan, providing the necessary flexibility to provide opportunity the lucrative demands,
- ✓ Arrangement of production and stocks based on sales forecast depending on potential customers in the production process,
- ✓ Checking whether the production processes are in the long-term production plan or the short-term production plan, making available resources full stock,
- ✓ The most important function that should be in every production plan is to minimize costs and maximize the quality and profit level by using the available production resources and facilities in the most optimal way.

The principles to be followed in the preparation of the production plan in the production process of an enterprise can be listed as follows.

- ✓ Determining the planning period,
- ✓ Appropriate material grouping,
- ✓ The restrictive factors have been determined and taken into account.

When the production plan has been prepared according to these principles, it becomes a tool for the distribution of time intervals according to production, the production process according to these plans, and the necessary machinery, control and equipment. During the production plans, after the distribution of tasks made after the main production planning, they are equally responsible for this process, from the person who physically performs the work on the machine to the manager at all levels during the sharing of the entire workload. Even a yarn or color change in production raw materials significantly has affected the quality of the process. Changes in the production process completely has affected the quality of the products.

In this study, the enterprise produces more than one product in the production process. In production, raw material products have gone through 15 different processes. Different production errors has occurred in each of these processes. In this study, the outputs of the variables in the production process are dealt with and the correlation results that cause the formation of faulty products have been presented. Boosting algorithms have been applied to the variables resulting from the correlation, and the error results have been tried to be determined before the error occurred at the end of the production process. At the same time, some faulty products are returned in this company. In this study, the prediction of refund products by boosting algorithms have been also performed. In this study, in the first paragraph it is expiained the manufacturing systems in the firm and its importance for the firm. In the second part, it has been given the literature review related with manufacturing in the firm the scope of machine learning and data mining applications. In the third part, the data and and methodology have been given. In the next part, findings, conclusion and further studies have been also given, respectively.

2. LITERATURE REVIEW

There are many studies about manufacturing and quality control process in literature. In this study, it has been studied boosting machine learning algorithms to decrease the number of faulty products.
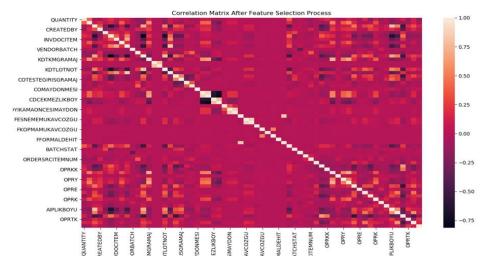
In one study, machines operating cold press in a production facility stop at certain intervals. This downtime has caused economic losses due to the process in which production has not been carried out. In order to reduce this time, prediction has been made in the optimized usage processes of the machines. Then, classification, decision tree and weighted learning algorithms, which are data mining techniques, have been used in the related study. These methods have been applied in Weka and Rapidminer programs. For another study, in a beverage production company, the data cube created according to the volume, sales channel, delivery location, activity type, product group, brand and time dimension has been analyzed first. As a result of the model created, it has been observed that clustering has gained importance according to the type of activity, product group, product size (volume) and daily cash receipt. In this study, data mining application has been carried out on the 5-month transaction data of the customers, which are producing regionally, specifically for the enterprise. Clustering method and decision trees algorithms have been used in practice. The customers have been divided into homogeneous groups by using the clustering method in order to determine the company's distribution strategies according to the customer groups by considering the daily transaction data of the customers. By applying decision trees algorithms, the data hidden in huge volumes could be useful for the company has researched. For another study, in a logistics company, the tables in which the main information regarding the orders received from the customers are kept and the charts kept in line with the demands of the customers have been used. In this study, it is aimed to find the limit value that separates the points that geometrically correspond to the customers who lost and continue to work by looking at the compliance and demand dates in two-dimensional space. In this model, classification and decision trees algoritms have been applied. Another study, supplier selection has been made in a production facility. Among the variables in the data set, the most important variable was chosen as the "delay" as the decision variable for the supplier selection. The relationship between each variable and the delay was examined in the established model and the variables affecting the delay were determined. While measuring success level, "delay = 1", "on time delivery = 0", respectively (Quality Certificate (IS0-TSE) - R&D - Guarantee certificate - Sector group - Number of administrative staff - Number of technical staff - Form of establishment - Firm type - City - Number of items variables were taken. In the analysis, artificial neural network, C&R Tree, C5.0, Chad, one of the decision tree algorithms, and K-Means, Kohonen and Two Step algorithms from the clustering algorithms were used in this study. However, K-Means from these algorithms have resulted a higher consistency ratio. This analysis was done with SPSS Clementine program. In another study, data mining has been discussed in detail in a production company. In the application part, the reasons for the improper separation of the products produced in a manufacturing enterprise have been determined. Variables of the study have been groupped as shift order, production month, production day, error groups (originating from the person, machine, machine and person), production frequency, quantity put in the package, factory, machine failure, product group, customer, staff, profile type and production type (mass production or not). A model was developed for decision analysis with decision trees and artificial neural networks algorithms. SPSS Clementine 11.1 application was used. In another study, data mining applications were carried out in a carpet manufacturing enterprise to improve the carpet manufacturing process and product quality. Product error data and machine data were modelled separately. An effective approach was used for variable reduction with feature compatibility analysis. The causes of carpet and machine faults have reached a standard with knowledge gaining technique. Error id, date, workbench, mail, employee, quality, pattern, color, width, height, error part, customer, type of error, cause of error, barcode order, barcode date variables have been studied. Along with statistical methods, C4.5 algorithm and artificial neural networks (Radial Based Function Networks) were used. For another study, optimization work was carried out in a leading company in the automotive field. The study took place in the cutting section of the company. Error variables such as stitch error, cut error, half piece, fabric error, end-of-eart piece, pattern shift, color difference were determined. Relationships between input and output variables were interpreted with Canonical correlation analysis. Then, with Multiple Regression Analysis, separate estimation equations were created for each output. The assignment of the most appropriate values for the input variables has been achieved by a hundred percent at the desired values for the width and height. Fifty-five percent improvement has been achieved for the hole diameter. With the Gams Program, error codes and error types have been defined then applied statistical methods and comparisons have been made with multiple regression analysis. For another study, it is aimed to estimate the production time. For this purpose, the water meter production process has been examined. The analysis was made in the WEKA program using the Decision Tree C4.5 Algorithm. For this purpose, a mathematical programming model has been created by using the association rules of data mining techniques. But a genetic, hybrid algorithm has been developed to apply to large-scale problems. Retailing sector was chosen as an application. Ms-sql, statistical methods, association rule and LINGO program were used in the study. For another study, data covering the first 9 months of 2009 on an operational basis belonging to an institution serving in the retail sector was used. The target audience of the company are retail businesses and catering businesses. The parameters in the data are classified as products, product codes, product packaging information, product main brand codes, sub-brand codes, information about the category and sub-categories, product category and customer information. The aim is to ensure that customers with high profits have been grouped and to focus on these factors. Data manipulation methods were applied and then EDA methods were used. Classification and regression models from data mining models, clustering and association rules

have been examined. At the same time, by using C&R Tree and CHAID algorithms, it has been tried to determine which factors are effective in gathering "high amount of profit" customers in this group. The studies were carried out with SPSS Clementine package program. Another study was developed a quadratic model in the field of production logistics, and made a multi-dimensional scaling analysis that was used to locate the products in the warehouse. In this study, traveling salesman problem used in optimizing routes during order picking was explained. This application was made with one-way analysis of variance, association rules and independent sample t-test. Also, paired sample t-test and statistical methods were used. Another study carried out in a company that does business in the production sector, in the quality control department. In the quality control process, measurement data is very high and in this sense, as the volume of the data increases, the rate that people understand decreases. Variations are the enemy of quality and there is variation in everything. In this study, a decision support system was implemented in the quality control process with classification algorithms, which are data mining methods. While preparing this study, Cross Industry Standard Process Model (CRISP) was used for data mining. NaiveBayes, SMO algorithms, C4.5 Random Forest Classification Algorithms, R, Weka and Minitab programs were used.
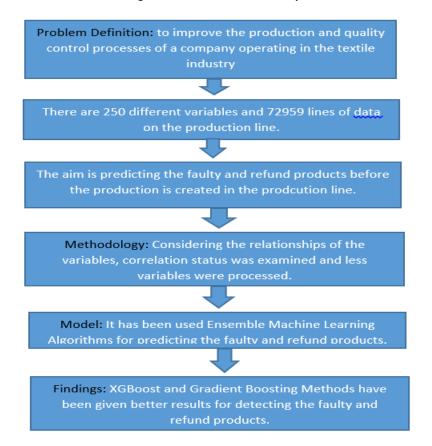
3. DATA AND METHODOLOGY

3.1. Data

In this paper, an estimation of the number of faulty products was made by considering the production variables in a company producing in the textile sector. While doing this study, simulation was done. The reason for the simulation was that it was easy to detect refund and faulty products in a small lot group, but the process have become more difficult as the amount of mass production increases. The aim is to determine the accuracy of the prediction algorithms in this determination point in the same way the increase the number of refund and faulty products when a larger lot production has been made.The question was tried to be investigated.In the simulation method, "oversampling" and "undersampling" methods were used. The concept of "oversampling" is referred to as a higher number of sampling. In other words, while the number of production lots is increasing at this simulation point, the question of whether it can be showed the same success in a more crowded data stack is increased by increasing the refund and faulty products within the same application. Likewise, according to the concept of "undersampling", it is aimed to determine the refund and faulty situations in a smaller lot by taking samples of refund and faulty products with less frequency.

In this paper, the sector was examined in the manufacturing sector. There are 250 different variables that affect the production result. Later, considering the relationships of the variables, correlation status were examined and less variables were processed. The reason for this is that some of the existing variables have more effect on the result, while other variables have less effect on the result. In addition, some variables include "multicollinearity" problem. "Multicollinearity" problem shows that there is a relationship between the variables and this relationship eventually leads to inconsistency. In order to reduce the correlation between the variables, the variables that have the most impact on the product were taken. The variables that have the highest effect on the product were taken from the variables that have high relation with each other. This process is called attribute selection in Figure 1.

Figure 1: Correlation Matrix After Feature Selection

As seen in Figure 1, the correlation has gone from -1 to 1. There is a positive correlation at point 1 and a negative correlation at point -1. There is a strong negative correlation from -0.5 to -1. There is a positive correlation from 0.5 to 1. Attributes that are strongly correlated with each other cause a multicollinearity problem. Therefore, instead of taking two variables, modeling has been done by taking the variable that affects the production the most. Figure 1. "VariantKey", "Kdthkmpus", "Kdtkmen", "Fetuvboycek", "Fph", "OutQuantityx", "Status", "Oprbs", "Oprsk", "Oprm", "Oprgrm","Oprd", "Hmkmbatch", "Oprfk" since these operation variables cause "Multicollinearity" problem, it should be disabled because it will affect our selection of feature. Considering Figure 1, the feature that is between the +0.5 and -0.5 correlation band has been selected by removing the mostly light colored parts and much darker colored parts from the feature selection. As a result, it is important that the variables to be modeled correlate with each other. If there is a high correlation between the variables as in the Figure 1, the variables that affect the production should be considered. Other variables should not be added to the model. After determining the required attributes, XGBoost, ADABoost and GradientBoost algorithms were applied. These algorithms make a difference in manufacturing in the literature. These algorithms are referred to as Ensemble Machine Learning Algorithms in the literature. The reasons for using these algorithms are that there is no work in the field and avoiding "oversampling", ie "over-learning" status of our data, enables ensemble machine learning algorithms. In our study, the condition of refund and faulty products constitutes 8% of a data of 72 thousand lines. Usual machine learning algorithms easily learn 92 percent of 72 thousand lines, and predict the rest at high prediction rate. The point to avoid here is that the machine does not memorize it. If it is memorized, the application made when more lots of products come is also likely to encounter errors. Therefore, every situation should be considered. Another reason for the need for an ensemble machine learning algorithm is the "precision" and "f" scores in the model. Next part of the study, ensemble machine learning algorithms will be discussed. And also the flow of the study can be seen in Figure 2.

Figure 2: The Content of the Study

Problem Definition: to improve the production and quality control processes of a company operating in the textile industry

There are 250 different variables and 72959 lines of data on the production line.

The aim is predicting the faulty and refund products before the production is created in the prodcution line.

Methodology: Considering the relationships of the variables, correlation status was examined and less variables were processed.

Model: It has been used Ensemble Machine Learning Algorithms for predicting the faulty and refund products.

Findings: XGBoost and Gradient Boosting Methods have been given better results for detecting the faulty and refund products.

3.2. The Proposed Methodology: Ensemble Machine Learning Algorithms

Gradient Boosting Algorithms: Gradient Boosting algorithm is an ensemble machine learning algorithm that works for both regression and classification problems. Gradient Boosting algorithm has used the boosting technique, combining a number of weak learners to create a power learner. The regression tree algorithms have used as the basic learners are based on the errors calculated by the previous tree, one tree in series. Since it comes from decision tree algorithm in Gradient Boosting algorithm, initial leaf is created first. Leaf nodes symbolize predictions; they are the extreme points of trees. This leaf holds our first prediction about the weights of samples. Then, new trees are created by considering the prediction errors. This situation continues until the number of trees decided or no further improvement from the model. First, the variable to be predicted (target variable) is averaged. This number is the initial leaf, our first predict attempt. By comparing this value with the target variable, it is observed how many incorrect predictions are made. Residual value is found by subtracting the predicted value from the observed value. The initial tree will be established as a model that predicts the errors obtained as a result of the initial leaf. If it has been added the first prediction and the results from the initial tree, it can be predicted the target variable 100% accurately. However, this situation is caused overfitting. Gradient Boosting algorithm has added learning rate to trees to overcome this overfitting problem.

XGBoost Algorithms: XGBoost (eXtreme Gradient Boosting) is a high performance version of the Gradient Boosting algorithm optimized with various arrangements. It has been introduced with the article titled "XGBoost: A Scalable Tree Boosting System" published by Tianqi Chen and Carlos Guestrin in 2016. This algorithm has not been implemented in the manufacturing field yet. The most important features of the algorithm are its ability to obtain high prediction rate, to prevent overfitting, to manage empty data and to do them quickly. According to Tianqi, XGBoost runs 10 times faster than other popular algorithms. Software and hardware optimization techniques have been applied to achieve better results using fewer resources. XGBoost Algorithm is a decision tree based algorithm and is shown as the best among decision tree algorithms. The first step in XGBoost is to make the first base score. This prediction number can be any number since the correct result will be achieved by converging with the next steps. This number is 0.5 by default. How well this estimate is made is analyzed with incorrect residuals of the model. Errors have been found by subtracting the predicted value from the observed value. At the next stage, the decision tree that predicts errors is established. The aim is to learn the errors and approach the nearest prediction. The similarity score is calculated for each branch of the tree created. Similarity score shows how well the data is grouped in branches. Accordingly, the formula is as follows.

Similarity score= $\dfrac{(The\ sum\ of\ incorrect\ predictions)^2}{Number\ of\ incorrect\ estimates + \lambda}$

The purpose after calculating similarity scores is whether a better prediction can be made. The value of lamda (λ) in the similarity score formula is the parameter of regularization. Trees of all possible possibilities are modelled for a better prediction. Similarity scores are calculated for all tree values. To decide which tree structure is better, the gain must be calculated. While the branches of the tree have been evaluated with the "similarity score", the whole tree has been evaluated with the "gain" score. The formula for the gain score is as follows.

Gain Score = Left Similarity Score + Right Similarity Score - Similarity Score of Previous Tree

After determining which tree has higher gain score and deciding to use the tree, pruning process begins. The value called "gamma" is selected for pruning. Gamma is an assessment brought to gain score. Branches with a gain score are lower than a gamma score require pruning. Therefore, increasing the gamma value only helps to keep the valuable branches in the tree and to prevent overfitting. Pruning continues from the last branch up. If it is decided not to prune the bottom branch, there is no need to examine the branches above. The default gamma value is 0. If the gamma value is 0, it means that if the gain score gets negative value, the branches will be pruned. If all the branches are pruned, the best prediction is 0.5 at the top. After these definitions and calculations, the lambda value mentioned in the similarity score can be examined in more detail. This value refines the model. According to the similarity score formula, the similarity score has been calculated as lambda increases, hence the gain score decreases. Only high-scoring branches can survive, get rid of pruning. Lamda value is 1 by default. As Lambda value increases, learning rate will become harder, but overfitting will decrease. Lambda affects only branches that have a single value due to its position in the formula. In other words, the more value in the branch, the lower lambda similarity score.

XGBoost Algorithm shows higher prediction rate using different techniques and is optimized to work on large data sets. The main issues that it diverges from other algorithms are as follows: regularization, pruning, working with null values, system optimization, parallel processing, high flexibility, maximum depth, internal cross validation.

ADABoost Algorithms (Adaptive Boosting): ADABoost algorithm is both the most basic and the simplest among the boosting algorithms. ADABoost algorithm is a method based on decision tree algorithm. Decision tree algorithms have been used since data mining and machine learning methods came out. But in some cases it does not give good results. For this reason, ADABoost Algorithms have been developed. In ADABoost algorithms, random values are called forests in big data. Normally, there is no restriction on the size of the tree. But the AdaBoost algorithm takes its name from a node called stump and two leaved trees. The algorithm steps and advantages of AdaBoost Algorithms are as follows. New nodes definitely perform worse than normal decision trees. This situation is referred to as weak learner. In addition, in large forests where random data are generated, the value (votes) of each tree is equal in random forests. However, the value (rating) of some stumps is greater in AdaBoost. Finally, it does not matter which tree was formed before when evaluating in random size data, namely in random forests. But sorting is important in AdaBoost algorithm. The mistake made in the first situation affects the second situation.

The value that comes out of AdaBoost is made through the following definitions.

$ht(x)$: the value of one of the stumps,

at: shows the weight of the value of that stump.

If the binary classification problem is solved, the weight is determined as in the equation below.

$$\ln[\frac{1 - \text{error of that stump}}{\text{error of that stump}}]$$
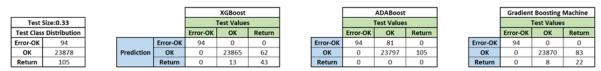
In the AdaBoost algorithm, multiple-row models are created, each of which corrects errors in the latest model. AdaBoost algorithm focuses on mispredicted observations and the next model works to predict accurately these values. The steps to implement the AdaBoost algorithm are as follows. Initially, all observations in the dataset are given equal weight. A model is based on a subset of data. Using this model, predictions have been made for the entire data set. Errors are calculated by comparing predictions and real values. When creating the next model, more weight is given to incorrectly estimated data points. Weights can be determined using the error value. For example, the higher the error, the greater the weight value assigned for observation. This process is repeated until the error function changes or until the maximum limit of the number of estimators is reached.

4. FINDINGS

The most challenging part of data science is data manipulation. While modeling the data, many important operations such as extracting the data, creating correlation matrices, choosing the features according to the correlation between the variables have been performed in this step. Later, "undersampling" and "oversampling" operations were performed as simulations. Afterwards, the separation of training data and test data was done. Then, sensitivity, f-score and complexity matrix values were analyzed as a whole. Later, cross validity analysis, an important step of machine learning, was carried out. The reason for the need for cross validation analysis is based on the foundation of machine learning. In machine learning, some of the data is divided into test and some as training data. If the same data is always divided into test data and training data, the machine will go to memorize. The problem of "overfitting" will arise. Later, the data was manipulated and the dependent and independent variables were separated. As a result of the simulation, undersampling was performed on the data. The extraction and visuality of the attributes related to the correlation matrix of the data are shown in Figure 1. Afterthat, it is aimed to obtain better results by applying Ensemble Machine Learning algorithms that give better results to simulated data. Then, by cross-validating, different data sets were created thanks to the data allocated for testing and training from each data of 72 thousand lines each time. Firstly, in Figure 3., the complexity matrix result is shown without reducing the sampling. Here, first of all, Ensemble Machine Learning techniques were applied, no simulation was performed and the results were examined. In Figure 4, a new model has been established by undersampling and simulation results have been obtained. In Figure 5, Ensemble Machine Learning algorithms were applied by simulating "oversampling" and the accuracy of the model was examined as a result of the complexity matrix.

Figure 3: Complexity Matrix Result Without Reducing Sampling

Confusion Matrix Results Without Undersampling

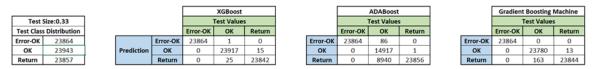| Test Size:0.33					XGBoost					ADABoost					Gradient Boosting Machine		
Test Class Distribution					Test Values					Test Values					Test Values		
Error-OK	94				Error-OK	OK	Return		Error-OK	Error-OK	OK	Return		Error-OK	Error-OK	OK	Return
OK	23878		Prediction	Error-OK	94	0	0		Error-OK	94	81	0		Error-OK	94	0	0
Return	105			OK	0	23865	62		OK	0	23797	105		OK	0	23870	83
				Return	0	13	43		Return	0	0	0		Return	0	8	22

Figure 4: Result of Complexity Matrix by Undersampling

Confusion Matrix Results With Undersampling (Algorithm:Instance Hardness Threshold Method)

Test Size:0.33				XGBoost				ADABoost				Gradient Boosting Machine		
Test Class Distribution				Test Values				Test Values				Test Values		
				Error-OK	OK	Return		Error-OK	OK	Return		Error-OK	OK	Return
Error-OK	101		Error-OK	101	0	0	Error-OK	101	0	0	Error-OK	101	0	0
OK	14203	Prediction	OK	0	14202	6	OK	0	14179	29	OK	0	14203	6
Return	79		Return	0	1	73	Return	0	24	50	Return	0	0	73

Figure 5: Result of Complexity Matrix by Oversampling

Confusion Matrix Results With Oversampling (Algorithm: Borderline SMOTE Method)

Test Size:0.33				XGBoost				ADABoost				Gradient Boosting Machine		
Test Class Distribution				Test Values				Test Values				Test Values		
				Error-OK	OK	Return		Error-OK	OK	Return		Error-OK	OK	Return
Error-OK	23864		Error-OK	23864	1	0	Error-OK	23864	86	0	Error-OK	23864	0	0
OK	23943	Prediction	OK	0	23917	15	OK	0	14917	1	OK	0	23780	13
Return	23857		Return	0	25	23842	Return	0	8940	23856	Return	0	163	23844

5. CONCLUSION

In this study, "Ensemble Machine Learning" methods, which are newly introduced to the literature, have been applied in the manufacturing sector. The purpose of this study is to determine the production line that can create an error that can predict the result when a new firmware data is entered for production. In the literature, mathematically applied Boosting Algorithms (ADABoost, XGBoost, Gradyent Boost), which appeared in 2016, were applied in the literature. Another difference of the study is the simulation dimension. In machine learning, simulation studies are carried out with "undersampling" and "oversampling" algorithms. "Undersampling" application has been referred to the literature as reducing the number of samples. This situation can be stated as follows: The error rate in the data of 72 thousand lines is lower in percent. The prediction results may not be very good. In order to test its accuracy, by "undersampling", that is, decreasing the total data set and keeping the incorrect data set constant, increasing the error rate as a percentage makes an important contribution to the accuracy of the results. It can also be seen from the graph of the final Figure 3 and 4. Later, the same Ensemble Machine Learning algorithms were applied to another simulation methodology, "oversampling". In this process, the prediction algorithm was made by increasing the sample number and error rate. As a result of the simulation, "Undersampling" and "Oversampling" methods were found to give better results than classic data mining techniques. To sort the simulation results in itself, the "Undersampling" methodology gave better results than the "Oversampling" methodology. Three complexity matrix is made in Finding section and results are analyzed better. In this study, products that are "returned", "faulty" and normal are analyzed in the same model with ensemble machine learning algorithms. It can be tried artificial neural network and deep learning algorithms in this model for future studies.

REFERENCES

Adam, E. E., & Ebert, R. J. (1986). Production and Operations Management: Concepts, Models and Behaviour. (3. Edition). UK: Prentice-Hall International Editions.

Afentakis, P., & Gavish, B. (1986). Optimal Lot-sizing Algorithms for Complexproduct Structures. Operations Research , 237–249.

Afentakis, P., Gavish, B., & Karmarkar, U. (1984). Computationally Efficient Optimal Solutions to the Lot-sizing Problem in Multistage Assembly Systems. Management Science, 222–239.

Anthony, R. (1965). Planning and Control Systems: A Framework for Analysis . Cambridge, Mass: Harvard University Press.

Bamford, D., & Wystouri, T. (2005). A Case Study of Service Failure and Recovery within an International Airline. Managing Service Quality: An International Journal, 15(3), 306-322.

Bell, C. R., & Zemke, R. E. (1987). Service Breakdown: The Road to Recovery (Cilt 76). New York.

Butler, J. K. (1991). Toward Understanding and Measuring Conditions of Trust: Evolution of a Conditions of Trust Inventory. Journal of Management, 17(3), 643-663.

Chang, J., Khan, M. A., & Tsai, C.-T. (. (2011). Dining Occasions, Service Failures and Customer Complaint Behaviours: an Empirical Assessment. International Journal of Tourism Research, 6(14).

Chase, R. B., Aquilano, N. J., & Jacobs, F. R. (2001). Operations Management For Competitive Advantage (9. Edition). New York: The McGraw-Hill Companies.

Crowston, W. B., Wagner, M. H., & W. J. (1973). Economic Lot Size Determination in multi Stage Assembly Systems. Management Science, 517–527.

Demir, E., & Dinçer, S. E. (2020). Place and solution proposals of data mining in production planning and control processes: a business application. Pressacademia, 189-193.

Demiröğen, O., & Güzel, D. (2009). Üretim Planlama Ve İş Yükleme Metotları. Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, Cilt: 23, Sayı: 4, 43-67.

Ekmekçi, N. (2015). SANAYİ İŞLETMELERİNDE ÜRETİM PLANLAMASI VE DOĞRUSAL PROGRAMLAMA İLE BİR SANAYİ İŞLETMESİNDE OPTİMİZASYON UYGULAMASI. Konya: T.C. SELÇUK ÜNİVERSİTESİ SOSYAL BİLİMLER ENSTİTÜSÜ İŞLETME ANABİLİM DALI ÜRETİM YÖNETİMİ VE PAZARLAMA BİLİM DALI-YÜKSEK LİSANS TEZİ.

Elmaghraby, S. (1978). The Economic Lot-scheduling Problem (ELSP): Reviews Andextensions. Management Science, 587–598.

Ennew, C., & Schoefer, K. (2003). Service Failure and Service Recovery in Tourism: A Review. The Tourist: A Psychological Perspective.

Eppen, G. D., & Martin, R. (1987). Solving Multi-item Lot-sizing Problems Using Variable Definition. Operations Research, 832–848.

Fleischmann, B. (1990). The Discrete Lotsizing and Scheduling Problem. European Journal of Operational Research, 337–348.

Fleischmann, B. (1994). The Discrete Lotsizing and Scheduling Problem with Sequence-dependent Setup Costs . European Journal of Operational Research, 395-404.

Florian, M., & Klein, M. (1971). Deterministic Production Planning with Concave Costs and Capacity Constraints. Management Science 18, 12–20.

Harris, F. W. (1913). How Many Parts to Make at once . Factory, The Magazine ofManagemen, 135-136.

Hax, A. C., & Meal, H. C. (1975). Hierarchical Integration of Production Planning and Scheduling. M. G. editor içinde, TIMS studies in Management Science (s. Chapter 1). New York: North Holland/American Elsevier.

Heizer, J., & Render, B. (2011). Operations Management (Global Edition / 10. Edition). Kanada: Pearson Publishing.

Hess, R. L., Ganesan, S., & Klein, N. M. (2003). Service Failure and Recovery: The Impact of Relationship Factors on Customer Satisfaction. Journal of the Academy of Marketing Science, 31(2), 125-145.

Hoffman, K. D., & Kelley, S. W. (2003). A CIT Investigation of Servicescape Failures and Associated Recovery Strategies. Journal of Services Marketing, 17(4).

İkizoğlu, S., Demir, E., & Atasoy, B. (2019). Machine Learning Based Feature Extraction for Determination of Balance Disorders.International Statistics Congress (s. 164). Bodrum, Muğla: Turkish Statistical Association.

Jacobs, F. R., & Chase, R. B. (2008). Operations and Supply Management: The Core. Indiana: McGraw-Hill/Irwin.

Jones, M. A., Reynolds, K. E., Mothersbaugh, D. L., & Beatty, S. E. (2007). The Positive and Negative Effects of Switching Costs on Relational Outcomes. Journal of Service Research, 9, 335-355.

Jünger, M., & Naddef, D. (2001). Mathematical Programming Models and Formulations for Deterministic Production Planning Problems. Y. Pochet içinde, Computational Combinatorial Optimization (s. 57-111). Berlin, Heidelberg: Springer.

Karmarkar, U., & Schrage, L. (1985). The Deterministic Dynamic Product Cycling Problem. Operations Research , 326–345.

Krajewski, L. J., Ritzman, L. P., & Malhotra, M. K. (2013). Operations Managemen: Processes and Supply Chains. Üretim Yönetimi: Süreçler ve Tadarik Zincirleri. Çeviri Editörü: Semra Birgün. 9. Baskıdan Çeviri. İstanbul: Nobel Akademik.

Kumar, S. A., & Suresh, N. (2008). Production and Operations Management (With Skill Development, Caselets and Cases) . India: (2. Edition). New Delhi: New Age International..

Levesque, T. J., & McDouggall, G. H. (2000). Service Problems and Recovery Stratégies: An Experiment. Canadian Journal of Administrative Sciences, 17(1), 20-37.

Lewis, B. R., & McCann, P. (2004). Service Failure and Recovery: Evidence from the Hotel Industry. International Journal of Contemporary Hospitality Management, 16(1).

Lewis, B. R., & Spyrakopoulos, S. (2001). Service Failure and Recovery in Retail Bankimg: The Customers' Perspective. International Journal of Bank Marketing, 19(1), 37-48.

Loo, P. T., Boo, H. C., & Khoo-Lattimore, C. (2013). Profiling Service Failure and Customer Online Complaint Motives in the Case of Single Failure and Double Deviation. Journal of Hospitalit Marketing & Management, 22(7).

Lorenzoni, N., & Lewis, B. R. (2004). Service Recovery in the Airline Industry: A cross- Cultural Comparison of the Attitudes and Behavirours of British and Italian Front-Line Personnel. Managing Service Quality: An International Journal, 14(1), 11-25.

Love, S. (1972). A Facilities in Series Inventory Model with Nested Schedules . Management Science, 327–338.

Nguyen, D. T., & McColl-Kennedy, J. R. (2003). Diffusing Customer Anger in Service Recovery: A Conceptual Framework. Australasian Marketing Journal, 11(2), 46-55.

Orlicky, J. (1975). Material Requirements Planning . New York: McGraw-Hill.

Özgen, H. (1987). Üretim Yönetimi. Ankara: Bizim Büro Basımevi.

Roy, R. N. (2005). A Modern Approach To Operations Management. India: New Age International.

Russell, R. S., & Taylor, B. W. (2011). Operations Management: Creating Value Along the Supply Chain (7. Edition). United States: John Wiley.

Salomon, M. (1990). Deterministic Lotsizing Models for Production Planning. PhD Thesis. Erasmus Universiteit Rotterdam, The Netherlands.

Salomon, M., Kroon, L., Kuik, R., & Van Wassenhove, L. (1991). Some Extensions of the Discrete Lotsizing and Scheduling Problem. Management Science, 801–812.

Slack, N., Brandon, -J. A., & Johnston, R. (2013). Operations Management (7. Edition). Kanada: Pearson Publishing.

Stevenson, W. J. (1996). Production / Operations Management (5. Edition). Kanada: Irwin Publishing.

Tekin, M. (2012). Üretim Yönetimi . Konya: Günay Ofset Cilt 1 (Yenilenmiş 8. Baskı).

Trigeiro, W., L.J., T., & McClain, J. (1989). Capacitated Lot Sizing with Setuptimes . Management Science, 353–366.

Veinott, A. (1969). Minimum Concave Cost Solution of Leontief Substitution Models of Multi-facility Inventory Systems. Operations Research , 262–291.

Vollman, T., & Berry, W. L. (1997). Manufacturing Planning and Control Systems . New York: Richard D. Irwin, Third Edition.

Wagner, H. M., & Whitin, T. M. (1958). Dynamic Version of the Economic Lot Size Model. Management Science 5, 89–96.

Weun, S., Beatty, S. E., & Jones, M. A. (2004). The Impact of Service Failure Severity on Service Recovery Evaluations and Post-Recovery Relationships. Journal of Services Marketing, 18(2), 133-146.

Wilson, R. H. (1934). A Scientific Routine for Stock Control. Harvard Business Review 1934, 116-128.

Zangwill, W. I. (1969). A backlogging Model and a Multi-Echelon Model of a Dynamic Economic Lot Size Production System – a Network Approach. Management Science , 506–527.