Graduation Project Report: Medical Representative Sales Analysis Using Machine Learning

---

Project Title

Medical Representative Sales Analysis Using Supervised Machine Learning

Team Members

- AbdelRahman AbdelHalem Helal Ahmed
- Mazen Mohamed Elsayed Elsafty
- Mazen Mahmoud Shaban Mohamed
- Rashad Mohamed AboElmkaram Shehab
- Mazen Mohamed Mohamed Ahmed
- Karim Ahmed Farhat Eid

Supervisor

Eng. Sherif Said

Track

IBM Data Science (ALX1_AIS3_S1e)

Organization

Digital Egypt Pioneers Initiative

Abstract

Medical representatives play a vital role in the pharmaceutical industry by bridging the gap between pharmaceutical companies and healthcare professionals. This project focuses on analyzing the factors that influence doctors' prescribing behaviours and developing a supervised machine-learning model to predict the likelihood of doctors prescribing drugs from a local pharmaceutical company. Using the AdaBoost algorithm, the model achieved an accuracy of 82% and an F1-score of 87%. Key insights regarding factors such as doctor age, specialty, and location were also analyzed, providing useful guidance for optimizing sales strategies.

---

1. Introduction

1.1 Background

Medical representatives are responsible for promoting and selling pharmaceutical products to healthcare professionals. These representatives must convince doctors to prescribe their company's products in a highly competitive market where multiple companies offer similar alternatives. Despite having the same active ingredients, doctors may choose different products based on various factors.

This project aims to leverage machine learning techniques to predict doctors' prescribing behaviours based on their profiles. By analyzing features such as **specialty**, **age**, **location**, and **class**, we develop a model that predicts whether a doctor will prescribe one of the drugs produced by a local pharmaceutical company.

1.2 Problem Statement

Medical representatives face a challenge in persuading doctors to prescribe their company’s drugs when several other brands offer the same active ingredients. Understanding the factors that influence prescribing decisions will allow companies to optimize their representatives' efforts. Therefore, the primary problem this project addresses is predicting whether a doctor, based on specific features, will prescribe a drug from the local company.

1.3 Objectives

The key objectives of this project are:

- To determine the factors influencing doctors’ prescribing behaviours.
- To analyze the relationships between different variables, including doctor age, specialty, and location.
- To build a supervised machine learning model capable of predicting whether a doctor will prescribe a local company’s generic drugs.

---

2. Literature Review

Several studies have examined how pharmaceutical companies can improve their sales strategies by understanding prescribing patterns. Studies indicate that doctor's specialty, experience, and location significantly impact prescribing habits. Various machine learning algorithms, including Decision Trees, Random Forest, and AdaBoost, have been applied to similar problems in healthcare. AdaBoost, known for enhancing weak learners, is one of the most efficient algorithms for classification tasks with imbalanced data.

---

3. Methodology

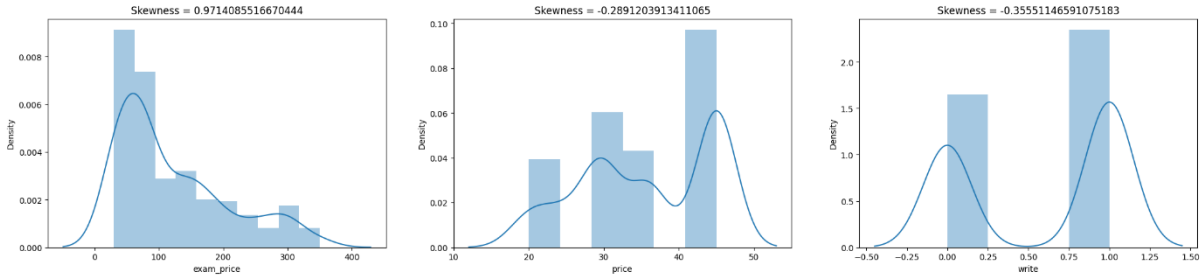
3.1 Data Collection

The dataset for this project was sourced from two tables: `medicine_table` and `doctor_table`. These tables contained details about doctors, their specialties, locations, and prescribing behaviours.

3.2 Data Preprocessing

Data preprocessing involved the following steps:

- **Handling Missing Values**: Missing data were imputed using mean, or dropping techniques.
- **Feature Encoding**: Categorical features such as specialty and location were encoded using one-hot and label encoding.
- **Normalization**: Continuous variables like age and years of experience were normalized using MinMaxScaler to bring all variables into comparable ranges.
- **Skewness**: Checked skewness of every column which was acceptable so no change was required.



---

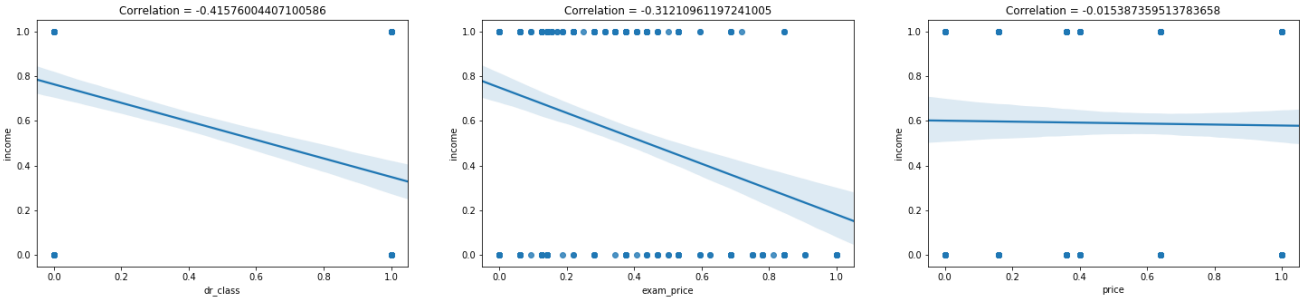
4. Exploratory Data Analysis (EDA)

4.1 Key Visualizations and Insights

4.1.1 Correlation Heatmap

A correlation plots were generated to identify relationships between the features. Key findings:

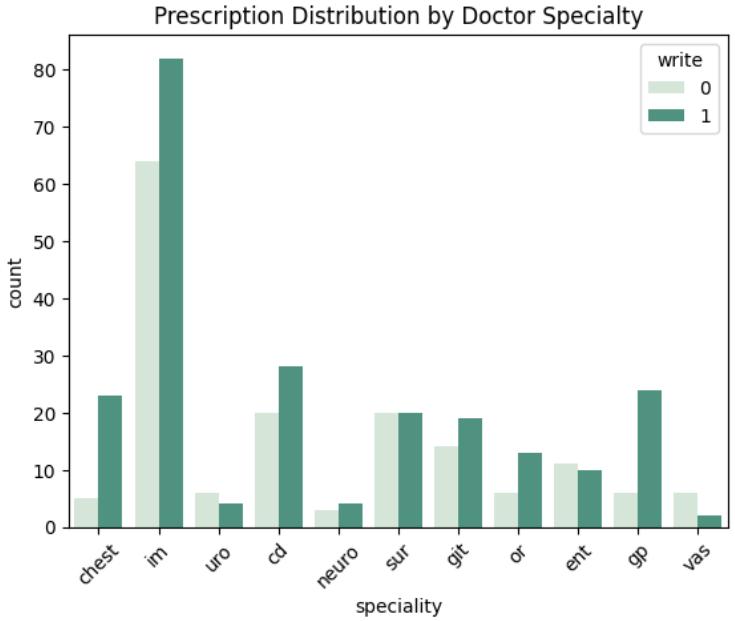
- **Doctor's Class and Examination Price** were highly correlated with prescribing behaviour.
- **Drug Price** showed a poor correlation with prescribing behaviour..

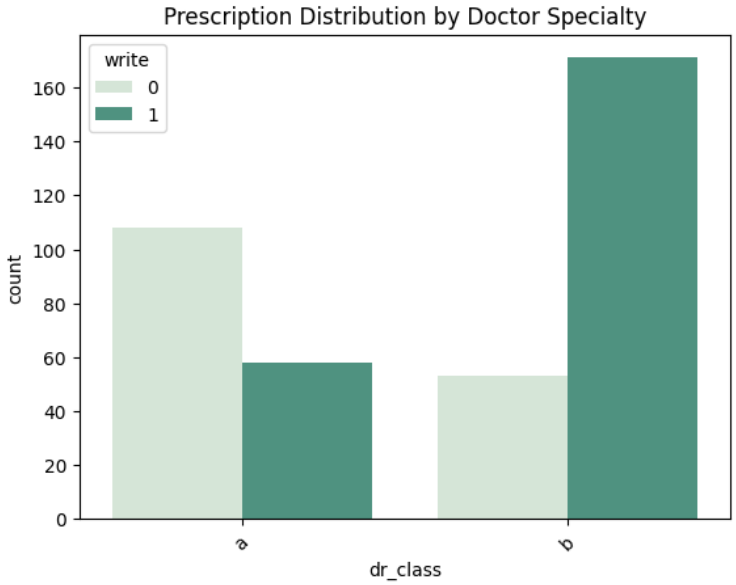


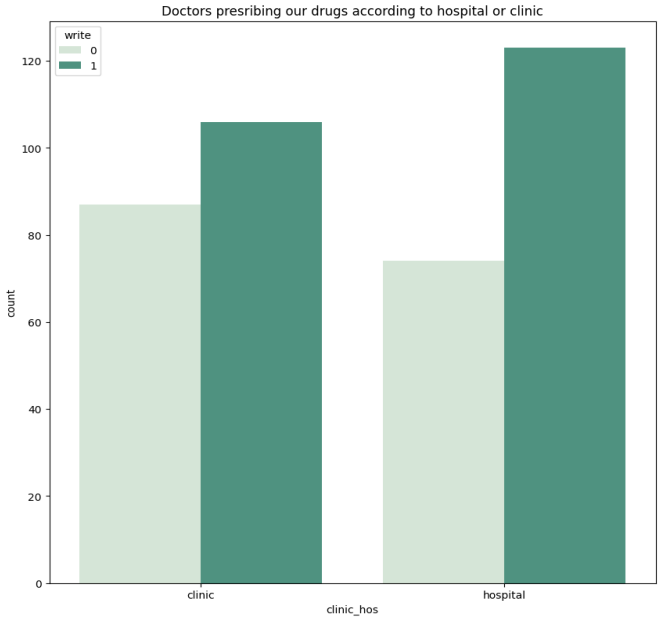
4.1.2 Bar Charts and Distributions

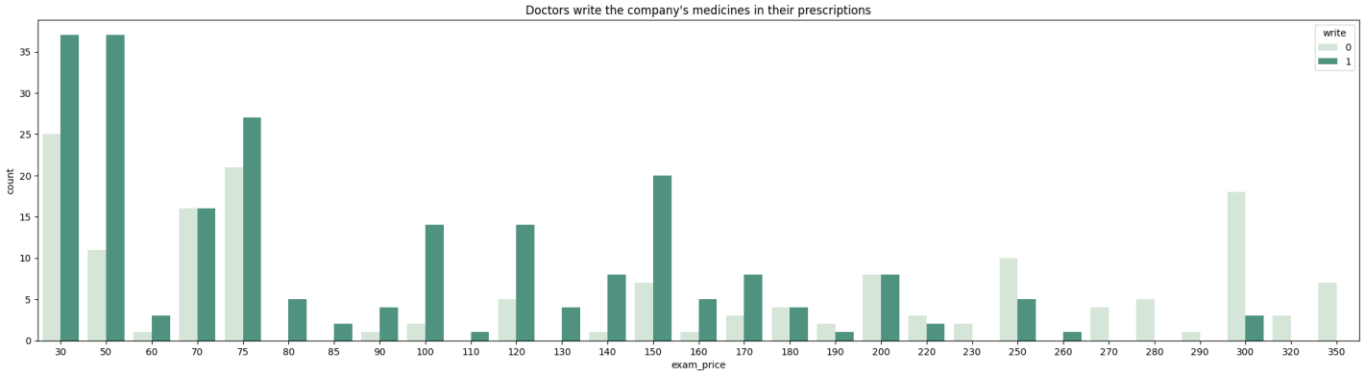
Bar charts were used to visualize the distribution of categorical features:

- **Specialties**: General Practitioners (GPs) and Chest specialized doctors were found to be more likely to prescribe the local company's drugs.
- **Class Distribution**: Class B doctors were the most frequent prescribers of generic drugs.
- **Drug Price**: prices like (50,100,120,150) were found to be more likely to prescribe the local company's drugs. Unlike prices higher than or equal to 250.
- **Place of work**: Clinic doctors were found to be more likely to prescribe the local company's drugs while Hospital doctors tend to be neutral.




Prescription Distribution by Doctor Specialty
write
0
1
count
dr_class
a
b


Doctors presribing our drugs according to hospital or clinic
write
0
1
count
clinic_hos
clinic
hospital

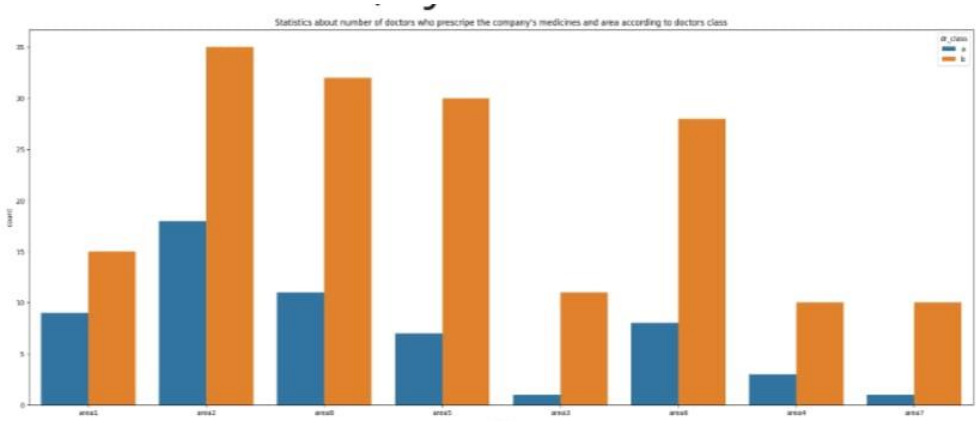

Doctors write the company's medicines in their prescriptions
write
0
1
count
exam_price

The previous graphs shows that:

1-Doctors with class 'b' are the most ones who prescribe the company's medicines for their patients.

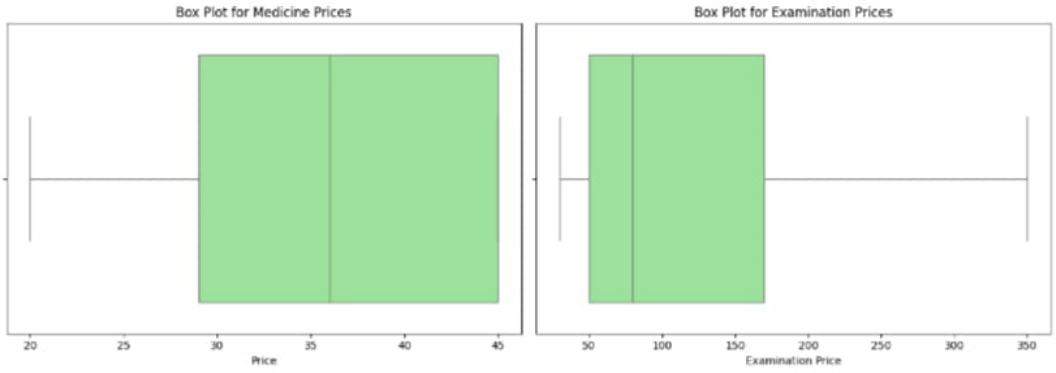
2-Doctors who work at hospitals are little bit more than doctors who work at clinics in prescriping the company's medicines for their patients

3-Doctors whose examination price is '30' or '50' are the most ones who prescribe the company's medicines for their patients, followed by '75'.



This bar chart shows that doctors of class 'b' are the most ones who prescripe the company's medicines for their patients at all areas

Which means that young doctos or doctors who are at the beginning of their career are more expected to prescripe the company's medicine than famous or expert doctors



Interpret the box plots Price

this box plot indicates that the prices of medicine are fairly consistent, with a median around 35 and a range from 20 to 45, without any extreme outliers.

Examination Price

The width of the box indicates the variability in examination prices but overall, this box plot indicates that examination prices have a median around 100, with a range from 50 to 200, and

show some variability but no extreme outliers. This suggests a consistent pricing structure for examinations within this dataset

4.2 Feature Importance

The most important features of the AdaBoost model were:

- **Doctor's Class**: The most influential factor in predicting prescription behaviour.
- **Doctor's Examination Price**: Doctors with relatively low examination price were more likely to prescribe local company products.

---

5. Model Development

5.1 Model Selection

We tested several machine learning models, including Decision Trees, Random Forests, Gradient Boost ,and Support Vector Machines. After extensive experimentation, the **AdaBoost algorithm** was selected due to its superior performance.

5.2 Model Training

The dataset was split into an 80:20 ratio for training and testing. The AdaBoost model was trained on the training set, with hyperparameters optimized through grid search. We evaluated the model using metrics such as accuracy, and F1-score.

5.3 Model Performance

The final AdaBoost model achieved:

- **Accuracy**: 84.6%
- **F1-Score**: 88.2%

---

6. Conclusion

This project successfully applied supervised machine learning to predict doctors' prescribing behaviours. By analyzing features such as **class**, **specialty**, and **examination price**, the AdaBoost model achieved an accuracy of 84.6% and an F1-score of 88.2%. These results can help pharmaceutical companies optimize their marketing strategies, save a lot of resources, and better target healthcare professionals.

---

7. Future Work

Future research could explore the following:

- **Model Improvement**: Further tuning of the AdaBoost model and testing other advanced models could improve accuracy.
- **Expanded Dataset**: Including more doctors and additional features, such as prescription frequency would improve model robustness.
- **Real-Time Recommendations**: Developing a real-time recommendation system for medical representatives based on the model's predictions.

---

8. References

- Scikit-learn documentation: https://scikit-learn.org/
