National College of Ireland

Utilizing Machine Learning Models to forecast sales for a Software Company

MSc Research Project
AI for Business

Linda Amado Orozco
Student ID: 22182101

School of Computing
National College of Ireland

Supervisor: Brian Byrne

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Linda Amado Orozco
Student ID:	22182101
Programme:	AI for Business
Year:	2024
Module:	MSc Research Project
Supervisor:	Brian Byrne
Submission Due Date:	12/08/2024
Project Title:	Utilizing Machine Learning Models to forecast sales for a Software Company
Word Count:	4340
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:	
Date:	11th August 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):	

Utilizing Machine Learning Models to forecast sales for a Software Company

Linda Amado Orozco
22182101

Abstract

The present study aims to train and compare different ML models that predict the sales of a company that provides software products for customers in the architecture company. The study involves real data about the company's customers which provides an understanding of the current market, and the predictions made by the model could be used by the sales and marketing team to improve and apply more data-driven strategies and sales approaches.

Keywords: Machine Learning, Sales Forecasting.

1 Introduction

The accuracy of sales forecasting has become a need to be addressed by companies to have a clear perspective of what path and decisions should be taken to achieve sales targets. Nowadays, the market is getting more competitive because of the application of Artificial Intelligence (AI) to handle problems that in the past took effort and a lot of resources impossible to cover especially for small companies. Sales is one of these areas where the application of AI models such as ML could help them to understand current patterns in the market to create more effective sales strategies. This study is based on a real company which provides software products and services for the construction, architecture, manufacturing, engineering, education, media and entertainment industries, the company has a digital sales team which is responsible to attend all their online customer needs. The data collected comes from current customers with active licenses, being licenses the goal of the prediction in the present study. Currently, the digital sales team along with the marketing team are facing challenges in understanding where the efforts should be the point, for this reason, an accurate model that helps to predict the sales based on current data could help the team to create data-driving strategies to improve the sales approach. During the exploration of the literature, it was found a gap covering the sales prediction in software products, this could be because the sales cycle, is hardly related to seasons which is the opposite of fashion or agriculture industries, and could be more related to the software quality itself, price, or external variables such as economy or inflation. Additionally, this software is a solution Business to Business that involves a more complex sales cycle. All these variables could affect the prediction of sales and its accuracy itself.

1.1 Research Objectives

- Provide data insights to the digital sales and marketing team that help them to bridge the gap between their current sales strategies with the current data sales.
- Train and compare which ML models are more capable of predicting the license sales of the company based on the type of the dataset collected.
- Understand the data from the company's customers to determine which ML model would be more suitable for sales forecasting.
- Application of ML models to address real needs that software/technology companies' sales teams are facing related to the prediction of sales and elaboration of marketing strategies.

2 Related Work

Forecasting helps businesses to make data-decision driving, the correct prediction of sales could help a company to grow and take risks but also make changes in the current process with the identification of patterns. B2B Software sales have a different sales process than the sale of a product, the lifecycle could be similar to a service however has different variables that affect it. During the research done, many studies were reviewed related to sales forecasting in different fields, such as fashion, agriculture automotive and software, but also it was found a study related to the time forecast of a construction project, using ML models. This exploration of related work helped to identify that even when some of the studies are related to sales forecasting one of them was specifically focused on the prediction of sales of software used in the construction and architectural field.

2.1 Sales Forecasting using Machine Learning models

Research in the sales forecasting field using ML models has had considerable advancement lately, including a variety of methods and approaches that have improved the discussion regarding achieving more accurate predictive models. The more remarkable methods and their applications include traditional models, new models and hybrid models. It was found that the researchers frequently compared different model evaluation results to measure the accuracy and their error, this comparison helped this study to map the different applications of the regression models and what fit better according to the data type being based on what techniques to process the data and to train the model could be replicable in the present study following others experience.

2.1.1 Sales Forecasting using regression models

Regression models such as Linear and multiple regression have been the subject of many studies because they are well-known models for prediction where the independent variables (X) and the dependent (Y) variable have a linear correlation. Kinaneva et al. (2021) evaluates and compares LR, multiple LR, polynomial regression, Support Vector for regression and Random Forest regression to estimate the wine quality based on variables such as acidity, residual sugar PH, etc. As they had 11 independent variables simple and polynomial Regression weren't considered. For multiple LR he had to use the backward

elimination technique as the model is sensitive to non-significant values same with SVR. On the Decision Tree method, the researchers found that feature selection wasn't needed as the model performed well with all the variables. The last method that took part in the experiment was Random Forest which gave better results for the type of the dataset and also it was found that as the number of trees in the forest the model gets more accurate. The study included plot visualizations that help understand the current values compared to the predicted however it is hard to see which model was the one with the worst accuracy which would be good to understand which of the models is not suitable for the data at all. Following the regression exploration for sales forecasting,Cheriyan et al. (2018) compares General LR, DT and Gradient Boost Tree to predict the sales of an e-commerce shop using three consecutive sales records. During the data exploration, it was noticeable that there was only one or two variables with positive correlation (quantity and Sales). The model with a better accuracy rate was Gradient Boost Tree (98%) and it is important to say that the dataset includes a huge amount of entrances (85,000) and the final output was an integer. In another study made to predict sales of various groceries Otulets were done by Bajaj et al. (2020) LR, K-neighbors Regressor, XGBoost regressor and Random Forest were compared and it was found that RF was the algorithm with an accuracy of 93,53%. Unfortunately, the authors didn't include a comparison of all the model's results to have a better understanding of the study itself.

2.2 Neuronal Networks

Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN) are well known for their capability to handle lots of data and non-linear data which is pretty normal in business and sales datasets related. Many studies have demonstrated the high accuracy of RNN with seasonal series and long-term predictions, which could be relevant for e-commerce and retailers to see seasonal patterns and trends. A study made by Loureiro et al. (2018) forecast the sales of new items in a fashion company, the only numerical variable was the price of the product, and the data was classified by independent variables which classify the products according to the sales amount (dependent variable) that a product could reach, this classification was done with the marketing team acting as domain experts. The study aimed to explore the use of Deep Neural Networks (DNN) to predict sales and compare it with LR, DT, RF SVR and ANN. The data mining technique applied to tuning the parameters was brute-force grid search, this process was avoided with the LR. The model with the best performance, having the accuracy (R square) as a reference evaluation model, was RF with 75,6% followed by the DNN that obtained 7,27%, and LR was the model with the worst performance which achieved 56,8%. Otherwise, if the models are evaluated based on error-related metrics (MAPE) RF and DNN obtained 34,5% and 37,8% respectively and performed better than the other five models. During the analysis, it was pointed out that even when DNN can have a good performance with both small and large datasets, in real business applications its application can be harder because of the complex training processes needed, otherwise RF can achieve acceptable performance and most time-efficient development. The researchers also referred that the inclusion of experts in the designation and evaluation of the data plays a crucial role in obtaining a forecast that can predict sales with a real business approach.

2.2.1 Support Vector Machine

The SVM are both well known for regression and classification, they work well with datasets that contain different variables. They can handle lineal and non-lineal data providing flexibility during its development. However in the analysis made by Wen et al. (2014) Pointed out that SVM can not deal with non-linear problems, the study pretended to predict the daily sales forecast for grapes with records of a year and the weather which is a high factor that affects fruit sales. The authors compared SVM with ANN and DT. The results showed that SVM (Day Absolute Error 21,24%) outperform the other systems but it doesn't predict the daily sale quantity effectively.

2.2.2 Decision Trees and Random Forest

DT and RF are techniques that split the dataset into smaller and more manageable subgroups which let the model obtain a higher accurate prediction. These methods are commonly used to identify key factors that affect sales through interpretable models. To predict the sales of Walmart, the researchers in Raizada and Saini (2021) compared LR, SVR, Extra Tree Regression, SVM, KNN and RF. Using the sales records of 45 stores for three years, however, the models are run yearly. In addition to the records from the stores they also identify and include external variables like the Consumer Price Index and Temperature. The results of the study show that Extra Tree Regression obtained the lowest Root Mean Squared Error followed by Random Forest. With the Mean Squared Error, the result showed that SVM obtained the highest error score.

2.2.3 Software Sales prediction using Machine Learning.

The sales process could vary depending on what is the product, the cycle of sale grapes has nothing to do with the process of selling a software product, mainly because nowadays the acquisition of software is in service base and it is known as Software as a service and it is similar to a Netflix subscription. Most software companies move to this type of sale cycle as it gives flexibility to the customer. As this study is based on the sales prediction of construction software it was hard to find another study related to, however the study made by Eitle and Buxmann (2019) gave an understanding of the ML models in the sales prediction of Customer Relationship Management (CRM) software. The model was trained with data from almost two years, RF, XGBoost, SVM and CatBoost were the models compared to predict the sales of the company based on the lead stage. The results showed that because of the data type (most of it categorical), the most accurate model was CatBoost. You are expected to provide a critical/analytic overview of the significant literature published on your topic. Comment on the strength and weakness/limitation of work in each reviewed paper.

2.2.4 Forecasting in the Construction field using ML

During the research it was found that other studies related to the construction field have deployed ML to forecast, even when the aim of this document is related to sales, it is also important to understand a little bit about other ML applications in the field. During this exploration it was found that no many studies have been developed around construction software sales forecasting, however, this sparks the creativity of how ML could impact the field. Predicting accurately the duration of a construction project can reduce the

amount of resources involved in the finalization of it and it has an endless number of benefits. In a study made by Wauters and Vanhoucke (2016) compared five models AI: DT, Bagging, RF, Boosting and SVM against the traditional tracking project Earned Value Management (EVM) model commonly used in construction. One of the most common problems with traditional tracking is the low accuracy during the early stages of the project, and the lack of tracking external variables such as weather, social-related changes, supply chain unexpected issues, etc. Unpredictable events are daily-based in the construction field and most of the time their appearance has so short reply window that affects the costs directly. According to the experiment made by the author, the combination of SVM and a fast messy genetic algorithm (fmGA) achieved a 30% more accurate prediction during the initial phase of the project than the prediction made by the traditional mode.

3 Research Methodology

This study aims to help the company sales and marketing team improve current sales strategies by forecasting the number of licenses that customers currently have. The research is based on quantitative data, however, the definition of the variables and data needed was established following the domain experts' feedback.

The research will start with related literature review about ML models deploy to sales forecasting, this will provide insights about what models are suitable according to the data type, followed by the data collection, understanding and pre-processing, finally the models are train, tested and evaluated.

3.1 Company Background

The present study is based on a real company that provides software products and services for the construction, architecture, manufacturing, engineering, education, media and entertainment industries. This product is provided trough annual licenses and the company has a portfolio which includes around 10 to 20 products. The acquisition of these licenses could be done online (website of the company) or through a partner. The Digital sales team is responsible for all the licenses sold online. The company is facing challenges in understanding sales patterns and how to improve the approach from the sales and marketing team to the customers to close sales more effectively. For this reason, the collection of the data is based on some specific variables (employees, revenue, industry) that could provide a better understanding of what specifications are shared by current customers that could help the company to increase its market share.

4 Design Specification

4.1 System Framework

Figure 1 shows the system framework of the present experiment. Including the data collection, which was done using internal and external tools of the company. All the data was imported on Google Colab, to be analyzed using visualizations, after this the data was pre-processed to finally split it in training (80%) and testing (20%) to run the models.

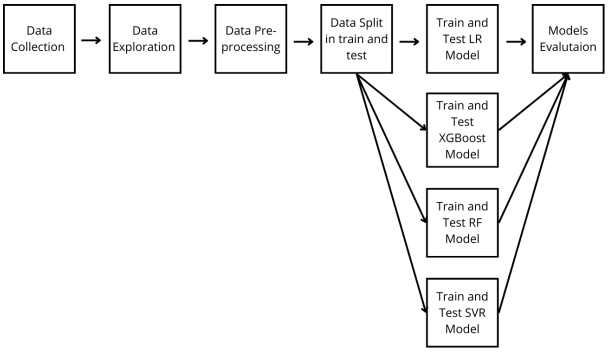

Figure 1: Experiment system framework

4.2 ML Models Selection

XGBoost and RF are regression models that handle predictions with non-linear data, this was mentioned before by Cheriyan et al. (2018) and Raizada and Saini (2021) for this reason, these models are selected to be trained and tested, additionally, the other three ML models were evaluated to understand what model fits better for the data collected, knowing that the dataset is relatively small, with no historical pattern which is a specification that most of the studies reviewed share, additionally the variables are not highly related as the correlation matrix showed. The five ML models selected were LR, RF, XGBoost and SVR. Three other models were tested (Lasso Regression, CNN and NN); however, after the evaluation, the predicted values were far from the real ones, so they are excluded from the present documentation.

4.2.1 XGBoost

This model is recognized for its flexibility with both regression and classification problems, the library implements decision trees in boosting techniques combining them to create a robust model, every new model is built over the mistakes of the previous model and with more weak models added the performance of the model improves.

4.2.2 Random Forest

RF assembles a bunch of decision trees during the training process, each tree is trained with a subset different from the dataset, and this subset is obtained with bootstrappin. Each tree node selects randomly a subset of characteristics to determine the best split and whit this avoids the overfitting to the training data. Once the trees are trained, they make predictions based on the divisions and for regression problems, the model calculates the mean of all the tree (forest) predictions.

4.2.3 Linear Regression

LR is one of the traditional models that predicts the relationship between a dependent variable (target) and one or more independent variables (predicted) through a linear function.

Variable Name	Type	Source
industry	Integer	Internal tool
company name	Object	Internal tool
location	Object	Internal tool
country	Integer	Internal tool
employees	Integer	External tool
url	Object	External tool
revenue	Integer	External tool
csn	Integer	Internal tool
licenses	Integer	Internal tool

Table 1: Data Description

4.2.4 Support Vector Regression

SVR is a variance of the Support Vector Machine (SVM), which is commonly used for regression problems. SVR aims to find the optimal hyperplane approaching a function that predicts continuous values between an acceptable margin.

5 Implementation

The libraries imported to visualize, analyse and manipulate the data were panda, numpy, matplotlibt and seaborn. The equipment used for this research is a MacBook Air 2020, equipped with an Apple M1 Chip and 16GB RAM.

As discussed in Section 4, the tool to process the data and run the models was Google Colab, which is a jupyter notebook hosted, providing access to GPUs and TPUs at not cost. It is commonly used in the ML and data science study areas because of its friendly-user design.

5.1 Data

The dataset includes 1425 entries and nine (9) variables, six (6) of the integer and the rest object. The variables are the name of the customer, the industry, the location and country of the customers, the employees and the revenue of each customer, their website link, the internal number reference and the number of active licenses that each customer has when the data was collected. Table 1 includes the name of the variables, type and source of the data.

It was necessary to transform the type of the variable's revenue, Industry and Country to integer, tables 2, 3 and 4 explain the new entries. This was necessary to have more numerical entries that could train the models and have a better prediction.

5.1.1 Data Collection

The data was collected from June until August 2024, using two main sources. The internal tool is where the company locates all its customers' records and an external tool is named LinkedIn Sales Navigator where records about revenue and employees can be collected. (Table 1 shows the resource of each variable). The dataset's original aim was provided to the Digital Sales team's potential customers; but for this study, customers

Revenue Previous Data	Revenue New Data
500,000 - $1M	1
$1M - $2.5M	2
$2.5M - $5M	3
$10M - $20M	4
$20M - $50M	5
$50M - $100M	6
$100M - $500M	7
$500M - $1B	8
1B	9

Table 2: Revenue data adjusted

Industry Previous Data	Industry New Data
Architecture	1
Construction	2
Design and Manufacturing	3
Engineering	4
Media and Entertainment	5

Table 3: Industry data adjusted

Country Previous Data	Country New Data
North European Countries: Denmark, Sweden, Norway, Finland, Ireland, United Kingdom, Estonia	1
South European Countries: Spain, Italy, Greece, Portugal, Croatia, Malta, Bosnia, Cyprus, Slovenia.	2
West European Countries: Germany, France, Luxembourg, Netherlands, Belgium, Austria, Switzerland.	3
East European Countries: Poland, Hungary, Czechia, Liechtenstein, Lithuania, Romania	4

Table 4: Country data adjusted

	industry	country	employees	revenue	csn	licenses
count	1425.000000	1425.000000	1425.000000	1425.000000	1.425000e+03	1425.000000
mean	2.733333	2.583860	74.937544	4.059649	5.195390e+09	6.990175
std	1.403413	1.264308	90.362610	2.018512	2.382350e+08	16.205186
min	1.000000	1.000000	3.000000	1.000000	7.022792e+07	1.000000
25%	1.000000	1.000000	24.000000	2.000000	5.108341e+09	1.000000
50%	3.000000	3.000000	42.000000	3.000000	5.138901e+09	1.000000
75%	4.000000	4.000000	93.000000	5.000000	5.155740e+09	6.000000
max	5.000000	4.000000	1000.000000	10.000000	5.501840e+09	376.000000

Figure 2: Basic Statistics to understand the variables

without purchase records were excluded from the dataset, as this can affect the model predictions. Additionally, it was found that some customers have bought licenses, but they are expired, thus only the active number of licenses were included to have a real-time perspective.

5.2 Data Understanding

To have a better understanding of the variables, they were summarised using basic statistics such as count, mean, std, min and max. Figure 2 shows this summary and it shows that licenses that is the dependent variable of this study have a mean of 6.99, meaning the average amount of active licenses that a company has. This analysis was done using the Python command: *df.describe().*

The Revenue and the number of employees can indicate the size of the customer, during the data exploration phase it was found that the majority of the customers (792) in the dataset have revenue between $1M-$50M. And the amount of employees of most of the customers is 17-36. However, the majority of the licenses number is 1 (737) and 2 (111) meaning that there is a huge sales gap between the number of employees and the amount of licenses that a company get. This found, will help the digital sales team to address conversations with current customers to increase the amount of licenses bought. Regarding marketing understanding, the customer industry, shows a clear pattern of good company market share, as most of the customers with active licenses are in the architectural industry (See Image 3), saying this the marketing team could take advantage of this and create campaigns that attract new customers using current customers reviews. On the other hand, the company with a lower amount To understand patterns and identify potential outliers, especially in the dependent variable (licenses) the data was visualized using a boxplot. Figure 3 shows the distribution of the number of licenses by industry. Additionally, Figure 4 shows the correlation matrix that helps to visualize and summarize the relationships between the numerical variables.

5.3 Pre-Processing

The dataset doesn't have missing values or duplicates, however during the exploration phase it was identified that it has outliers that to train the model have to be handled. Two methods of pre-processing (Transformation and winsorizsation) were carried out and compared to see which one would fit better for the models. Winsorization reduces the impact of the outliers replacing them for range values without removing them, the winsorirization only modifies the outliers while the transformation modifies all the values not only the outliers. During the evaluation phase, the models achieved a higher per-

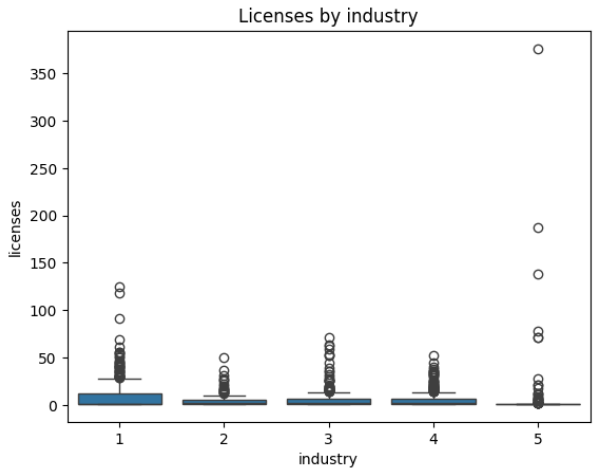

Figure 3: Number of Licenses by industry boxplot

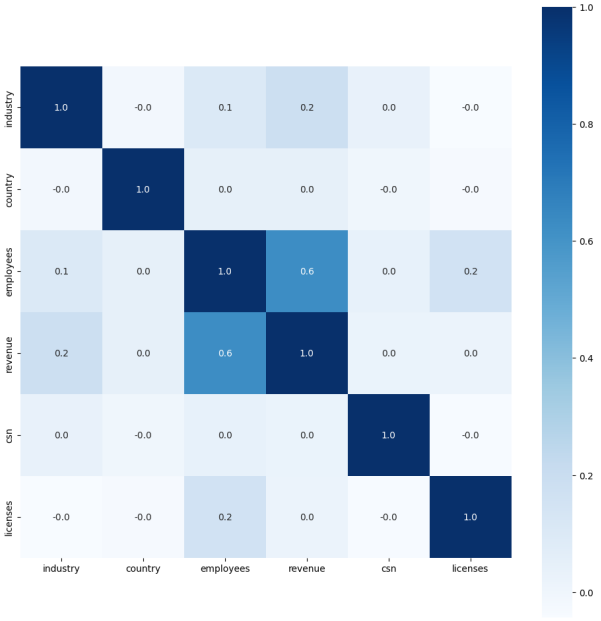

Figure 4: Correlation matrix

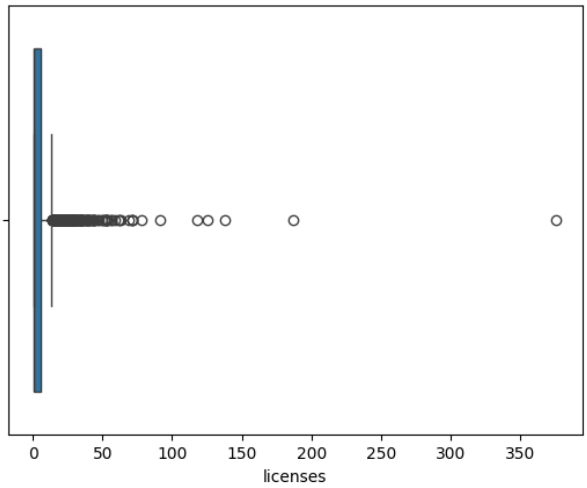

Figure 5: Licenses boxplot before the data transformation

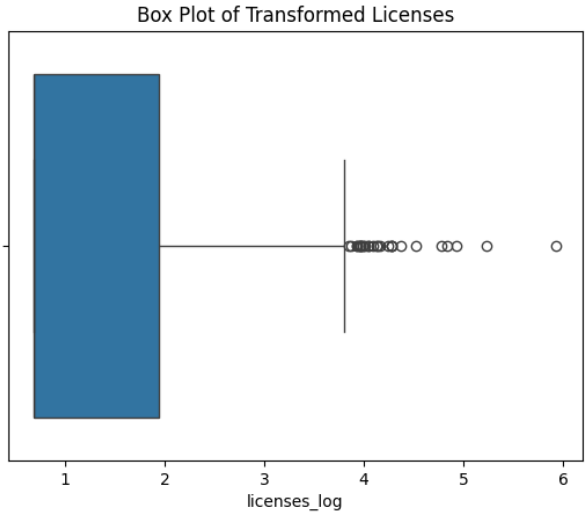

Figure 6: Boxplot of transformed licenses

formance with the transformation method. Figure 5 and 6 show the boxplot before and after, respectively, the transformation.

With the data preprocessing affected the next step is the selection of the model, training and evaluation.

5.4 Python Workflow

With the machine Learning models (LR, RF, XGBoost and SVR) libraries imported from the "Sklearn" the Python workflow for each model is the following:

- **First:** Data preparation; in this step, the dependent variable (Y=licenses) is extracted from the Dataframe while the other variables that are settled are the independent variables.
- **Second:** The categorical data is handled using one-hot encoding. The non-numeric data in the independent variables is identified to encode them.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Figure 7: MSE formula

- **Third:** The code integrates the original numeric data with the created encoded categorical data.
- **Fourth:** The data is split into 20% testing data while the remaining 80% is training data.
- **Fifth:** Each model is created and trained using the training data the model is trained. With this, the model will be able to learn the relationship between the features and the dependent variable (licenses)
- **Sixth:** After the training, the model is used to predict the target variable using the test data.
- **Seventh:** MSE and R-squared methods are used to evaluate the model evaluated

The following link inlcudes the Google Colab notebook where the data was process and the models where train.

6 Evaluation

To assess the models but also the pre-processing of the data, the evaluation metrics used were Mean Square Error and the R-Squared, all four models included in the present report and the other three models that were included (CNN, NN and Lasso Regression) were evaluated and compared with objective to understand which combination between pre-process method (Transformation or Winsorization) performs better with which ML model and the dataset. The selection of these two metrics for the present study is because of their frequent use to evaluate regression problems, additionally, most of the studies reviewed previously mentioned (Raizada and Saini (2021) and Bajaj et al. (2020)) use them too.

6.1 Evaluation Metrics

MSE: This evaluation metric calculates the accuracy of the model, taking the mean of the square errors, which means the difference between the predicted values and the real values. The metric calculates the difference between each real value and the model value, then to eliminate the negative signs and give more weight to the bigger errors, the differences are squared and finally, the mean of these squared errors is calculated. Lower values indicate that the model has low errors meaning better predictions. The MSE formula is as follows:7

R-Square: Measure how well the predicted values fit with the variability of the real values by comparing the square residuals to the sum of the total squares. Higher values indicate that the model explains better the variability of the data meaning a good fit. The r-square formula is as follows:8

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Figure 8: R-Square formula

Model	R-Squared	MSE
XGBoost	0.999628	0.000390
Random Forest	0.999437	0.000591
Linear Regression	0.579199	0.441817
SVR	-0.370939	1.439406

Table 5: Results Comparison

6.2 Results

As mentioned before, a total of seven ML models were released to do the present study, however because of their low performance three of them were excluded in the present project. The results show that the best performer model combined with the data transformation method, in both metrics R-Squared and MSE, was XGBoost and the score (the highest score) was XGBoost followed by RF.Table 5 shows all the model scores ranked by the highest R-Squared performance.

6.3 Discussion

XGBoost and RF obtained an incredible fit between the predicted data and the original one, while the accuracy of the model outstanding LR and SVR showed a very poor performance. This low performance could be because of their low fit with the data type, in the LR model or because of its sensitivity to data characteristics or hyperparameter settings in the SVR model.

7 Conclusion and Future Work

The present study aimed to train an ML model that helps the digital sales team of a construction software company to predict the sales of its licenses, using real and updated data on the company sales. This objective was completed successfully, showing that XGBoost and RF are models that perform outstanding results when comparing the values predicted with the original values (active licenses). During the data recollection and processing it was found that data related to sales forecasting needs to be transformed to get more accurate models, Without this transformation the forecasting wouldn't be viable. Additionally, the experiment phase comparing different ML models to predict the sales helps to have a better understanding of the data type and what could be the best combination between ML model, preprocessing method and evaluation method to achieve an accurate prediction of the sales. From the business understanding, this study will help the digital sales and marketing team to narrow their efforts and strategies to overachieve their sales target based on what industries, what type of companies and what is the location of customers with higher buying potential. Following the purpose of the

current study, the variables collected provide a general understanding to train the model according to current customer data, however for future work, and knowing that sales are a complex environment that could be affected by so many actors, add more data related with an economy index such a Gross Domestic Product, the unemployment rate in each industry, Foreign Direct Investment, etc. Including this data to train the model could have more accurate sales forecasting.

References

Bajaj, P., Ray, R., Shedge, S., Vidhate, S. and Shardoor, N. (2020). Sales prediction using machine learning algorithms, *International Research Journal of Engineering and Technology (IRJET)* **7**(6): 3619–3625.

Cheriyan, S., Ibrahim, S., Mohanan, S. and Treesa, S. (2018). Intelligent sales prediction using machine learning techniques, *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, IEEE, pp. 53–58.

Eitle, V. and Buxmann, P. (2019). Business analytics for sales pipeline management in the software industry: A machine learning perspective.

Kinaneva, D., Hristov, G., Kyuchukov, P., Georgiev, G., Zahariev, P. and Daskalov, R. (2021). Machine learning algorithms for regression analysis and predictions of numerical data, *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, IEEE, pp. 1–6.

Loureiro, A. L., Miguéis, V. L. and Da Silva, L. F. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail, *Decision Support Systems* **114**: 81–93.

Raizada, S. and Saini, J. R. (2021). Comparative analysis of supervised machine learning techniques for sales forecasting, *International Journal of Advanced Computer Science and Applications* **12**(11): 102–110.

Wauters, M. and Vanhoucke, M. (2016). A comparative study of artificial intelligence methods for project duration forecasting, *Expert systems with applications* **46**: 249–261.

Wen, Q., Mu, W., Sun, L., Hua, S. and Zhou, Z. (2014). Daily sales forecasting for grapes by support vector machine, *Computer and Computing Technologies in Agriculture VII: 7th IFIP WG 5.14 International Conference, CCTA 2013, Beijing, China, September 18-20, 2013, Revised Selected Papers, Part II 7*, Springer, pp. 351–360.