FACTORS AFFECTING PERFORMANCE RATE IN POSTAL SECTOR

Case: Posti Group Oyj

Lappeenranta–Lahti University of Technology LUT

Master of Science in Technology, Master’s thesis

2022

Aleksandra Perekhozhikh

Examiners: Professor Pasi Luukka
Post-doc Researcher Jyrki Savolainen

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Engineering Science

Computational Engineering (Business Analytics)

Aleksandra Perekhozhikh

Factors affecting performance rate in postal sector

Case: Posti Group Oyj

Master's thesis

2022

52 pages, 10 figures, 13 tables and 2 appendices

Examiner(s): Professor Pasi Luukka and Post-doc Researcher Jyrki Savolainen

Keywords: postal service, delivery performance, random forest regression, generalised linear regression

This study focuses on delivery performance in postal services by quantitatively exploring data of operational measurements. The goal is to lay out the most important factors that affect delivery performance. The data is provided by Postal Service reporting team, Posti Group Oyj.

First, the thesis presents the theoretical background of machine learning usage and algorithms that were used to get information about factors that affect the most on delivery performance rate. The data of this study was analysed using two regression models (Generalized Linear Regression and Random Forest Regression models) in order to have models with good interpretation possibilities. The models were evaluated by RMSE and R2 error metrics. Next, the thesis describes the factors that have been observed to have affect the work performance in the past literature: training, information and communication technologies, infrastructure level, automation level, supply chain complexity, health, and work-related factors (age, type of contract).

The identifies key factors as: sickness rate, difference of planned and actual volume, difference of planned and actual number of employees, indoor efficiency, percentage of overwork hours, route master application rate. Both regression model types applied in this study produced near to similar results. However, the models had low percentage of described variance and relatively high RMSE that indicated the need of further exploring other factors that might have an impact on the delivery performance rate.

ACKNOWLEDGEMENTS

First, I would like to thank LUT University for the great opportunity to study master's degree in Business Analytics (Computational Engineering). My deepest gratitude to Post-doc Researcher Jyrki Savolainen for his essential recommendations and guidance during master thesis writing.

I also want to express gratitude to my colleagues at Posti – for granting the permission to use the data, assistance during the whole process and support. My sincere thanks goes especially to Jäntti Harri, Luomajoki Pekka and Ehtamo Jyri.

Finally, I would like to thank my family and my friends who always believe in me and support me a lot.

Aleksandra Perekhozhikh

19.06.2022

ABBREVIATIONS

BD	Basic Delivery
ICT	Information and Communication Technology
KNN	K-Nearest Neighbour Algorithm
ML	Machine Learning
MSE	Mean squared error
Posti	Posti Group Oyj
PR	Performance Rate
R2	R-squared coefficient
RMSE	Root-mean-square deviation

Abstract

(Acknowledgements)

(Symbols and abbreviations)

Table of Contents

Appendices

1 Introduction

1.1 Background and Motivation

Data is an important asset that allows companies to get insights about customer needs and operational processes hence improve services and be more competitive. The information technologies are developing continuously that leads to changes of business processes as well as generated data volume increase. While volume of the collected data is increasing, the expectation around data is also increasing (Mahanti, 2021). This thesis focuses on exploring factors that affect the performance rate (PR) in the postal sector. Tracking a measure independently does not give the opportunity to find dependencies with other variables. Therefore, to get maximum usage of data, it is important to find the impact of these factors on PR (delivery rate of the items).

The postal chains operate on enormous amount of data and for a company operating in this industry, it is crucial to find the best way to use data. Historically, Finnish postal service company Posti Group Oyj (Posti) has followed technology transformations: from delivering mail by rowing boat in 1638 to building digital, web-based application OmaPosti. The main service provided by postal sector is still delivering letters and parcels to customers. The postal chain can be described by four steps: clearance (collecting items and delivery to distribution center), sorting, intercity transport (delivery to the closest to customer distribution center), and delivery (Parcu, Brennan, Glass, 2020) where each next step in the chain depends on the result of previous step.

The motivation behind this research came from Postal Service reporting team of Posti who highlighted that they need to know the most important factors to be able to reallocate the resources for their development. Maritan, Lee (2017) emphasized the need for a “resource budget” and considered resource allocation as an essential element of a strategic plan. Well planned resource allocation allows a company to efficiently use their assets. The result of this study will benefit Posti as it will get to know the factors that positivity and negatively affect the performance so that they can further develop the factors that has positive affect and investigate into negative factors and take preventive measures.

1.2 Research questions and limitations

This study is aimed towards addressing the following research questions:

1. How and which machine learning methods are applied to identify the most important variables in datasets and are they applied in postal service data?
2. What factor from the dataset affect the delivery performance rate the most in the case of Posti and which machine learning model is the most suitable for identifying relationship between performance rate and those factors?

The study is focusing on exploring already created performance measures and defining most important factors from the list affecting on performance. Therefore, the ways of defining and calculating the performance measure in postal sector are out of the thesis scope. As the second limitation, the data quality can be highlighted: the provided data contains already aggregated measures that were extracted from different sources. This transferring process could lead to data corruption. As the result, it cannot be guaranteed that the dataset presents the real picture of business processes in Posti.

1.3 Data and methodology

The data for this thesis was extracted from PowerBI report that contained information regarding different touchpoints of postal chain and it was provided by Postal Service reporting team at Posti. The analysis was done using Python 3.10.

Data preprocessing steps such as data cleaning, data normalization, data transformation, missing values imputation, and noise identification were done to adapt data to algorithms. Missing values were filled using KNN imputation method, ordinal and nominal variables were encoded using ordinal and one-hot encoding respectively. Described data manipulations were done using following libraries: NumPy, pandas, ProfileReport (pandas_profiling), KNNImputer (sklearn.impute), MinMaxScaler, OneHotEncoder (sklearn.preprocessing). For data visualisation, matplotlib and seaborn python packages as well as PowerBI were used.

For the analysis two supervised machine learning algorithms were applied to the tabular dataset: Generalized Linear Regression model and Random Forest Regression model. The models were evaluated by two error metrics: Root Mean Squared Error (RMSE) and coefficient of determination (R2). To get information about the most important factors, features´ coefficients (Generalized Liner Regression model) and features´ importance (Random Forest regression model) values were used. The following libraries were used to build models and evaluate: RandomForestRegressor (sklearn.ensemble), LinearRegression (sklearn.linear_model), mean_squared_error and r2_score (sklearn.metrics), train_test_split (sklearn.model_selection).

1.4 Thesis Structure

The first section briefly describes the research background and questions to be explored. The second section defines ML models and justifies the selection of these models. This section also describes the theoretical background of the algorithms used. The third section contains literature review. It gives definition to postal service and delivery performance. Next, the section presents the factors that were observed to have affected the work performance in the past literature. The fourth section describes implementation part of the research which consists of dataset description, preprocessing steps, performance, and evaluation of ML models. The fifth section presents the research results and implications for the company. The sixth section presents main findings by answering on the research questions as well as ideas for further studies.

2 Theoretical Background

This section presents theoretical background of the methodologies used in the implementation part of the study as well as the rationale for those choices. First, the overview of the fundamentals of machine learning (ML) is described. Secondly, the data preprocessing steps are described in detail. Next, two regression models (Generalized Linear and Random Forest Regression models) are introduced. Finally, the selection of error metrics is reported: RMSE and R2.

Ronchetti (1997) stated that the model selection is essential when performing an analysis. Therefore, when choosing the models for exploring the factors affecting the performance rate, following factors were considered: type of problem, simplicity of model interpretation and usage for business side as well as constraints regarding the target variable.

2.1 Machine Learning

During recent years, ML research made the drastic progress on solving complex problems due to availability of large amount of generated data and computer power (Rebala, Ravi, Churiwala, 2019). Alloghani, Al-Jumeily, Mustafina, Hussain, Aljaaf (2019) define ML as a component of artificial intelligence that involves learning of hidden patterns in historical data and as a result, perform predictions for the new events.

Supervised and unsupervised learning

ML algorithms can be divided into two groups: supervised and unsupervised algorithms. The supervised machine learning algorithms request labeled data, and on the other hand unsupervised ones use unlabeled data when one is solving an analytical task (Alloghani et al., 2019). Two classical tasks that belong to supervised learning algorithms are classification and regression. Classification task requests target variable in a form of finite and categorical value whereas regression works with numerical target variable. The commonly used technique in unsupervised learning is clustering where only input data without labels is available (García, Luengo, Herrera, 2015; Rebala et al, 2019). Posti provided labeled dataset to get insights about PR. It means that the study is focusing on supervised ML techniques.

As an example of supervised learning, prediction of person's weight (target variable, dependent variable) by observing the information about height, gender, activity level and illnesses (independent variables) could be considered. In this scenario, the model is to be trained on labeled dataset that contained both independent and dependent variables. The performance of the model is to be measured by comparing predicted and true values. In unsupervised learning task, the clustering of customers having different buying behavior could be considered. In this case, there is no ground truth target variable as model is to be trained on unlabeled data.

Machine learning process

García et al. (2015) defined machine learning process in the following steps: problem specification, problem understanding, data preprocessing, data mining, evaluation and result exploration. Figure 1 presents the process visually. Next, the steps are shortly described.

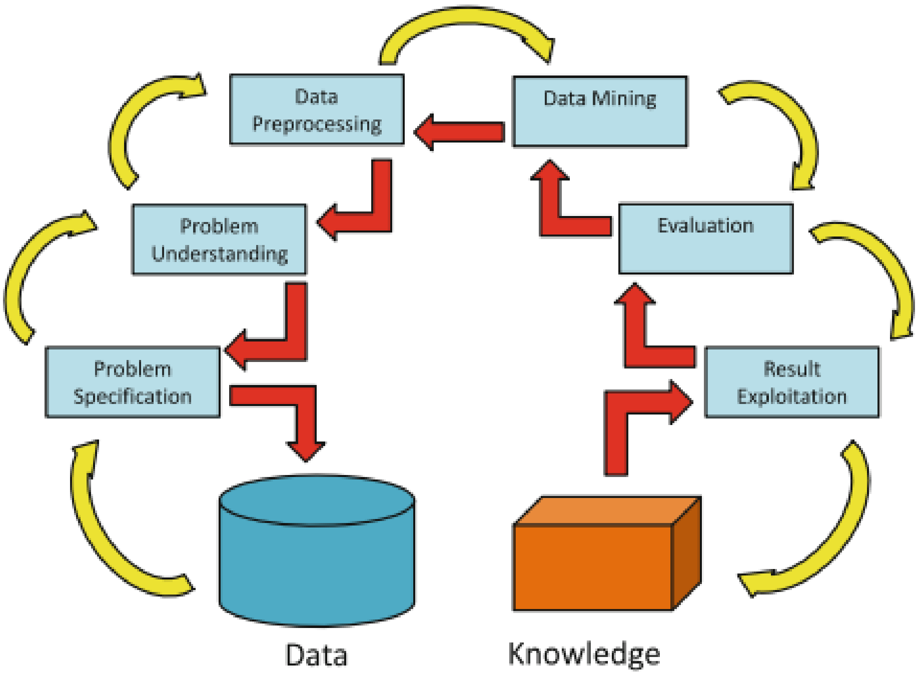

Figure 1. Data Mining Process, García et al. (2015)

The problem specification and understanding steps involves close cooperation with business side to design the application domain and get the expert knowledge about the selected data. The data preprocessing step involves transforming data in an appropriate form for specific data mining task. The data mining step involves running the algorithm for solving the data mining task. Evaluation step involves interpreting the mined patterns based on the measures.

The last step i.e., result exploration involves using the obtained knowledge from the data directly (García et al., 2015).

2.2 Data Preprocessing

Data preprocessing step is highlighted as a powerful step as it adapts data to algorithm to get the best performance. Data preprocessing involves data cleaning, data normalization, data transformation, missing values imputation, data integration and noise identification (García, Ramírez-Gallego, Luengo, Benítez and Herrera, 2016).

2.2.1 Feature Encoding

In many cases, dataset contains both types of variables: numerical and categorical. The categorical variables could be divided into two categories: nominal and ordinal. Nominal variables describe labels that do not have particular order, i.e., variable values that could not be compared with each other. Ordinal variables, on the contrary, have defined order, but in contrast to numerical variables they are not quantitative. Consequently, mode is the only descriptive statistics are not defined for nominal variables, whereas for ordinal variables it is also possible to calculate median (middle value) along with the mode (Kim, Hong, 2017).

Machine learning models primarily work with data consisted of real numbers; hence transformation of categorical variables is required. Ordinal variable could be trivially mapped onto [1, …, N] subset of whole numbers with preserving model interpretability. For nominal variables, such mapping is not appropriate, because it implicitly induces order between values of nominal variables. For dealing with that issue, one-hot encoding is a suitable method. It replaces a nominal variable with k binary variables, where k is the number of distinct categorical values that were assigned to chosen nominal variable (Hardy, 1993). As the main disadvantage of using this method, the increase of predictors is stated (Kim, Hong, 2017).

2.2.2 Data Normalising

The normalisation is used to avoid overweighing for attributes with large range when calculating distance measure. To “eliminate” the unit of measurement, the feature can be rescaled to [0, 1] range where 0 and 1 are assigned to the minimum and maximum values respectively (min-max normalisation). For rescaling feature v to $[0\ ,1]$ range, the formula should be applied:

$$v' = \frac{v - min_A}{max_A - min_A} \qquad (1)$$

where max_A and max_A are original minimum and maximum values of the feature (Shalabi, Shaaban, Kasasbeh, 2006).

2.2.3 Finding Redundant Attributes

García (2015) defined redundant attributes as attributes that can be derived from other attributes. Having those attributes in dataset can lead to overtrained model. Redundant attributes can be detected by correlation test. García (2015) defined Pearson’s product moment coefficient as the most popular correlation coefficient that can be calculated by the following formula:

$$r_{A,B} = \frac{\sum_{i=1}^{m}(a_i - \bar{A})((a_i - \bar{B})}{m\sigma_A\sigma_B} \qquad (2)$$

where m is number of instances, a_i and b_i are attributes values of A and B, $\bar{A}$ and $\bar{B}$ the mean values of A and B, σ_A and σ_B are the standard deviations from A and B.

Asuero, Sayago, González (2006) described the interpretation of correlation value in a scale from little if any correlation to very high correlation (Table 1).

Table 1. Correlation value interpretation

Correlation value	**Interpretation**
0.90 to 1.00	Very high correlation
0.70 to 0.89	High correlation
0.50 to 0.69	Moderate correlation
0.30 to 0.49	Low correlation
0.00 to 0.29	Little if any correlation

García (2015) mentioned that highly correlated features should be removed from the dataset to avoid building unreliable model.

2.2.4 Missing Values Handling Methods

In practice, collected data for analysis in most cases contains some number of missing values due to unavailability of an information (Faisal, 2018). As presence of missing values are not accepted by most analysis methods, these values should be handled by using appropriate method (García, 2015).

There are two traditional methods of dealing with missing values: deletion method and missing value imputation. First, deletion method is the simple but an unsafe approach especially when missing data are not randomly distributed. The method involves ignoring missing data by deleting records that have at least one missing value. When the number of missing values is high, usage of the method can lead to loss of valuable information and selection bias (Lin, Tsai, 2019).

Second, the imputation method involves replacing missing values by some estimated values. The method performs better than previous method as it returns the complete dataset with true characteristics to some extent. Values can be calculated using one of two types of imputation techniques: statistical such as mean/mode and regression, and machine learning based techniques such as k-nearest neighbours (KNN), artificial neural network and support vector machine (Lin, Tsai, 2019).

K-nearest neighbours

In comparison with statistical procedures, machine learning based techniques provides more accurate imputation of missing values (Faisal, 2018). Therefore, KNN which involves building predictive model to estimate missing value is explored.

In this research, KNN is chosen as the imputation method due to its advantages such as preservation of the original data structure as well as opportunity to do the estimations for qualitative and quantitative data. It is important to be aware that missing attributes are not considered when the distance calculation is done.

The simplified imputation process of KNN is described in Figure 2.

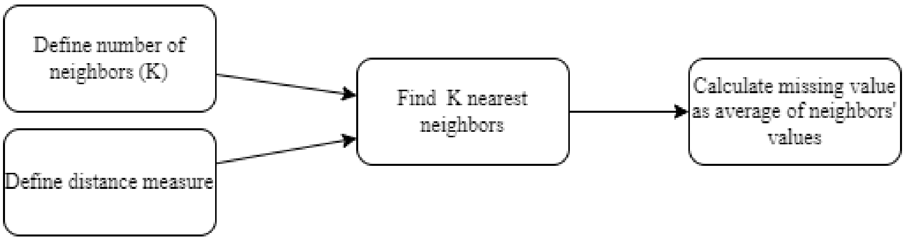
Define number of neighbors (K)
Define distance measure
Find K nearest neighbors
Calculate missing value as average of neighbors' values

Figure 2. Missing value imputation using process influenced by Faisal (2018)

The Euclidean distance is the usual distance measure (Lin et al, 2019) and can be calculated by the following formula (Faisal, 2018):

$$d_q(x_i, x_j) = \left[\frac{1}{a_{ij}}\sum_{s=1}^{p}\left|x_{is} - x_{js}\right|^q I(o_{is} = 1)I(o_{js} = 1)\right]^{1/q} \quad (3)$$

where a_{ij} is number of considered attributers, $I(o_{is} = 1)$ and $I(o_{js} = 1)$ imply attributes that are not missing in x_i and x_j.

Number of neighbours can be defined by researcher or through cross-validation. After finding k closest neighbours based on the closest distance, the missing value is replaced by feature average value of those neighbours (Faisal, 2018).

2.3 Data Mining and Evaluation

2.3.1 Data mining

The data mining steps means building the machine learning algorithm. In supervised machine learning, the dataset is split into training, validation and testing datasets to avoid model overfitting and get well-generalized model. The training dataset is used to train the model, validation dataset allows to set model parameters whereas testing set is used for model evaluation (Xu, Goodacre, 2018). The process of metrics' selection for model assessment is described later in the section.

2.3.2 Selection of Error Metrics

Different machine learning models perform differently when working on different dataset: there is no one best model that works for every dataset. To be able to choose the most accurate model for the specific dataset, each model should be evaluated using some measure. This part of the thesis describes choice of error metrics that are used for assessing the models.

James, Witten, Hastie, Tibshirani (2021) defined mean squared error (MSE) as a commonly used measure for assessing the quality of model fit in regression tasks. The measure can be interpreted in the following way: small value demonstrates that the predicted values are close to true values whereas large value shows significant difference between predicted and true values. When comparing models, the one with lower MSE value performs better.

MSE is calculated using by the formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad (4)$$

where $\hat{f}(x_i)$ is predicted value and y_i is true value for observation i.

To get the error measure in the same unit as target variable (not square of target variable), Root Mean Squared Error (RMSE) is used. RMSE is calculated as square of MSE.

At the same time, it is stated that using only one metric does not show full picture of the model performance. The coefficient of determination (R2) is defined as an informative and truthful measure with no limitations in interpretability (Chicco, Warrens, Jurman, 2017). R2 shows the proportion of deviation that is “explained” by the model and gets value from 0 to 1 where 1 is the best achievable result (Nagelkerke, 1991).

As a result, two commonly used error measures are chosen for model assessment: RMSE and R2.

2.4 Regression

2.4.1 Linear Regression

When it comes to modelling of a response variable in a domain of real numbers, linear regression is a comprehensive yet straightforward approach. Gupta, Sehgal (2021) stated linear regression as one of the most interpretable and easy explained algorithms. Having n samples of k predictor variables $(x_1, \dots, x_k)_i$ and corresponding response variables y_i, linear regression model is given by

$$y_i = w^T x_i + b \qquad (5)$$

where w is a vector of weights of length k, and b is an additive bias factor. Such model has good interpretability: for example, keeping other variables constant, an increase of a predictor variable x_i by m increases the output of a model by $w_i * m$.

The definition of linear regression makes it convenient to transfer it to matrix form. In this case, let X be a n by $k+1$ matrix representing our dataset, where the last column consists of ones, and let w be a vertical weight vector of length $k+1$ (including bias factor). Then, the same model in a matrix form for our dataset is:

$$y = Xw \qquad (6)$$

The optimal solution for weights w is found by minimizing $E(y|x)$, the expected value of response variable conditional on predictor variables, which is in turn is equal to minimizing squared errors given by

$$Q(w, X) = \left|\left|Xw - y\right|\right|^2 \qquad (7)$$

This solution for this optimization problem can be found via iterative optimization methods as well as analytically (Davison, 2003).

The main obstacle for using linear regression in our study is the type of the response variable. Ordinary linear regression does not take account of a possible range of values for the response variable, which in our case is bounded in [0, 1]. While certain techniques (e.g. clipping) make it possible for outputting proper bounded values, it prevents the model from generalizing on unseen data observations.

2.4.2 Generalized Linear Model

Generalized linear models are a family of models that could handle response variables that are neither quantitative nor qualitative. In our study the response variable takes values inside the unit interval [0, 1]. Similar to linear regression described above, generalized linear models attempt to model the mean of response variable as a function of predictor variables. However, in order to enforce various limitations on the response variable, a custom link function g is defined for generalized linear models (James, 2021):

$$g\big(E(y|x)\big) = w^T x + b \qquad (8)$$

Considering the domain of the response variable, the choice for link function g is a logit function (Crawley, 2012):

$$g(x) = \log\frac{p}{1-p} \qquad (9)$$

This function is also called log odds (Norton, Dowd, 2018). To conclude, the learning process is similar to the standard linear regression problem, however in the described setting we try to model not the original response variable, but the transformed one from the link function. Additionally, to obtain a prediction for the original response variable, the inverse function g^{-1} is applied. For logit function, the inverse function is the logistic function:

$$g^{-1}(x) = \frac{1}{1+e^{-x}} \qquad (10)$$

2.5 Tree-based Model

This subsection presents the theoretical knowledge about random forest as a tree-based method. First, the definition and fitting procedure of decision tree are described. Next, the subsection describes random forest model.

2.5.1 Decision Tree

Decision tree is a supervised non-linear model that constructs a tree structure (similar to flow chart) and predicts an outcome by traversing through tree nodes: from top (root) to the bottom (leaf). Each node consists of a logical statement (rule), typically of a form $[x_i \leq t]$

where x_i is a value of ith feature variable, t is some threshold value that is chosen during a decision tree fitting procedure. Depending on the outcome of a given logical statement, the next node is considered, finally until the leaf o a tree is observed (Maimon, Rokach, 2005).

Decision trees are much more suitable for datasets with non-linear relationships, because in contrast to the linear regression models, decision trees construct a piecewise decision function, which are able to approximate non-linear dependencies without extensive feature engineering. Furthermore, decision trees can process datasets with categorical features by building node rules that consist of compare operations (similar to equality operator in various programming languages) (Wilkinson, 1992).

Decision trees can solve both regression and classification problems and can be visualised that makes their presentation and interpretation easy. At the same time, building one tree to solve regression task is not robust approach. It is noted that minor change is data can affect significant on the estimated tree. Using random forest method improves the model performance as well as robustness (James et al., 2021).

Decision Tree Fitting procedure

Decision tree fitting procedure in their study. Before defining the decision tree fitting algorithm, it is crucial to establish the appropriate node impurity measure Q. It indicates the degree of homogeneity of the labels in a particular node. Since all optimization methods are based on minimizing some measure, the impurity measure is set to be minimized as well.

Decision trees are built in a greedy manner. That is, at each step of decision tree fitting, the algorithm tries to find such parameters of the split so that the impurity measure is minimized:

$$Q(X_m, i, t) \rightarrow \min_{i,t} \qquad (11)$$

where i and t represent the parameter if chosen split that is defined as $[x_i \leq t]$ logical statement and X_m is a subset of training samples X. After the best split parameters are found, the node is split into two descendant nodes, left and right. Each descendant node contains subset of training data samples according to a found split criterion (Podgorelec, Kokol, Stiglic, Rozman, 2002).

The defined procedure is set to repeat until the end of the tree, i.e. finding tree leaves that contain training samples with the same value of response variable, or some predefined stopping criterion, for instance reaching some limit in tree depth.

Since the study concerns regression task, the impurity measure for a tree node is defined as a weighted sum of variance in each child node after splitting node of X_m. Weights are proportional to the number of samples in each child node and normalized so that their sum is equal to one:

$$Q(X_m, i, t) = \frac{|\mathrm{X_{left}}|}{|\mathrm{X_m}|} Var(X_{left}) + \frac{|X_{right}|}{|X_m|} Var(X_{right}) \qquad (12)$$

where X_{left} and X_{right} are subsets from child nodes of X_m after splitting (Loh, 2014).

It could be observed that having a perfect split would result in impurity measure equal to zero, because variance in both nodes would be equal to zero. If a tree has an imperfect split, some number of tree leaves would have multiple training samples. In this case, for regression task the prediction for an unseen data observation is a mean value over all training samples in a final tree leaf.

It is important to point out that in case of a decision tree regression model we do not adjust or transform our response variable to match with specific domain constraints (percentage value) as it was with linear regression model. Indeed, it is obvious that the prediction of a decision tree would not result in a value out of our domain range [0, 1] given that values of response variables in training dataset are also in range [0, 1]. Because the prediction of a decision tree is either a mean over some data samples or a value from a single data sample, it could not fall outside of a response variable range.

2.5.2 Random Forest

Random forest improves the model performance but makes interpretation of the results more complex (James, 2021). Decision tree fitting procedure usually involves building a tree until leaf nodes has a single data observation, or multiple observations with identical value of the response variable. This property allows to get a comprehensive decision tree structure that makes no prediction error on a training dataset. However, this fact does lead model to overfitting. Overfitting means, in effect, that our model learns patterns in the training dataset to an excessive degree and is prone to errors in the predictions for the unseen data observations (Hastie, Tibshirani, Friedman, 2009).

This phenomenon is described by the high variance of a particular model, in our case a decision tree. The overfitting of a decision tree can be illustrated by a simple example, in which the decision tree is fitted on a cubic function with an additive normally distributed noise (Figure 3).

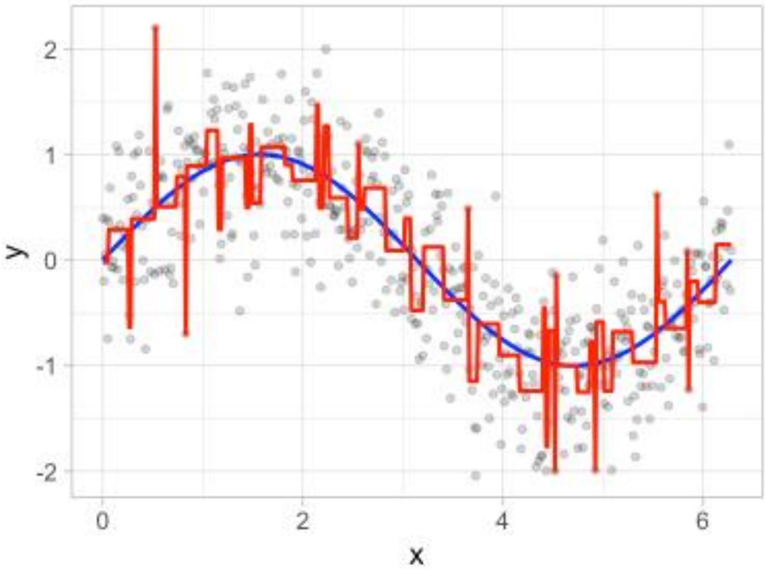

Figure 3. Decision tree overfitting on a cubic function, Boehmke, 2020

Random forest family of models allow to overcome the problem of overfitting by training an ensemble of individual decision trees and outputting an average prediction value in case of regression task. Each individual decision tree in the ensemble is trained on randomized subset of data samples. This technique of randomization, called bagging, decreases model variance, which in turn, prevents from overfitting and improves the overall predictive capability of the random forest (Breiman, 2001).

3 Literature Review

The keywords for the literature review's focal areas were established. The following were the keyword phrases: *delivery performance in the postal sector, variables influencing performance in the postal sector, delivery rate in the postal sector*. Google Scholar was used to do the search. Literature was found that applied machine learning algorithms on the postal sector data: demand forecasting (Munkhdalai, Park, Batbaatar, Theera-Umpon, Ryu, 2020), handwritten digit recognition (Shamim, Miah, Angona Sarker, Al Jobair, 2018), postal delivery areas analysis (Han, Yu, Na, Jung, Heo, Jeong, Kim, 2022). However, there is lack of research related to the thesis subject: studies that are analysing delivery performance using data collected through all postal chain.

3.1 Postal Service

Oxford dictionary defines postal service as “the national organization in many countries that is responsible for delivering letters, etc”. Collection, transportation, and delivery of all forms of letters, documents, printed materials (books, newspapers, and magazines), and parcels by all types of public and private operators are referred to as postal services. They are universally acknowledged as playing a significant role in society and as a key part of a country's economic and communication infrastructure. They are an important tool for communication and information sharing (Otsetova, Dudin, 2018).

Postal services (particularly courier services) are classified as a type of logistics service in today's world. Postal services have a minimal amount of consumer interaction, which dictates the service delivery process's predicted high efficiency. Customers, on the other hand, play an important part in service delivery, owing to contemporary technology's capabilities, which enable traceability, changing movement direction, and so on (Otsetova, et al., 2018).

In the marketing of postal services, the factor of time is extremely important. Technology is now having a strong influence on several phases of the postal process, namely in the sorting and delivery operations. Barcodes, for example, help speed up the sorting of postal products since they can be automatically categorized. Similarly, the use of Information and

Communication Technology (ICT) in the delivery of postal services is a source of efficiency. Customers in a self-service setting, for example, do a specific task, increasing the efficiency of the process and saving labour expenses for the postal service provider. For postal services, technology will become increasingly vital. The new digital means cannot replace the conventional distribution, but they can improve process efficiency and flexibility while lowering transaction costs (Otsetova et al., 2018).

3.2 Delivery Performance

Delivery performance has become a key criterion of success (Milgate, 2001). There are links between delivery performance and both the complexity of the product/process and the unpredictability of the management systems.

Increased product diversity and more complicated supply networks, on the other hand, did not appear to affect performance. (Vachon, Klassen, 2002). While delivery performance is an essential component of the overall logistics scorecard, a study of the literature revealed little attempts to objectively quantify the magnitude to which chain elements influenced delivery performance. The scarcity of literature, according to Vastag, Kasarda, Boone (1994), is due in part to the fact that delivery is the result of a series of upstream activities and administrative choices. Furthermore, downstream processes, such as inadequate logistical preparations, might have a bad effect on delivery quality. As a result, rather than viewing delivery performance via the lens of a single manufacturer, explicit acknowledgement of upstream and downstream supply chain activity is required.

The research acknowledged the need of paying more attention to the client side of the value chain. According to the data, last mile activities in the postal value stream, which are the final stage between the recipient and the post, have seen poor performance (Macioszek, 2017).

Some of the effects of the last-mile difficulties were discovered in the study. One of the obstacles of successful last mile operation was the continued reliance on old postal concepts, which are outdated in today's economic world. As a result, customers and established market areas were being lost to private rivals. The rivals value innovation and are fast to adapt new trends into their business models. Last mile activity was also discovered to be difficult because of improperly addressed locations, resulting in huge numbers of undelivered mail and packages (Laseinde, Mpofu, 2017).

It is worth noting that almost every product and service offered by government-owned posts has been imitated by private competitors, who are generally more efficient when it comes to service delivery time, collection methods, post-service follow-up, and customer support, even at increased service rates. However, the competitive advantage may be due to lesser quantities handled by private rivals in comparison to national postal service companies. As a result, postal operators must optimize their competitive edge over new competitors (Laseinde et al., 2017).

3.3 Work Performance in Literature

In this subsection, analysis of factors affecting the performance for the general delivery sector will be discussed.

Performance of courier service industry

Nyaga (2017) explored factors affecting on work performance in courier service industry. The data was collected from 134 courier firms by conducting the survey. The author explored effect of training, motivation, customer service, transport infrastructure, information, and technology on performance. In the study, the data was analysed using the *descriptive statistics.*

It was noted that the couriers’ companies were invested a lot into the employees´ training that could lead to success in the sector. Another element contributing to the rise of courier service companies was information and communication technology. Next, the author noted that the infrastructure level should be taken to use right type of vehicle and meet clients´ demand.

Automation and performance

Bloss (2013) stated that automation plays significant role in postal sector helping with sorting, reading addresses, and transporting mail containers at postal sorting centers. Lu Y., Tu, Lu S., Wang (2010) in the study mentioned that nowadays manual sorting cannot cover postal sector needs due to required time as well as high cost. Authors mentioned that higher recognition performance of automated sorting could help increase the efficiency of postal sector. Frohm, Lindström, Winroth and Stahre (2006) explored advantages and disadvantages of automation in manufacturing companies. Authors collected data by interviewing people with expertise in automation field and use *descriptive statistics* analysis. As the result, increased efficiency and improved quality were noted as the main automation advantages. As the main disadvantage, difficulties in automation caused of variation of products is noted.

Supply chain complexity and delivery performance

Milgate (2001) explored the linkage between the supply chain complexity and delivery performance. In the study, the delivery performance measure constituted of speed and reliability. To build the regression models, the following preprocessing steps were done: standardization to assist the interpretation and natural logarithmic transformation to process the large variation of delivery speed and average lateness variables. Two separate *regression analysis* of delivery reliability and delivery speed were done. In the study it was observed that the late delivery by suppliers significantly negatively affect to delivery time. In case of Posti, delays in clearance step of postal chain (collecting items and delivery to the distribution center) can negatively affect the performance. Also, it was noted that larger manufacturers provide less reliable deliver that leads to higher average lateness. It was noted that the complexity of the supply chain is linked to the delivery performance. At the time, no relationship between delivery speed and technological complexity.

Health and work-related factors and productivity loss

Heuvel, Geuskens, Hooftman (2010) studied health and work-related characteristics associated with productivity loss at work. In the study, the performance assessment was done by employees using the survey. *Logistic regression analysis* was used to study the association of demographic, work, and health related features to performance of work. It was noted that there is low linkage between sickness absence and performance at work. The multivariate analysis showed that employees with temporary contact increase the probability of reporting low performance at work. Also, it was noted that older workers report low performance at work less than younger workers.

4 Data modelling

This section describes the process of data modelling to identify patterns in the data. There are five subsections. In the first subsection, the case description is presented. In the second subsection, selected data for analysis is described. In the third subsection, the diagram visually presents data cleaning steps. In the fourth subsection, the data preprocessing steps are described in detail. In the fifth subsection, Generalized Linear and Random Forest Regression models are built to explore factors affection on performance rate.

4.1 Case Description

Company background

This master thesis is done in collaboration with Posti; one of the leading delivery and fulfillment company with operations in Finland, Sweden and Baltics. Posti's operations are divided into following businesses: Postal Services, Parcel and eCommerce, Freight, Transval, Aditro Logistics. These 5 businesses are operated by approximately 21,000 people. In 2021, company's net sales were amounted to EUR 1.6 billion and the adjusted EBITDA amounted to EUR 181.6 million. The company is owned by the Finnish state (Posti in Brief, 2022).

Case description

In Posti, Postal Service reporting team is responsible for tracking the overall performance of postal chain parts by creating needed measures and reports in PowerBI. As delivery time is highlighted as one of the factors affecting customer satisfaction in postal sector (Jucha, Stalmachova, Zilincikova, Jaculjakova, Corejova, 2020), that is why reporting team needs to track the delivery rate without any delay. As delivery performance is measured at the last step of postal chain and previous steps influence it, so, the dataset with variables collected from all steps is extracted. Also, it is important to mention that there are factors that affect the workers' performance such as work engagement, sickness, etc. Those factors are also considered and added to dataset.

4.2 Data Description

The dataset was created by Postal Service reporting team and used for description analysis in PowerBI. The data contain weekly metrics tracked for post offices in Finland. The records were done in period from 1st week of year 2020 to 11th week of 2022. Overall, there are 31 variables and 20880 observations. The list of 31 variables with description can be observed in Table 2. Variables´ description was extracted during meetings with owners of the Power BI report and independent exploring of Data Analysis Expressions (DAX) programming Language that were used for creating measures in PowerBI.

In original dataset, columns are named in Finnish and translated to English for analysis. The list of original column names can be found in Appendix 1. The target variable is colored orange.

Table 2. Data Description. The coloured cells indicate the dependent variable ("PerformanceRate").

Column Name	Description
AutoSortingPerc	Percentage of automated sorting.
CostCenter	Cost center information (cost center id, type, area). There are three cost center types: BD (basic delivery), OPS RD BD and PS P (internal abbreviations).
CostCenterSize	Size of the cost center: S (small), M (medium), L (large). Value is calculated based on actual WTI (WTIActual).
FeedbackAnsPerc	Percentage of feedback that were processed and feedback ticket is closed.
FeedbackPerc	Percentage of received customer feedback by number of distribution points.
IndoorDelayTic	Number of tickets created with information about volume of items in delivery office which were not delivered on time.
IndoorDelayVol	Volume of items in delivery office (indoor work) which was not delivered on time.
IndoorEfficiency	Efficiency of the indoor work (number of processed items in an hour).
OutdoorDelayTic	Number of tickets created with information about volume of items from delivery route returned to delivery office which were not delivered on time.
OutdoorDelayVol	Volume of items from delivery route (outdoor work) returned to delivery office which were not delivered on time.
OverworkPerc	Percentage of overtime hours out of all total working hours.
PerformanceRate	Delivery performance shows percentage of delivered items on time. More detailed information can be found below the table.
PulseScore	Average pulse score shows the overall spirit in the cost center's team. Values are collected by survey.
ResidialTotalTic	Total number of tickets created with information about delayed mail items. Calculated as sum of OutdoorDelayTic, TransportDelayTic and IndoorDelayTic.
ResidialTotalVol	Total volume of delayed mail items. Calculated as sum of OutdoorDelayVol, TransportDelayVol and IndoorDelayVol.
RMUsageRate	Percentage of users that used Route Master application (RM).
ShipmentDelayTic	Number of tickets created with information about transportation delays.
SicknessPerc	Percentage of sickness hours out of all total working hours.
SortingErrorTic	Number of tickets created with information about sorting errors.
SuccessRate	Success delivery rate is percentage of successfully delivered mail items from basic delivery and early morning delivery.
TransportDelayTic	Number of tickets created with information about volume of items from transportation which were not delivered to delivery office on time.
TransportDelayVol	Volume of items from transportation (delivery outdoor transportation work) which were not delivered to delivery office on time.
VolumeActual	Actual number of items to deliver (actual volume).
VolumeDiffPerc	Difference between actual and planned volume (percentage). Positive number indicates underestimation of planned volume value.
VolumePlan	Planned number of items to deliver (planned volume).
WorkCommitment	Average work commitment (1-10 scale). Data collected by survey.
WTIActual	An actual value of full-time working days (used work power).
WTIDiffNum	Difference between actual and planned WTI (number). Positive number indicates underestimation of planned WTI value.
WTIDiffPerc	Difference between actual and planned WTI (percentage). Positive number indicates underestimation of planned WTI value.
WTIPlan	An estimate of required full-time working days.
YearWeek	Year and week number as the aggregation date.

Performance Rate

Performance Rate is the variable that is examined by building models in this study. This subsection describes the variable numerically and visually. PR shows percentage of delivered items on time (only trackable items including domestic and outbound items). PR can get real values from 0 to 1, where 0 means unsuccessful delivery and 1 – successful delivery. Table 3 contains numerical description of the variable and Figure 1 shows the distribution.

Table 3. Success Rate Numerical Description

count	9154.00
mean	0.947232
std	0.043222
min	0.480000
25%	0.927000
50%	0.955000
75%	0.977000
max	1.000000

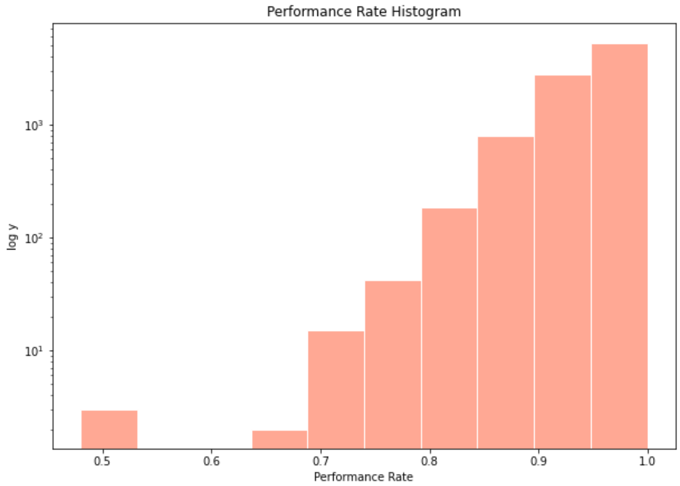

Figure 4. Performance Rate distribution

From numerical and visual description, it can be observed that the variable has low standard deviation value (0.043) and mean value close to 1 (0.947) which means the distribution is left skewed. The scale of axis (Figure 1) is changed from linear to logarithmic axis as the number of values in each bin has large value range. Around 90% of values are between 0.9 and 1.

4.3 Data Processing Diagram

Figure 5 represents the view of data preprocessing steps that are described and justified later in subsection 4.4.

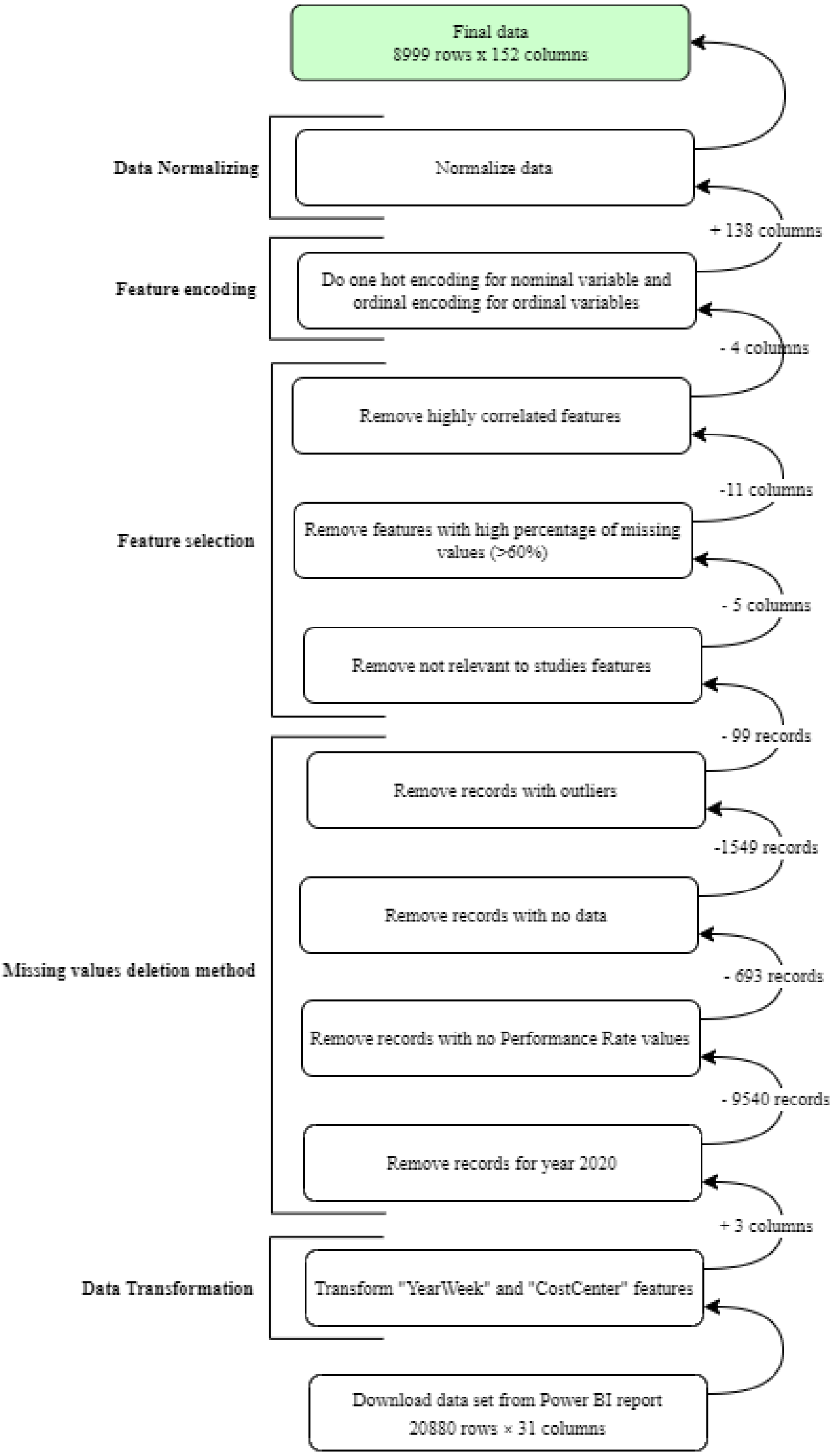
Final data
8999 rows x 152 columns
Data Normalizing
Normalize data
+ 138 columns
Feature encoding
Do one hot encoding for nominal variable and ordinal encoding for ordinal variables
- 4 columns
Remove highly correlated features
-11 columns
Feature selection
Remove features with high percentage of missing values (>60%)
- 5 columns
Remove not relevant to studies features
- 99 records
Remove records with outliers
-1549 records
Remove records with no data
Missing values deletion method
- 693 records
Remove records with no Performance Rate values
- 9540 records
Remove records for year 2020
+ 3 columns
Data Transformation
Transform "YearWeek" and "CostCenter" features
Download data set from Power BI report
20880 rows × 31 columns

Figure 5. Data Processing Diagram

Firstly, to get detailed information, data transformation was done for features that contain aggregated values. Secondly, irrelevant and noisy data was removed. As the next step, categorical variables were encoded by one-hot and ordinal encoding. Finally, the data was normalised.

4.4 Data Preprocessing

Data preprocessing step takes great importance as presence of irrelevant, noisy, and unreliable data makes data mining algorithm usage difficult to conduct.

4.4.1 Data Transformation

Two features in dataset were created by merging few values ("CostCenter" and "YearWeek"). Cost Center information such as id, type and area are recorded in one feature in the following format: "ID Type Area". These three values were split into three features such as "CC_ID", "CC_Type" and "CC_Area" that show id, type, and area of cost center respectively. Also, year and week number of the record are described in one feature in the "YYYYWW" format. Week number and year were split into two features "Year" and "Week" that show year and week respectively. The result of this step gives opportunity to do analysis by year, area of cost center, etc. separately. Table 4 and Table 5 show the data before and after preprocessing, respectively.

Table 4. Features before pre-processing

CostCenter	**YearWeek**
111111111 BD Klaukkala	202201

Table 5. Features after pre-processing

CC_ID	**CC_Type**	**CC_Area**	**Year**	**Week**
111111111	BD	Klaukkala	2022	1

4.4.2 Data Cleansing

The dataset contains 28800 records and 31 features. Totally, 43% of values are missed and each record has at least one missing value.

As high number of missing values as well as presence of irrelevant features leads to the production of less reliable model and time-complexity for model implementation (García et al., 2015), detection and handling of those should be done.

Missing values deletion method

Some of the features were added to the dataset only in 2021 and that leads to having all missing values for the 2020 records for these measures. Overall, records from 2020 contained 54% of missing values when 2021 and 2022 contains 37% and 38% respectively (Figure 6). On agreement with business side, it was decided to drop 9540 records that were done in 2020.

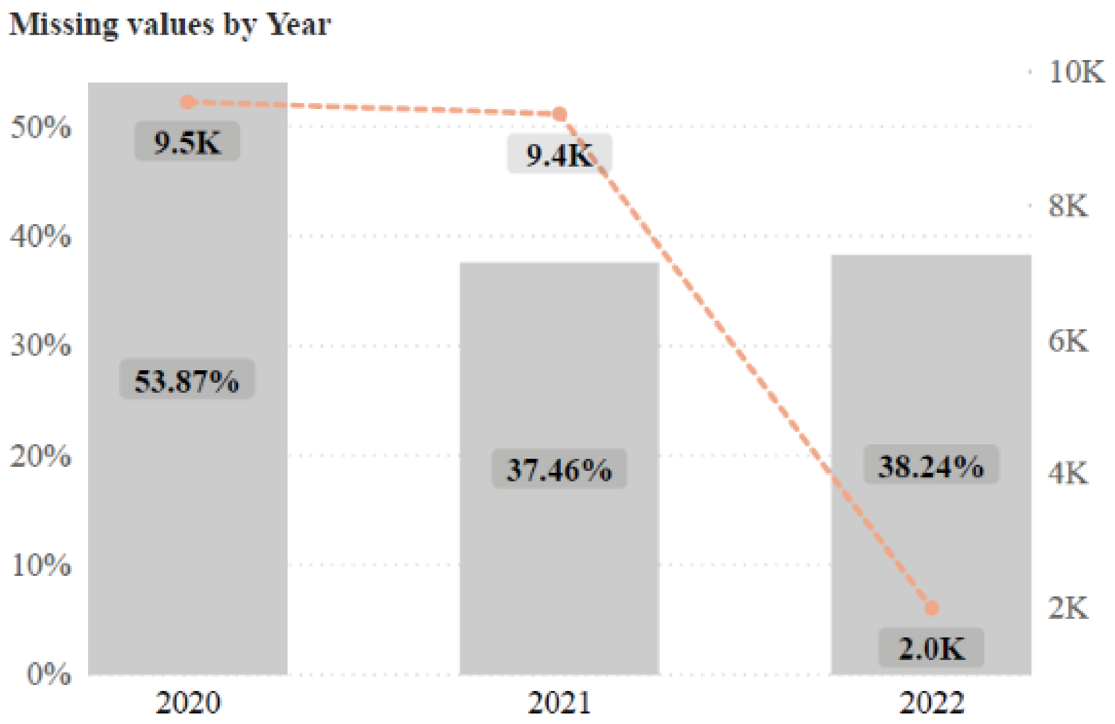

Figure 7. Missing values by years

It was observed that 11 cost centers did not have any feature values except information about themselves (cost center id, type, and area) and date of the record. Those 1549 records were deleted from the dataset. The reason of having these records were described by business side as adding irrelevant (administrative in this case) cost centers to the dataset.

To explore factors affecting on performance rate, supervised machine learning models will be built in the next section. As supervised learning requests labeled outcome measure (Hastie et al 2017), 693 records with missing values in “PerformanceRate” feature were removed.

As the result of deletion method usage, number of records decreased from 20880 to 9098.

Corrupted data

“RMUsageRate” displays the application usage rate in a scale from 0 to 1. It was observed that 93 records had value more than 1. It indicates that the data may have been corrupted when transferred from the source data base. As the result, those values in records were clipped to 1 on agreement with business side.

Outliers

In the next step, the outliers as improbable values were detected through the visualization. As a result, 99 records were removed: 24 records based on “PerformanceRate”, 56 records based on “AutoSortingPerc”, and 19 records based on “IndoorEfficiency”. At this stage, 8999 records were left for future analysis.

Feature selection

After removing irrelevant records above feature selection was done. As the first step, “FeedbackPerc” and “FeedbackAnsPerc” features were removed, as values could be received only after completion of delivery. It means that those features cannot affect the PR. Also, the feature “SuccessRate” was removed as it overlapped with the target variable (“PerformanceRate”). Secondly, “CC_Type” feature was removed as the study focused only on basic delivery offices; the dataset after removing irrelevant records contained only BD records. Next, “CC_ID” feature was removed as id values were generated for each office.

As a further step, fraction of missing values in each feature was explored. Figure 8 shows distribution of features with missing values.

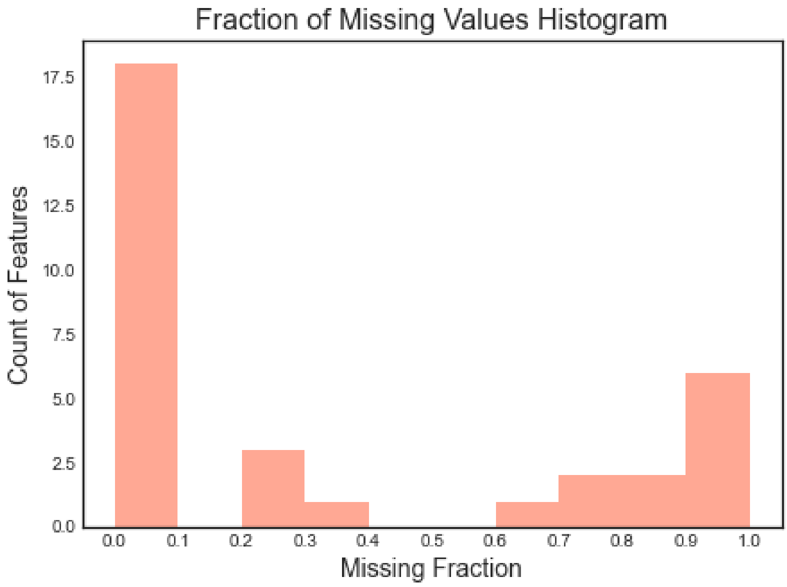

Figure 8. Fraction of Missing Values Histogram

It can be observed that there are 11 features with more than 60% of missing values and 6 features with more than 90% of missing values. Detailed information about percentage of missing values in each variable can be found in Appendix 2.

Correlation

In the next step, the correlation between variables was checked to remove variables with strong linear relationship. In the studies the moderate (correlation value from 0.50 to 0.69), highly (correlation value from 0.70 to 0.89), and very highly (correlation value from 0.90 to 1.00), correlated variables were explored. Table 6 displays those variables with correlation values. Features “VolumeActual”, “WTIActual”, “VolumePlan”, “WTIDiffNum” were removed.

Table 6. Highly correlated features in the data. Removed variables are coloured in red

	WTIPlan	**WTIActual**	**WTIDiffNum**	**WTIDiffPerc**	**VolumePlan**	**VolumeActual**
WTIPlan	1					
WTIActual	0.99	1				
WTIDiffNum	0.86	0.86	1			
WTIDiffPerc	0.041	-0.056	0.54	1		
VolumePlan	0.86	0.86	-0.32	-0.048	1	
VolumeActual	0.88	0.88	-0.33	-0.043	0.97	1

4.4.3 Feature Encoding

Apart from the numerical (or quantitative) variables, the dataset contained categorical features. In the studied dataset, cost center area (“CC_Area”) was a nominal variable that contained 139 labels for locations; cost center size (“CostCenterSize”), year (“Year”) and week (“Week”) features were ordinal.

In the dataset, cost center size feature (“CostCenterSize”) was transformed in the following manner: “S”:1, “M”: 2, “L”:3. The nominal feature transformation led to the addition of 138 new columns where each column name was the name of the area and values in the column were binary (0 or 1) where 1 showed that the record is done is this area.

4.4.4 Data Normalising

As the next data preparation step, data min-max normalization was done using equation 1. This type of normalization rescaled feature to [0, 1] range where 0 and 1 were assigned to the minimum and maximum values respectively. Later KNN was used as a missing values imputation technique. As KNN is a distance-based algorithm, all the features needed to be rescaled into one fixed range.

4.5 Model implementation

The subsection describes the process of model implementation. Next, subsection presents the implementation of Generalized Linear Regression model and Random Forest Regression model.

The process of building models was divided into following steps in the study (influenced by Zhang, 2010):

1. First, the preprocessed dataset was split into training and testing data using hold-out method. For training, 80% of data was used, remaining 20% of the data was used for testing.
2. In the second step, KNN imputation method was applied to the training and testing data to fill missing values. It is important to mention that the missing values in testing data were filled based on training data. KNN was chosen as the imputation method due to its advantages: machine learning based technique provides more accurate imputation of missing values, the method preserves the original dataset structure and do estimations for both: for qualitative and quantitative data (Faisal, 2018).
3. Third step was training the Generalized Linear Regression model and Random Forest Regression model on the training dataset and getting the information about most important factors in models.
4. Finally, the models' performance was assessed by RMSE and R2 coefficients after doing prediction for testing dataset. To get stable error metrics, the dataset split, and model running were done 5 times and average values of the coefficients were calculated.

At this stage, there were 8999 records and 152 features (139 features were displaying areas). The data stayed in a normalised form as KNN missing values imputation method was used to fill missing values.

4.5.1 Generalized Linear Regression model

Coefficients

Firstly, the features which display the areas were examined. Coefficients have values between 1.09 and 2.55. This indicates that the areas with higher corresponding coefficient have stronger increasing effect on the performance rate value. To examine the overview of area coefficients for whole Finland, the map was built in PowerBI and presented to reporting team.

In the next step, the date features were examined. “Week” and “Year” features´ coefficients have both positive with values 1.30 and 1.16, respectively. Reflecting it in terms of business, delivery performance of the company improved over time. Figure 9 demonstrates performance rate values against date and confirms the logic described above.

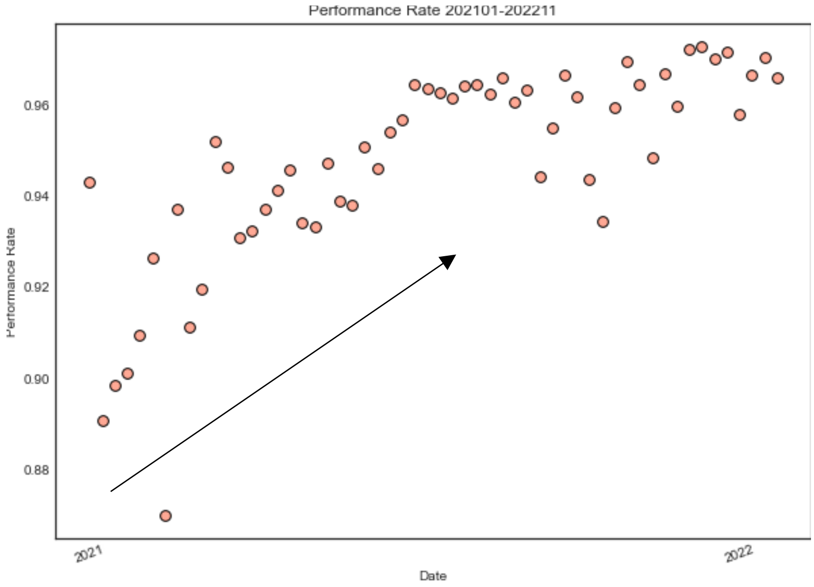

Figure 9. Performance Rate VS Time

Finally, the measures´ coefficients were examined and presented in the order from the most important to least important in Table 7.

Table 7. Feature coefficients, Generalized Linear Regression model

Feature name	Coefficient
IndoorEfficiency	0.585074
OverworkPerc	-0.57538
SicknessPerc	-0.57324
RMUsageRate	0.430455
WTIDiffPerc	-0.38457
AutoSortingPerc	0.275618
CostCenterSize	-0.14514
WorkCommitment	0.033858
VolumeDiffPerc	-0.02675

As the five most important factors, “IndoorEfficiency”, “OverworkPerc”, “SicknessPerc”, “RMUsageRate”, “WTIDiffPerc” can be considered. Because the linear model connects predictor variables with response variable via logit link function, the coefficients obtained are not subject to the standard interpretation of linear regression coefficients. However, the logit link function is monotonically increasing, therefore it is possible to conclude that larger coefficient implies stronger increase in the final prediction for “PerformanceRate” (name of response variable). As all the predictor variables were transformed with min-max normalization, model coefficients reflect the respective impact of each variable: variables with corresponding coefficients closer to 0 give less impact compared to the variables with a large coefficient.

Error metrics

In the study, two error metrics were examined to assess the model performance (RMSE and R2). After running the model 5 times, the following results were observed (Table 8):

Table 8. Error metrics, Generalized Linear model

Coefficient	Mean value	Standard deviation
RMSE	0.073	0.002
R2	0.239	0.016

Relatively low standard deviation of R2 and RMSE values indicate that the model is resilient, meaning that the predictive performance does not drastically change due to the differences in each training data split. Value of R2 coefficient shows that 0.239 (proportion) of response variable variability explained by the model. As R2 value equal to 1 is the best possible result, it can be concluded that the model describes relatively low proportion of variability.

4.5.2 Random Forest Regression model

Coefficients

To get information about features affecting PR the most, the importance of features were examined and can be seen in Table 9. For each feature in each built tree, the measure showing the decrease of impurity of the split were calculated. The final feature importance was calculated as average value of those values for all built trees.

Table 9. Feature Importance, Random Forest Regression model

Feature name	Importance
Week	0.200746
Year	0.09032
RMUsageRate	0.08019
SicknessPerc	0.063044
VolumeDiffPerc	0.059963
IndoorEfficiency	0.057612
OverworkPerc	0.056073
WTIDiffPerc	0.055914
AutoSortingPerc	0.053986
WTIPlan	0.050537
WorkCommitment	0.044389
CostCenterSize	0.004091

From the Table 9 it can be observed that the most important features are Week and Year. At the same time, when checking five most important measures, “RMUsageRate”, “SicknessPerc”, “VolumeDiffPerc”, “IndoorEfficiency”, “OverworkPerc” are highlighted. The list of features that are observed as the most important ones in Random Forest Regression model is close to the observed important features in Generalized Linear Regression model.

Figure 10 shows one of the trees to make the interpretation process easy (area variables are blurred). If the described condition is met, the left leaf should be chosen for moving to the next node. By visualizing tree with tree-depth of three, it can be observed that the features with high importance are presented as nodes.

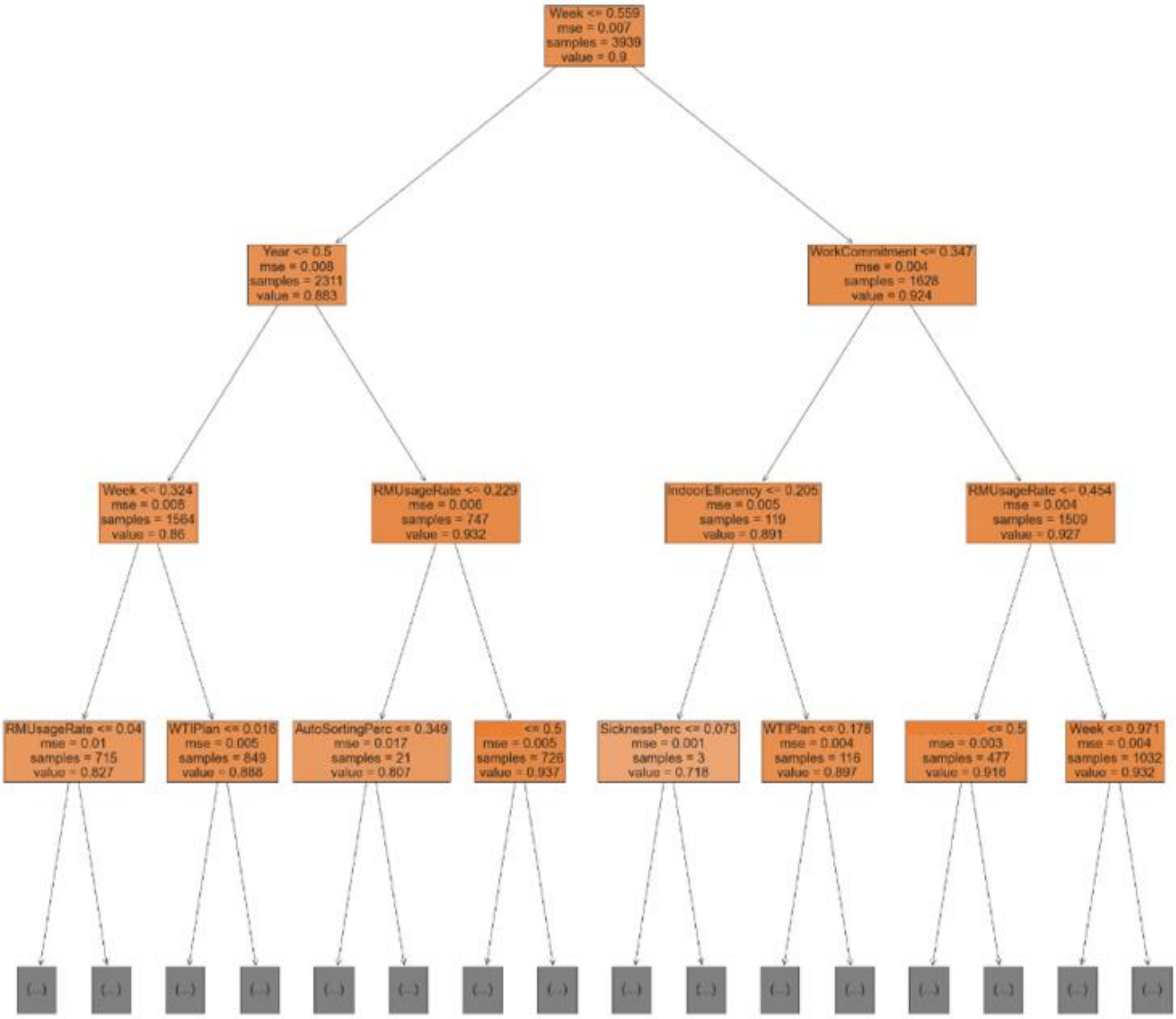

Figure 10. Decision Tree, Random Forest. The blurred words represent area names which are confidential information.

Error metrics

In the study, two error metrics were examined to assess the model performance (RMSE and R2). After running the model five times, the following results were observed (Table 10):

Table 10. Error metrics, Random Forest Regression model

Coefficient	Mean value	Standard deviation
RMSE	0.068	0.002
R2	0.315	0.012

The Random Forest Regression model also has low standard deviation values of R2 and RMSE coefficients. It indicates that the model is resilient. Value of R2 coefficient shows that 0.315 (proportion) of response variable variability explained by the model. The model has R2 value higher than in Generalized Linear Regression model as well as lower RMSE value. From this, it can be concluded that this model performs better as well as explains variability better.

5 Result Analysis

The thesis described the factors that were observed to have affected the work performance in the past literature. Posti’s Postal Services reporting team requested to explore already created measures and lays out the most important factors that affect delivery performance and how it allows the company to reallocate its resources to improve the performance. Most of the past research on this topic was conducted using qualitative methods and, in those studies, the work performance was either evaluated by managers or employee surveys. Nyaga (2017) performed descriptive analysis to explore factors affecting the performance in courier service industry. The author noted that training of employees and proper development of Information and Communication Technology (ICT) play important role. Also, unwise selection of type of vehicle delivery can negatively affect the performance. As automation is now popular, Bloss (2013) highlighted that it plays significant role on postal sector too. Frohm et al. (2006) highlighted increased efficiency and improved quality as the main advantages of automation. Heuvel et al. (2010) explored health and work-related factor associated with the productivity loss by performing logistic regression. Authors noted weak linkage between sickness absence and performance at work. Also, older workers and workers with temporary contracts reports low performance more often.

For the study, two regression models were built (Generalized Linear Regression model and Random Forest Regression model) and evaluated by R2 and RMSE error metrics. When choosing models, the following factors were considered: type of problem, simplicity of model interpretation and usage for business side as well as constraints regarding the target variable. For Generalised Linear Regression model, the predictor variables were connected with response variable via logit link function in order to enforce various limitations on the response variable (unit [0, 1]). Each model was run for five times, and average value and standard deviation of coefficients were used for assessment. Table 11 presents the values for models’ comparison:

Table 11. Error metrics, Models´ Comparison

	Coefficient	Mean value	Standard deviation
Generalized Linear Regression model	RMSE	0.073	0.002
	R2	0.239	0.016
Random Forest Regression model	RMSE	0.068	0.002
	R2	0.315	0.012

It can be observed that both models have low standard deviation values for R2 and RMSE coefficients. It indicates that the models are resilient. Value of R2 coefficient shows that Generalized Linear Regression model explains 0.239 (proportion) of response variable variability whereas Random Forest Regression model explains 0.315. Generalized Linear Regression model has value of RMSE equal to 0.073 whereas Random Forest Regression model has value 0.068. From the described above metrics, it can be concluded that Generalized Linear Regression model performs better and describes variability better as well.

To get the five most important features, the features´ coefficients were checked for Generalized Linear model (Table 12) and features´ importance for Random Forest model (Table 13).

Table 12. Features´ coefficients, Generalized Linear Regression model

IndoorEfficiency	0.585074
OverworkPerc	-0.57538
SicknessPerc	-0.57324
RMUsageRate	0.430455
WTIDiffPerc	-0.38457

Table 13. Features´ importance, Random Forest Regression model

RMUsageRate	0.08019
SicknessPerc	0.063044
VolumeDiffPerc	0.059963
IndoorEfficiency	0.057612
OverworkPerc	0.056073

In Generalized Linear Regression model, the response variable is transformed via logit link function, the coefficients obtained are not subject to standard interpretation of linear regression coefficients. Coefficients reflect the respective impact of each variable: variables with corresponding coefficients closer to 0 give less impact compared to the variables with larger absolute value. Positive and negative sign indicates positive and negative effect on the performance rate. In Random Forest Regression model, the importance measure shows the average decrease of impurity for features.

It can be noted that both models listed similar five most important features but in a different order. It can be observed that sickness percentage negatively affected the performance as less people were working. Also, when overwork percentage was increasing, the performance was decreasing. It could be due to the tiredness of employees or high volume of items to deliver. Furthermore, it was crucial to do accurate prediction for both: number of employees needed and expected volume to be able to reallocate the resources. Date features significantly affected the performance rate in a positive way that indicated that the company had been improving the performance over the time. Finally, the usage of the Route Master application, that allows to seek assistance from other employees to finish the work, affected the performance rate.

Generalized Linear Regression Model defined indoor efficiency (variable "IndoorEfficiency") as the most important feature. However, the Random Forest Regression model assigned the highest feature importance value to Route Master usage rate (variable "RMUsageRate"). It means that the company should pay attention to those identified factors.

Posti needed to make sure that the indoor work processes are well organised in distribution centres as it leads to possibility of handling larger number of items per hour and as a result, positively affects the performance rate. Milgate (2001) also noted that the complex chain has a link with delivery performance, so Posti needed to be sure that the process is simple and clear for all employees. Usage of Route Master application allowed to seek assistance from other employees to finish the work and that played a significant role on performance rate. It was also noted in literature that the use of ICT in the delivery of postal services makes the process more efficient (Otsetova, et al., 2018). Hence, it is advisable to promote useful features to the employees. Also, the training helps a lot for the work performance (Nyaga 2017). Therefore, Posti needs to train their employees for mobile application usage as well as indoor work process. As a result, they will have more transparency of the processes and are comfortable with the use of digital tools, which will lead to better performance.

6 Conclusions and discussion

This thesis outlines the main factors that have the greatest impact on delivery performance and explains how this enables the business to better deploy its resources. The data used in this study comes from PowerBI report that contained information regarding different touchpoints of postal chain. The Postal Services reporting team at Posti requested to investigate the measurements that had already been established. It is crucial to understand the aspects that influence delivery performance because it has evolved into a crucial success criterion (Milgate, 2001).

In the study, Generalized Linear Regression model and Random Forest Regression model were applied. These models defined indoor efficiency and Route Master usage rate as the most important factor respectively. Random Forest Regression model is more suitable model for identifying relationship between performance rate and factors as it has lower RMSE and higher R2 values.

Answering the research questions

1. *How and which machine learning methods are applied to identify the most important variables in datasets and are they applied in postal service data?*

Machine learning involves learning of hidden patterns in historical data and as a result, perform predictions for the new events. To identify relationships between variables in datasets, supervised and unsupervised ML algorithms can be applied.

When it comes to modelling of a response variable in a domain of real numbers, linear regression models are comprehensive yet straightforward approach. The model has good interpretation possibilities; most important factors can be identified by highest absolute values of coefficients. Also, the coefficients´ sign indicates direction of the relationship (positive and negative). Tree-based models are “white-box” models that assign feature importance for each variable. The score represents the most important factors; a higher value indicates larger effect on the model.

There is a scarcity of literature in this specific domain of postal services. Therefore, this thesis provides a starting point into this domain by applying simple machine learning methods for exploration of factors that affect the delivery performance of all trackable items.

2. *What factor from the dataset affect the delivery performance rate the most in the case of Posti and which machine learning model is the most suitable for identifying relationship between performance rate and those factors?*

For the study two regression models were built: Generalized Linear Regression model and Random Forest Regression model. The Random Forest Regression model describes 0.315 of the performance rate variability, R2 value is equal to 0.068. At the same time, Generalized Linear Regression model describes lower proportion of the performance rate (0.239) with higher value of R2 (0.073). Answering on the second part of the research question, it can be stated that Random Forest Regression model is more suitable model for identifying relationship between performance rate and factors as it has lower RMSE and higher R2 values. Both models featured nearly the same five most significant factors, but in a different sequence. In response to the first part of the research question, the Generalized Linear Regression model ranks indoor efficiency (variable "IndoorEfficiency") as the most important, whereas the Random Forest Regression Model ranks Route Master usage rate (variable "RMUSageRate") as the most relevant factor.

Limitations and future research

The methods for defining and calculating the performance metric in the postal industry were outside the scope of this thesis. The data quality might be noted as a drawback because the offered data already included aggregated measures that were taken from various sources. Data corruption could result from this transfer technique. Because of this, it cannot be ensured that the dataset accurately depicts Posti's business processes.

As it was noted, the dataset contained high percentage of missing values that could significantly affect the model performance. The reporting team could focus on handling those missing measures as well as avoid their occurrence. As models are not describing high proportion of performance rate variance, it is a good idea to explore and add other factors to dataset that could affect the performance rate. Also, more sophisticated machine learning models can be built. It leads to the complexity of the model interpretation but at the same time can give better performance than simplistic models.

References

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A. and Aljaaf, A., 2019. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. Unsupervised and Semi-Supervised Learning, pp. 3–21.

Asuero, A., Sayago, A., González, A., 2006. The Correlation Coefficient: An Overview. Critical Reviews in Analytical Chemistry, 36(1), pp. 46–47.

Bloss, R., 2013. Automation Pushes the Envelope of Postal Mail Handling Efficiency. Assembly automation 33.1, pp. 3–7.

Boehmke, B., Greenwell, B., 2020. Hands-On Machine Learning with R.

Breiman, L., 2001. Random Forests. Machine Learning 45, pp. 5–32.

Chicco, D., Warrens, M. J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, p. 1.

Davison, A., 2003. Statistical models. Cambridge, U.K.: Cambridge University Press, pp. 41–44.

Faisal, S., 2018. Nearest neighbor methods for the imputation of missing values in low and high-dimensional data, pp. 9–17.

Frohm, J., Lindström, V., Winroth, M., Stahre, J., 2006. The Industry's View on Automation in Manufacturing. Automated Systems Based on Human Skill and Knowledge, France.

Garcia, S., Luengo, J., Herrera, F., 2015. Data Preprocessing in Data Mining, pp. 1–3, 6–7, 41–42, 59–60.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. and Herrera, F., 2016. Big data preprocessing: methods and prospects, pp. 3–4.

Gupta P., Sehgal N., 2021. Introduction to Machine Learning in the Cloud with Python, pp. 5–9.

Han, K., Yu, Y., Na, D. G., Jung, H., Heo, Y., Jeong, H., Kim, J., 2022. Understanding postal delivery areas in the Republic of Korea using multiple unsupervised learning approaches. ETRI Journal, 44(2), pp. 232–243.

Hardy, M., 1993. Regression with Dummy Variables, pp. 8–9.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning.

James, G., 2021. An Introduction to Statistical Learning, pp. 29–30.

Jucha, P., Stalmachova K., Zilincikova M., Jaculjakova S., Corejova T., 2020. What factors affect customer satisfaction with postal services?

Kim, K., Hong, J., 2017. A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis. Pattern Recognition Letters, 98, pp. 39–45.

Laseinde, O. T., Mpofu, K., 2017. Providing solution to last mile challenges in postal operations. International Journal of Logistics Research and Applications, 20(5), pp. 475–490.

Lin, W., Tsai, C., 2019. Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review, 53(2), pp. 1487–1509.

Loh, W. Y., 2014. Fifty years of classification and regression trees. International Statistical Review, 82(3), pp. 329–348.

Lu, Y., Tu, X., Lu, S., Wang, P.S., 2010. Application of pattern recognition technology to postal automation in China. Pattern Recognition and Machine Vision.

Macioszek, E., 2017. First and last mile delivery–problems and issues. In Scientific and technical conference transport systems theory and practice, pp. 147–154.

Mahanti, R., 2021. Data Governance and Data Management: Springer Singapore, pp. 5–8.

Maimon, O., Rokach, L., 2005. Data Mining and Knowledge Discovery Handbook, pp. 165–172

Maritan, C. and Lee, G., 2017. Resource Allocation and Strategy. Journal of Management, 43(8), pp. 2411–2420.

Michael J. Crawley, 2012. The R Book, p. 572

Milgate, M., 2001. Supply chain complexity and delivery performance: an international exploratory study. Supply Chain Management: An International Journal, 6(3), pp. 106–118.

Munkhdalai L., Park K. H., Batbaatar E., Theera-Umpon N., Ryu K. H., 2020. Deep Learning-Based Demand Forecasting for Korean Postal Delivery Service.

Nagelkerke, N., 1991. A note on a general definition of the coefficient of determination. Biometrika, 78(3), pp. 691–692.

Norton, E. C., & Dowd, B. E., 2018. Log Odds and the Interpretation of Logit Models. Health services research, 53(2), pp. 859–878.

Nyaga, J., 2017. Factors Affecting The Performance of Courier Service Industry: A Survey of Courier Companies In Kenya. International Journal of Supply Chain and Logistics, p. 44.

Otsetova, A., Dudin E. (2018). Postal services in the conditions of fourth industrial revolution, pp. 1–13.

Oxford dictionary. [www document]. [Accessed 9 June 2022]. Available https://www.oxfordlearnersdictionaries.com/definition/english/postal-service

Parcu, P. L., Brennan,T. J., Glass, V., 2020. The Changing Postal Environment: Market and Policy Innovation, pp. 40.

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I., 2002. Decision trees: an overview and their use in medicine. Journal of medical systems, 26(5), pp. 445–463.

Posti in Brief. [www document]. [Accessed 9 June 2022]. Available https://www.posti.com/en/group-information/posti-in-brief/

Rebala, G., Ravi, A. and Churiwala, S., 2019. An Introduction to Machine Learning, pp. 2–4, pp. 19–28.

Ronchetti, E., 1997. Robustness aspects of model choice, p. 327.

Shalabi, L. A., Shaaban, Z., & Kasasbeh, B. (2006). Data Mining: A Preprocessing Engine. Journal of Computer Science, 2(9), pp. 735–739.

Shamim, S. M., Miah, M. B. A., Angona Sarker, M. R., Al Jobair, A., 2018. Handwritten digit recognition using machine learning algorithms. Global Journal Of Computer Science And Technology.

Vachon, S. and Klassen, R.D., 2002. An exploratory investigation of the effects of supply chain complexity on delivery performance. IEEE Transactions on engineering management, 49(3), pp. 218–230.

Van den Heuvel, S. G., Geuskens, G. A., Hooftman, W. E., Koppes, L. L. J., van den Bossche, S. N. J., 2009. Productivity Loss at Work; Health-Related and Work-Related Factors. Journal of Occupational Rehabilitation, 20(3), pp. 331–339.

Vastag, G., Kasarda, J.D., Boone, T., 1994. Logistical support for manufacturing agility in global markets. International Journal of Operations & Production Management.

Wilkinson L., 1992. Tree structured data analysis: AID, CHAID and CART, pp. 2–5

Xu, Y., Goodacre, R., 2018. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. J. Anal. Test. 2, pp. 249–262.

Zhang, Y., 2010. New Advances in Machine Learning. p. 22.

Appendix 1. Dataset columns' translation from Finnish to English language

Column Name (Finnish, origin)	**Column Name (English, translated)**
Vuosivko	YearWeek
Kustannuspaikka	CostCenter
Jakelupalauttet suhteutettu%	FeedbackPerc
Jakelupalauttet onnistuminen%	SuccessRate
Toimistusvarmuus%	PerformanceRate
Vastaus% jakelupalautteet	FeedbackAnsPerc
Esityöjäämävol	IndoorDelayVol
Esityöpoikkeamat	IndoorDelayTic
Jakelujäämävol	OutdoorDelayVol
Jakelupoikkeamat	OutdoorDelayTic
JAKELUNKULJETUSpoikkeamat	TransportDelayTic
Kuljetusjäämävol	TransportDelayVol
Kuljetuksen myöhästyminen poikkeamat	ShipmentDelayTic
Lajitteluvirhepoikkeamat	SortingErrorTic
Poikkemat yht	ResidialTotalTic
Jäämä yht	ResidialTotalVol
Ohittavuus (J-taso) %	AutoSortingPerc
Sitoutumisaste	WorkCommitment
RM-käyttöaste%	RMUsageRate
Sairaus%	SicknessPerc
Ylityö%	OverworkPerc
Tehokkuus BD sisätyö (2264&2263) kpl/h	IndoorEfficiency
Suunnitellut WTI (WFM)	WTIPlan
WTI (Time)	WTIActual
WTI vs Suun	WTIDiffNum
WTI vs Suun %	WTIDiffPerc
TMP koko	CostCenterSize
Jakelun toteumavolyymi	VolumePlan
Jakelun ennustevolyymi	VolumeActual
Volyymi vs. Ennuste %	VolumeDiffPerc
Sum of Pulse average score	PulseScore

Appendix 2. Fraction of missing values for variables

Variable	Missing_fraction
TransportDelayVol	0.997222
SortingErrorTic	0.986999
TransportDelayTic	0.986443
ShipmentDelayTic	0.984665
IndoorDelayVol	0.981887
IndoorDelayTic	0.960885
OutdoorDelayVol	0.861985
ResidialTotalVol	0.846539
PulseScore	0.785421
OutdoorDelayTic	0.7033
ResidialTotalTic	0.677186
IndoorEfficiency	0.3037
AutoSortingPerc	0.27792
SicknessPerc	0.269363
FeedbackAnsPerc	0.227136
PerformanceRate	0.093344
WorkCommitment	0.071786
VolumeDiffPerc	0.071786
VolumeActual	0.036004
VolumePlan	0.02178
FeedbackPerc	0.011223
WTIDiffPerc	0.006223
OverworkPerc	0.006223
RMUsageRate	0.006223
WTIDiffNum	0.002667
WTIActual	0.000222
WTIPlan	0.000222
CostCenterSize	0
Week	0
Year	0
CC_ID	0
CC_Type	0
CC_Area	0