National College of Ireland

Forecasting the Return Rate of Products in the Textile Industry through Feed Forward Neural Networks

MSc Research Project
Data Analytics

Florian Demir
Student ID: x20216301

School of Computing
National College of Ireland

Supervisor: Mohammed Hasanuzzaman

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Florian Demir
Student ID:	x20216301
Programme:	MSc in Data Analytics **Year:** 2024
Module:	Research Project
Supervisor:	Mohammed Hasanuzzaman
Submission Due Date:	12/12/2024
Project Title:	Forecasting the Return Rate of Products in the Textile Industry through Feed Forward Neural Networks
Word Count:	6125 **Page Count** 14

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Florian Demir
Date:	12/12/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Forecasting the Return Rate of Products in the Textile Industry through Feed Forward Neural Networks

Florian Demir
x20216301@student.ncirl.ie
MSCDA
National College of Ireland

Abstract

The clothing industry appears as a sector that has existed for centuries, developed every year and will never end. The biggest losses in this sector are the cost of returned products to companies and the transportation and storage costs of products. Considering real-life conditions, less return means less waste, and less waste means more profit for companies. To best address these challenges and obtain objective, scientific results, this study introduces a new approach to predicting return rates using Feed-Forward Neural Networks, a new type of machine learning technique. In this study, to achieve optimal results in unbalanced data sets, samples in the minority class were increased with the Synthetic Minority Oversampling Technique, hyperparameter optimization was performed with RandomSearchCV, and a Feed-forward Neural Networks model was established with an accuracy rate of 88.7% to estimate the return rates.

Keywords: Feed-Forward Neural Networks, SMOTE, Predicting Return Rates, Hyperparameter Optimization

1 Introduction

This study aims to address this question: How can Feed-Forward Neural Networks combined with oversampling techniques improve the prediction of return rates in the textile industry using synthetic and historical sales data?

According to the study conducted by Statista, the cost of returned products in the USA to companies working in the textile field has reached 743 million dollars. And the leading products returned in the USA are textile products, with a rate of 24%. Studies in this field have increased significantly, especially after the 2000s. Many methods have been tried in different sectors and data to estimate the return reasons and return rates of products sold in recent years, and reliable results have been obtained. Identifying products that are likely to be returned, especially by using product data in companies' own databases, can enable the manufacturer to reduce the production cost of the products, make stock planning more efficiently and gain an advantage over its competitors. Additionally, the carbon footprints left by companies on the world can also be reduced.

In this academic study I conducted, a model was established to predict the return rates of products through Feed Forward Neural Networks, which combines much more useful and modern techniques compared to classical regression-based predictions, with a success and accuracy rate of 87.8 percent, by using the most up-to-date, advanced machine learning techniques. When the literature is examined, many machine learning-based models, from Lasso regression to Recurrent Neural Networks models, have been established to predict return rates. However, studies conducted to find the reason for return rates have revealed that the biggest factor is demand, return policies and product reviews. In a study, it was emphasized that Recurrent Neural Network methods created using DMLP gave good results, especially in predictions made on time series. [4] In a study, the relationships between the sales and return processes were evaluated using Bayesian estimation techniques and the Poison distribution method during the reproduction

phase of the products to be sold. [8] In a study, a two-stage optimization model was established in the field of Closed Loop Supply Chain, which is based on Adaptive Network-Based Fuzzy Inference System techniques and offers a different perspective combining Neural Network methods with the Fuzzy Logic concept. [9] Urbanke et al. identified products with high return rates in their data sets using Mahalanobis Feature Extraction and Stochastic Gradient Descent, but it turned out that the established model ignored some products with low return rates in the predictions made. [13] In a study, return rates of products were estimated with the XGBoost algorithm and Gradient Boosted Regression Tree algorithms using product photographs for the textile industry [6]In a study conducted by Tuylu et al., the return rates of products were estimated with the M5P algorithm by combining Decision Tree and Multiple Linear regression techniques in order to improve the optimization of stock management and distribution processes. [1] In addition, using the Matrix Factorization method, new features that show customers' body measurements and similarity scores of the textile products to be sold were created and these features were blended with the Bayesian Personalized Ranking and Skip-gram model. [7] In another study, a weighted hybrid model called Hygraph was built using customers' return reasons and purchase data. [15] Gradient Boosting and Random Forest techniques were used to compare with Lasso Regression to estimate return rates. Lasso Regression has been shown to have the best accuracy for large data sets where both variable selections and adjustments are made simultaneously. [5] Another study showed that product returns constitute a significant part of reverse logistics activities in retail.[16] In her latest study, Tuylu tested M5P, M5Rules, Linear Regression+SMOreg algorithms using Stacking and Vote algorithms from EML algorithms and reached a correlation coefficient of 86.07%. [12]

Although some models have been created in the literature for the textile industry using Neural Networks techniques to calculate return rates, these models need photographs of the products sold to make predictions. Considering the research conducted, it has been seen that very few studies have been conducted on imbalanced data sets and this problem has not been addressed with Neural Networks methods. As a result of this study, which went beyond these limits, a model that predicts product returns with a high accuracy rate was established using Feed Forward Neural Networks.

2 Related Work

In this section, I examined the methods used to determine the return rates of sold products by scanning the existing literature and decided that they could be classified under 5 headings, as listed and examined below.

2.1 Regression and Decision Tree Models used to estimate Return Rates

The foundations of machine learning models come from regression models. Although regression-based models have lost their former popularity, they still give very successful results. When the literature was examined, the model with the highest accuracy rate was a Lasso regression model. In a study, M5P algorithms were used to determine the return rates of products and to help companies, especially in stock management, and successful results were obtained. In the study, higher accuracy and lower error values were obtained in complex data sets with the M5P algorithm, which was created by blending Decision Tree and Multiple Linear Regression methods, compared to other established algorithms. In their studies, they also tried Linear Regression Supported Vector Regression and Decision Table methods, and it was determined that the best result was obtained from the M5P algorithm with a correlation coefficient of 82.35%. [2] Another study used the Least Shrinkage Operator method, which is an alternative to the last squares method, to estimate product return rates. In their studies, the machine learning method called Lasso Regression, which has a high accuracy coefficient and is easier to use than other models, was used. Another reason for using Lasso Regression in the study is that the data set used is large and both editing and variable selection can be done simultaneously while creating the model. Additionally, Random Forest and Gradient Boosting methods were tried to compare the results of the study, and it was observed that the Lasso Regression method gave the best result with an R^2 value of 0.918. [5]

2.2 Neural Networks and Advanced Analytics Techniques

In the model created using more than 1 million e-commerce data, 3 different dimension reduction methods called Mahalanobis Feature Extraction, Principal Components Analysis and Linear Discriminant Analysis were used. The

reason for using the Mahalanobis Feature Extraction method in particular is due to its ability to enable the extraction of important features from large and sparse data matrices. At each step of their study, they worked on 1 randomly selected data and reached the optimum point by constantly changing the selected points. The main reason why they chose this method is that the Stochastic Gradient Descent method can provide faster results than the Batch Gradient Descent method. As a result of the study, the Pearson Correlation Coefficient was reached to 0.409, which indicates a moderate positive correlation value. The most important result of the study is that the established model is effective in predicting products with high return rates due to its sensitivity value, but the model overlooks some products when predicting returns due to low recall values. [13] In another study, they estimated the return rates with new features obtained using Gabor Filters and deep learning methods in the model created using photographs of textile products (trousers, cardigans, jackets, etc.). In the study, a total of 6 different models were created and different features were added to each model. While the R^2 value obtained in the predictions made according to the price and category of the products sold in the first model was 33.68%, in the sixth, or last, model, the R^2 value was 46.38%. According to these values, it was observed that there was a 37% performance increase between the first model and the last model. The purpose of the model is to determine the return rates of new products to be launched in the textile industry before they are introduced to the market, and it has also helped decision makers in companies make serious strategic decisions. This study is important for companies to manage their costs and reduce these costs. [6]

2.3 Bayesian and Fuzzy Logic Methods

When the literature was examined, many studies based on Bayesian and Fuzzy Logic Methods were found. A study focused on the return and sales processes, evaluated the relationship between sales data and return data using Bayesian Estimation Techniques, and tried to determine the timing of returned products. The main purpose of the study is to help companies manage their inventories and support manufacturers in the re-production phase. Poison Distribution was used in the study and the results reached a 50 percent higher accuracy rate than other compared methods. In the established model, the Log-Likehood loss function was used in the optimization phase and the Stochastic Gradient iterative method was used in the training phase. The accuracy of the model was evaluated with 0.09 TIC metric. [8] In another study, a model was created using the Adaptive Network-Based Fuzzy Inference System method, which is one of the Fuzzy Logic methods that form the basis of artificial intelligence methods, and it was aimed to minimize future uncertainties by predicting the return rates of the products. The biggest advantage of this method used is both the concept of fuzzy logic and Artificial Intelligence. Since it combines Neural Networks methods, it can benefit from the features of both methods. A two-stage optimization method was used for supply chain optimization and a model was developed to be used in cases where return rates and product return times are uncertain. In the first optimization phase of the model, the number of products returned by the user was estimated using the Adaptive Network-Based Fuzzy Inference System method. In the second stage, the capacity of the production and storage facilities and the capacity of these production and storage facilities were determined by using the product return rate estimates obtained by the Adaptive Network-Based Fuzzy Inference System method. Closed Loop Supply Chain network was used to optimize their location. [9]

2.4 Time Series Prediction through Neural Networks

When the literature was examined, many studies were found that made predictions based on time series. In a study conducted with data consisting of time series, models were created with Neural Network methods, which are more modern than other traditional machine learning methods. In the established models, Feed Forward Neural Networks and Recurrent Neural Networks methods were compared, and it was stated that Recurrent Neural Networks produced better results in predictions made in time series. In the study, it has been shown that when DMLP is used in the model, the Recurrent Neural Network model works better on time-varying data than the Feed Forward Neural Network model, and more accurate results are obtained by using fewer layers. The biggest difference between these two neural networks is that in Recurrent Neural Network models, the information is transferred over time. It is transmitted both forward and backward. In feedforward Neural Network models, information obtained from previous layers cannot be transferred back. [4] In another study, a hybrid model was established using page click data containing customers' purchasing and return behaviour and photographs of textile products to be sold. The Skip-Gram model and the Bayesian Personalized Ranking model were blended in the model created to predict the return rates of textile products

before they go on sale. In addition, the properties of the textile products to be sold were determined by the Matrix Factorization method. Among the new properties determined, the similarity scores between the products to be sold, the dimensions of the products to be sold, there were new features such as the customer's body shape. By incorporating these newly discovered features into the new model, a 74% Precision value and a 34% Recall rate were achieved. [7]

2.5 Factors Affecting Product Returns

In addition to estimating return rates, many studies have attempted to understand why customers return products. In a study, it was observed that product reviews affected the likelihood of product returns by reducing product uncertainty. In the study, a linear connection was established between customers' attitudes towards purchasing and returning products and the accuracy of the information in product comments. When consumers' purchasing behaviour is examined when purchasing a product, it is seen that if the comments about a product are not descriptive, the customer also purchases substitute products along with these products. It has been revealed that customers protect themselves from uncertainties. However, it has been determined that if the evaluations of the products are shown higher than their real value, the return rate of these products also increases.[11] It has also been revealed that product return rates have increased over the years due to dynamic and unbalanced consumer demands in the textile industry, and demand generally directs consumer behaviour.[14] In the literature review on product returns, it was determined that consumer behaviour, product quality, various marketing strategies and return policies are among the most important factors in product returns.[17] The most important point that complicates the problem structure in product returns is uncertainty in demand. The demand forecast is the prediction of product returns. Accurate demand forecasting for returned products provides the company with strategic benefits in many key areas such as production, distribution and stock. Recent studies have shown that artificial intelligence and machine learning methods are more accurate than classical prediction methods in large complex data sets. [12]

3 Research Methodology & Design Specification

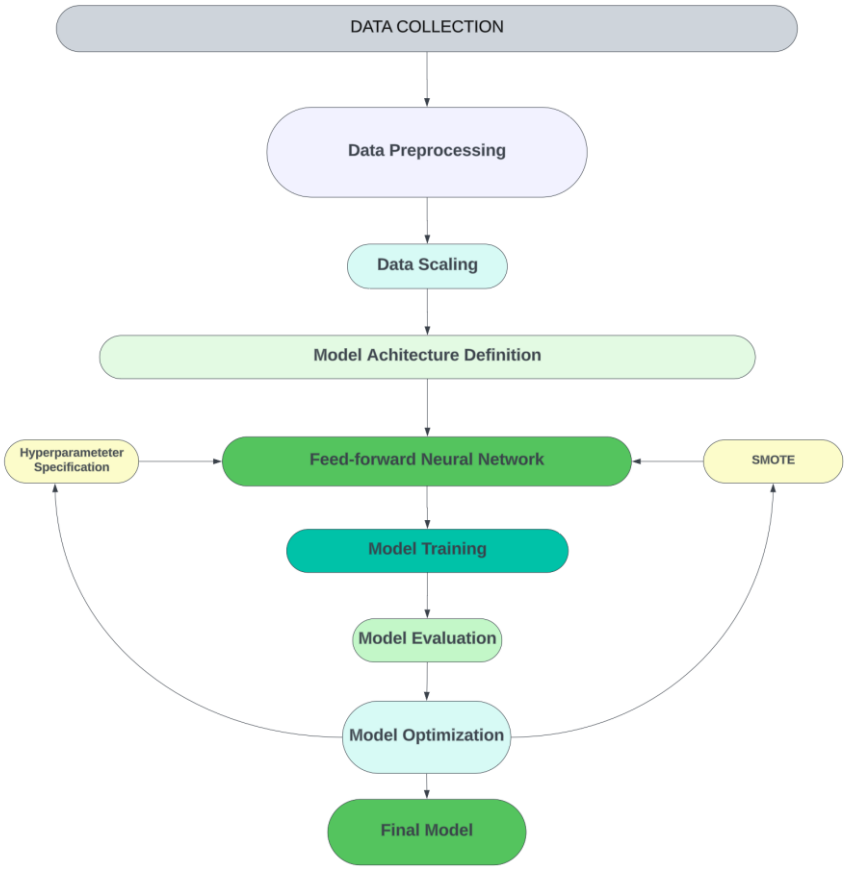

Figure 1: Project Implementation Flow Chart

In this academic study, a Feed-forward Neural Networks model was developed to be used in the textile industry to predict the return rates of the products to be sold. This model uses historical trade sales data to make accurate predictions. The model used features obtained from the data. The main purpose here is to increase customer satisfaction of commercial companies, reduce product return rates, help companies make more profits, reduce companies' logistics costs and help companies with their stock management. In this study, the data was prepared for the model and subjected to pre-processing to make it compatible with the model. As a result of the examinations, the most suitable model was selected, and the focus was on eliminating the class imbalances encountered during the education phase of this model. To achieve the best prediction result, a model with a deep learning base called Feed-forward Neural Networks was chosen. This model used the status of whether the products could be returned or not and the features of the products sold. Synthetic Minority Over-sampling Technique was used to increase the accuracy rate and performance of the Feed-forward Neural Networks model I produced and, most importantly, to eliminate class imbalance in the data set used. With this technique, class weighting has been done to further increase the performance of the model. After all these procedures, the success rate increased from 65% to 88.7%.

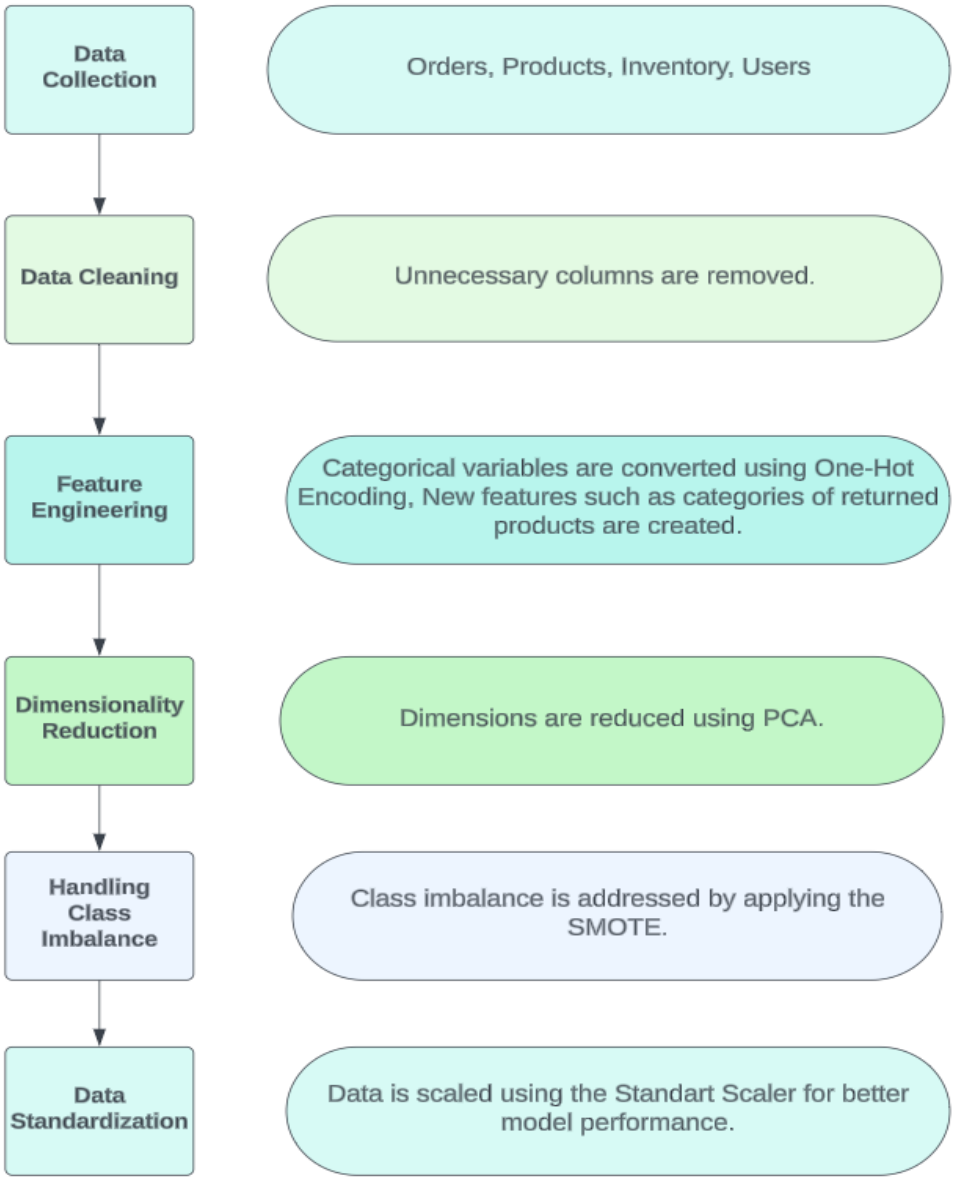

Figure 2: Data Preprocessing Flow Chart

3.1 Datasets

A total of 6 datasets were used in this study. The data used was synthetically produced and made available. The data was combined by determining primary keys. A total of 490705 lines of data were used for this study. While associating the data sets, emphasis was given to the names of the products, user information of those who ordered and inventory information. In the first data analysis, it was seen that 10% of the purchased products were returned and 15% were cancelled after the order. Since the data set was created synthetically, the use of this data does not create any ethical problems. The order items dataset contains 181,759 rows of data. This data set shows the information and features of the ordered products. When using this data set, I focused on the date and time information required for the model and created a new dummy variable from this data set. The orders dataset consists of 125,226 rows in total and contains data showing the name of customers' orders, the time of the orders, the brand, category and prices of the products in the order. The products dataset shows the information of the products to be sold and contains a total of 29,120 rows of data. The distribution centres dataset contains 10 lines of data. This data set shows latitude and longitude data of 10 different distribution centres. The inventory items dataset covers the general features of the products, from when they were sold, to the cost of the product, to the category of the product, to the type and brand of the product. This dataset contains 490,705 rows of data. The users dataset contains a total of 100,000 rows of data showing the personal data of customers who purchased the products.

3.2 Data Pre-processing and Processing

For the Feed-forward Neural Networks model to work better and obtain consistent results, returned products were first identified. I combined the Orders and Order items data sets and combined the information of the customers who bought the product and the product information in another data set. After examining the data after this combination, I cleared some columns that were not important for the model. After the review, I created new dummy variables. A very small number of products did not have either their brand information or their names in the data set. These lines were removed during the data preparation stage. After this stage, other missing data were identified. At this stage, missing numerical data was filled in using the median. Missing categorical data was removed from the data sets in rare cases, and in frequently found cases, the value was filled with the most recurring categorical values. NumPy and Pandas libraries were used to perform these operations. I also created two new features that include the categories of returned products and the brands of returned products. Then, I visualized the two features I created.

3.2.1 Determination of the target variable required for the model

At this stage, missing data was identified, the data set was examined, and a new dummy variable was created in this data set. At this stage, dependent and independent variables are determined. While creating the dependent variable, the “returned at” column was used to assign the value 1 to another column if there was any date in this column, or 0 otherwise. A value of 0 indicates that the sold products were not returned for any reason, and a value of 1 indicates that the sold products were returned for any reason. The main purpose here is to determine the target variable for the Feed-forward Neural Networks model.

3.2.2 Quantification of categorical data for the model

At this stage, the One-Hot Encoding method was used. Variables such as the distribution centre of the product, the category of the product, the platform from which the product was purchased, and the gender of the customers were converted into numerical data with the One-Hot Encoding method.

3.2.3 New feature selection and dimension reduction for the model

Until this stage, the data sets were first cleaned, then the target variable was determined, and then the categorical data was digitized. When the entire data set obtained was re-examined, unnecessary columns that reflected the characteristics of the data were removed, as some columns would slow down the operation of the model. In addition, those that aim to increase the accuracy of the model (for example, the category of the product purchased and returned by the customer) have been determined. Since the data set used contained approximately 500,000 pieces of data, multi-

dimensional data were transformed into lower-dimensional data by using the Principal Component Analysis method, which is a dimension reduction technique. In the feature selection, some of the columns that were not directly related to the selected target variable were removed. In this way, the model showed a faster performance. In addition, the number of samples in the minority class was increased in order to make better predictions in the minority class of the model.

3.2.4 Standardization of data for the model

At this stage, the data to be used was standardized so that the Feed-forward Neural Networks model could learn faster and more robustly than other models. To do this, I used the Standard Scaler method in this study.
After all these data cleaning and data pre-processing stages, a new data set with a total of 158895 lines was obtained.

3.2.5 Dividing the dataset into training and testing datasets for the model

Data set separation was made to ensure that the performance of the model under global conditions gives reliable results. The data set used in the study was divided into two as training and test data sets to be used in the training phase. The training data set resulting from this separation contains 111140 data in total. The test data set has 47632 data. The training data set contains 70% of the data set that has gone through all pre-processing processes, and the test data set contains 30%. I trained the model using the training data set and tested the model with the test data set.

3.3 Data Modelling

3.3.1 Feed-forward Neural Networks

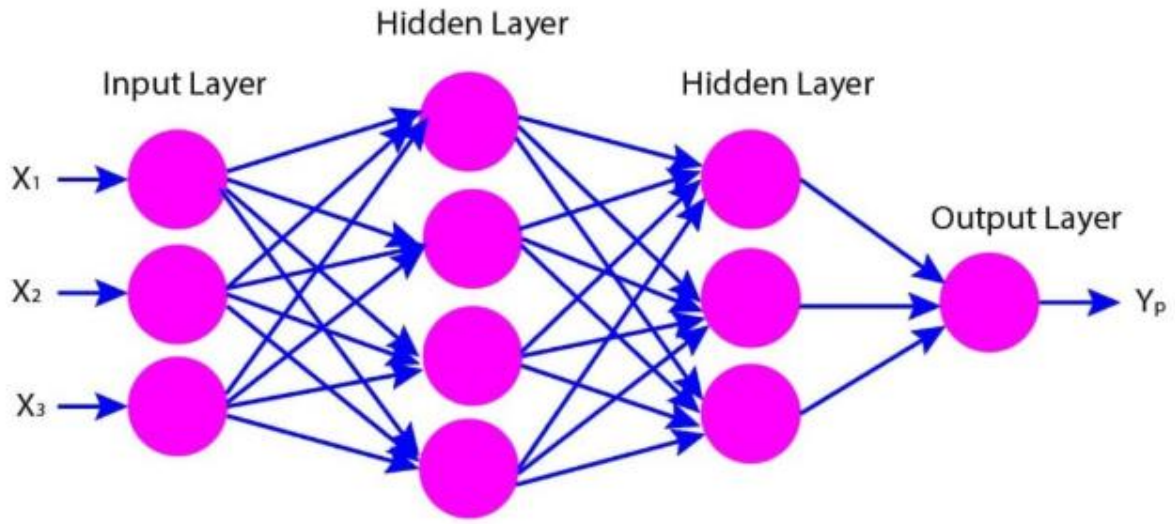

Figure 3: Structure of a Feed-forward Neural Network

There are multiple reasons for using this model. If we want to touch on these briefly. Feed-forward Neural Networks models have proven themselves successfully, especially in tabular data. This model showed better performance when the selected features were independent, regardless of relationships such as Long Short-Term Memory or Convolutional Neural Networks. Additionally, the model is trained faster than other deep learning models. Additionally, the Feed-forward Neural Networks model is an ideal choice for Binary Classification problems. The result aims to find out whether the product to be sold will be returned or not. Here, 0 indicates that the product was not returned and 1 indicates that the product was returned. Therefore, Sigmoid activation function was used in the output layer. Feed-forward Neural Networks were used in this academic study. You can see the mathematical formulas of the methods used during the model below.

$$z(l) = W(l)a(l-1) + b(l)$$

$W(l)$: It refers to the weight matrix in the layer.
$a(l-1)$: Represents the outputs in the previous layer.
$b(l)$: It refers to the bias value in the layer.

$$a(l) = \sigma\big(z(l)\big)$$

σ: It represents the activation function
$z(l)$: It refers to the total entries in the layer.

$$L = -\frac{1}{m}\sum_{i=1} m = [y^i \log(\dot{y}i) + (1 - yi)\log(1 - y^i)]$$

m: It refers to training examples in the training phase
yi: It represents the real value
$\hat{y}_i$: It expresses possibility.

$$y = a(L)$$

$a(L)$: It refers to the activation function located in the last layer.
$y^{\wedge}$: It refers to output

$$\partial z(l)\partial L = \partial a(l)\partial L \odot \sigma'\big(z(l)\big)$$

$\partial a(l)\partial L$: It refers to the derivative of the loss function on the activation function.
$\sigma'\big(z(l)\big)$: It refers to the derivative over the activation function.

$$\partial W(l)\partial L = \partial z(l)\partial L\big(a(l-1)\big)T$$

$\partial z(l)\partial L$: It refers to the error signal in the layer.
$\big(a(l-1)\big)T$: It refers to the transpose of activation functions in the previous layer.

$$\partial b(l)\partial L = \sum i = 1m\partial z(l)\partial L$$

∂z(l)∂L: It refers to the error signal in the layer.
$\sum_{i=1} m$: It refers to the error signals found in all training examples.

$$\partial a(l-1)\partial L = \big(W(l)\big)T\partial z(l)\partial L$$

$\big(W(l)\big)T$: It represents the transpose of the weight matrix.
$\partial z(l)\partial L$: It represents the error signal.

3.3.2 Binary Cross-Entropy Loss

Binary Cross Entropy loss function and Adam optimizer function were used during the optimization of the model.

$$L = -\left(\frac{1}{N}\right)\sum_{i=1}^{N}[y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)]$$

N: It represents the total number of data.
yi: It represents the real value
$\hat{y}_i$: It expresses possibility.

3.3.3 Weighted Binary Cross-Entropy Loss

$$L = -N1i = -1\frac{1}{N}\sum_{i=1}^{N}\left[w1 \cdot y_i \cdot \log\left(\hat{y_i}\right) + w0 \cdot (1 - y_i) \cdot \log\left(1 - \hat{y_i}\right)\right)$$

w_1: It expresses the weight of the minority class.
w_0: It refers to the weight of the dominant class.

3.3.4 Sigmoid Activation Function

The sigmoid function was used in this project because it relates the output of the model to probability values between 0 and 1 in prediction models.

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

3.3.5 Synthetic Minority Oversampling Technique

The Synthetic Minority Over Sampling technique had to be applied to the combined data set. Thanks to this technique, the number of samples in the minority class representing returned products was increased.

$$x_{new} = x_i + \lambda \cdot \left(x_j - x_i\right)$$

$x_j - x_i$: It refers to two samples in the same class.
λ: It represents Lambda.

4 Implementation

Python was used as the programming language in this research. Python libraries NumPy, Pandas, Seaborn, Tensorflow/Keras were used to run the models and analysis. The NumPy library allows us to quickly operate on mathematical data such as arrays and matrices. Pandas library has been a useful library in terms of fast and effective use of data frames during the data pre-processing phase, especially in time series data analysis. The Seaborn library, which has been used specifically to convert data into visuals, has been used to create statistical graphs. This library is based on the MatPotLib library. The reason why this library is used instead of other data visualization libraries is that it is compatible with the Pandas library. TensorFlow/Keras library is a library where the model for machine learning applications will be created and the neural networks it creates will be trained and run. Processes requiring GPU were performed on a system with RTX3080.

The established model has three layers. These layers are the Input Layer, Hidden Layers and Output Layer respectively. The characteristics of the data are selected at the input layer. Dense Count was started at 128 per layer and changes were made to this number based on the model's output performance. Hidden layers were created with the non-linear ReLU activation function to enable the model to learn. The dropout method was used to reduce the overfitting problem. The rate used is 30%. The last layer is the Output Layer, which displays the output of the Feedforward Neural Networks model. This layer shows us the classification result as 0 and 1. The sigmoid function was used in this project because it relates the output of the model to probability values between 0 and 1 in prediction models. Binary Cross Entropy loss function and Adam optimizer function were used during the optimization of the model. The learning rate value was set to 0.001 and the model was tried again by changing it to 0.002 according to the output of the model. While training the model, the class weights of the class containing the returned products were calculated and the model was run by giving higher weight to the class with a small number of returned products. This process enabled the model to learn better. For the same purpose, the class containing a small number of returned products was oversampled using the Synthetic Minority Oversampling Technique method. There was a class imbalance problem in the data sets used in this study. Returned product rate, the target variable used in the data set, represented only 10 percent of total products. That's why the remaining 90% consisted of non-returnable products. While working on such data sets, it has been observed that the model to be built focuses only on the dominant class

of non-returned products, does not care about the returned products or does not make accurate predictions.To overcome this problem, the Synthetic Minority Over Sampling technique had to be applied to the combined data set. Thanks to this technique, the number of samples in the minority class representing returned products was increased. In this way, the model more accurately predicted the products returned by the customer. The new data added with this technique was used in the training phase of the model to enable the Feed-forward Neural Networks model to make a more balanced classification. The model was trained under these conditions for 5, 10, 20 and 50 epochs. Synthetic Minority Over Sampling Technique and class weight parameters were used to eliminate the class imbalance problem during training phase.

Hyperparameter optimization was performed on the model to reach the final model. For this, Grid SearchCV and RandomSearchCV techniques were used to provide an extra contribution to the accuracy of the model. As a result of hyper parameter optimization, the best hyper parameters were found to be 0.01, best dropout rate was found to be 0.2, best epochs were found to be 10, and best batch size was found to be 32, respectively. The accuracy measure was used as the measure of success of the created model. The main purpose of using this metric is to use the success metric as an accuracy metric in studies on this subject and to enable us to reach more accurate results when compared to other studies.

5 Evaluation

The accuracy rate obtained from the techniques and training and validation sets increased from 69% to 88.7%. As a result of hyper parameter optimization, the best hyper parameters were found to be 0.01, best dropout rate was found to be 0.2, best epochs were found to be 10, and best batch size was found to be 32, respectively. Additionally, the success of the result was checked and the Feed-forward Neural Networks model gave consistent accurate outputs with low losses (0.352) in all sets (training data set and test data set). I would especially like to point out that working with unbalanced data sets should not be considered as a problem in this academic study. Considering real life, no company would expect half of its products to be returned. For this reason, comparisons were made only with models that performed better in unbalanced data sets.

Model	**Accuracy**
Feed Forward Neural Network	88.7%
Random Forest	86,7%
ADASYN+ CatBoost	15%
Voting (FNN +LGBM)	88%
Smote + LightGBM	23%
XGBoost	40%
Gradient Boosting	86.4%
XGBoost+Smote	67%
Logistic Regression	49%
Decision Tree + Smote	62%
Naïve Bayes	59%
Linear Discriminant Analysis	49%
Ridge Classifier	49%

Table 1: Comparison of Accuracy Rates of Models.

Best Hyperparameters	**Values**
Best learning rate	0.01
Best dropout rate	0.2
Best epochs	10
Best batch size	32

Table 2: Best Hyperparameters

6 Conclusion and Future Work

As a result of the data analysis, it was observed that 10 percent of the orders in the data set were returned, and 15 percent were cancelled after they were ordered. Although we do not know why the products were returned or cancelled, the Feed Forward Neural Networks model was established with the help of the Synthetic Minority Over Sampling Technique, and the success rate obtained from the techniques and training and validation sets increased from 67% to 88.7 percent. Also, the success of the result was checked, and the Feed-forward Neural Networks model gave consistent accurate outputs with low losses in all data sets (training data set and validation data set).
After this study, it was revealed that Feed Forward Neural Networks together with Synthetic Minority Over Sampling Technique gave strong results to predict the return rates of products, especially in textile data. Being able to predict whether a product will be returned will significantly reduce costs such as transportation, production and stocking. Thanks to the established model, companies can reduce their costs and increase their profits.

In the next stage, it is still unknown how the data to be analysed is not synthetic and what results the models will give, especially with data taken from seasonal discount periods (Black Friday, Christmas, etc.). This model can also be adapted within different sectors. Research on this subject can be continued.

References

[1] Adıgüzel Tüylü, A.N. and Eroğlu, E., 2019. 'Alphanumeric Journal: an exploration of operations research, statistics, econometrics and management information systems', The Journal of Operations Research, Statistics, Econometrics and Management Infor mation Systems, 7(1).\

[2] Adiguzel Tuylu, A.N and Eroglu, E., 2019. 'Using aMachine Learning Algoriths for Forecasting Rate of Return Product In Reverse Logistics Process', Alphanumeric Journal, 7(1), pp.143-156

[3] Agrawal, S., Singh, K.R. and Murtaza, Q., 2014. 'Forecasting product returns for recycling in the Indian electronics industry', Journal of Advances in Management Research, 11(1), pp.102-114.

[4] Brezak, D., Bacek, T., Majetic, D., Kasac, J. and Novakovic, B., 2012. 'A comparison of feed-forward and recurrent neural networks in time series forecasting', in: 2012 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), pp. 1-6, IEEE.

[5] Cui, H., Rajagopalan, S. and Ward, A.R., 2020. 'Predicting product return volume using machine learning methods', European Journal of Operational Research, 281, pp.612-627. Available at: https://doi.org/10.1016/j.ejor.2019.05.046.

[6] Dzyabura, D., El Kihal, S. ve Ibragimov, M., 2018. 'Leveraging the power of images in predicting product return rates', SSRN Electronic Journal.

[7] Kedia, S., Madan, M. ve Borar, S., 2019. 'Early Bird Catches the Worm: Predicting Returns Even Before Purchase in Fashion E-commerce', CoRR, abs/1906.12128. Available at: https://arxiv.org/abs/1906.12128.

[8] Krapp, M., Nebel, J. ve Sahamie, R., 2013. 'Forecasting product returns in closed-loop supply chains', International Journal of Physical Distribution & Logistics Management, 43(8), pp.614-637.

[9] Kumar, D.T., Soleimani, H. ve Kannan, G., 2014. 'Forecasting return products in an integrated forward/reverse supply chain utilizing an ANFIS', International Journal of Applied Mathematics and Computer Science, 24(3), pp.669-682.

[10] Saraswati, D., Sari, D.K., Puspitasari, F. ve Amalia, F., 2023. 'Forecasting product returns using artificial neural network for remanufacturing processes', in: AIP Conference Proceedings, 2485(1), AIP Publishing.

[11] Sahoo, N., Dellarocas, C. and S., 2018. 'The impact of online product reviews of product returns', Information Systems Research, 29(3), pp.723-738

[12] Tuylu A.N.A and Eroglu, E., 2022. 'The prediction of product return rates with ensemble machine learning algorithms', Journal of Engineering Research.

[13] Urbanke, P., Kranz, J. ve Kolbe, L., 2015. 'Predicting product returns in e-commerce: the contribution of Mahalanobis feature extraction',

[14] Wen, X,. Choi, T.M and Chung, S.H., 2019. 'Fashion retail supply chain management: A review of operational models', International Journal of Production Economics, 207, pp.34-55.

[15] Zhu, Y., Li, J., He, J., Quanz, B.L and Deshpande, A.A., 2018. 'A Local Algorithm for Product Return Prediction in E-Commerce', in:IJCAI, pp.3718-3724

[16] Ambilkar, P., Dohale, V., Gunasekaran, A. and Bilolikar, V. 2022. 'Product returns management: a comprehensive review and future research agenda', International Journal of Production Research, 60(12), pp.3920-3944

[17] Ma, S. and Wang, W., 2024. 'Proactive Return Prediction in Online Fashion Retail Using Heterogeneous Graph Neural Networks', Electronics, 13(7), p.1398

[18] Duong, Q.H., Zhou, L., Meng, M., Nguyen, T.V., leromonachou, P. and Nguyen D.T., 'Understanding product returns: A systematic literature review using machine learning and bibliometric analysis'.

[19] Pei, Z. and Paswan, A., 2018. 'Consumers' legitimate and opportunistic product return behaviours in online shopping'. Journal of Electronic Commerce Research, 19, pp.301-319