Healthcare Partners Inc

Check Amount Predictive Model

Contents

1. Introduction

Healthcare Partners Inc is a group purchasing organization in the healthcare industry, that helps members (the buyers) pool the purchases of goods and services, from suppliers. The GPO Organization, which is Healthcare Partners, Inc. provide its members with immediate access to leveraged pricing. In return, charges an administrative fee from the members and commissions from suppliers for all purchases made by the members.

This report is the summary of the data science project undertaken to create a predictive model for forecasting monthly revenue.

2. Problem statement

How can the forecasting team at Healthcare Partners predict Administrative Fees forecast for the next Fiscal Year, with 80% accuracy for its market types, Food Services, Surgical and Pharmaceuticals, so that the senior leadership can prepare the company budget to present in 3 months.

3. The Data Science Method (approach)

This section provides the systematic explanation of how the problem was approached from data collection to model evaluation.

3.1. Data Collection

A single CSV file was provided to the data science project. The data that was included in the CSV file was extracted from multiple systems.

3.2. Data cleaning and preprocessing

The data set was essentially capturing details for each contract for the length of the contract, which also presented the future months. Therefore, there were several blank rows in the dataset.

Additionally, there were several other data points that were not relevant to the project, such as names of personnel and contract details. All the non-essential rows are data points were removed before loading the dataset to the project.

Preprocessing

This included dealing with missing values, imputation and removing any column that had over 50% values missing.

As the final step of preprocessing, the ‘Market Category’ features were simplified into a fewer important categories.

3.3. Exploratory Data Analysis (EDA)

This step helps identify characteristics and patterns, such as the check amounts received through each month during the Fiscal Year, and the variation of check amounts for different market categories.

Various visualizations help identify patterns. The visualization, in fact showed there is significant variance in the amounts based on the market category.

A critical finding from the EDA was that the check amount varies quite significantly within the market category.

3.4. Feature Engineering

To prepare the dataset for machine learning models, two feature engineering techniques were used.

One-Hot Encoding:

This project is considered a regression project. For better performance it is important that categorical features are represented as binary vectors. Additionally, with hope of improving model performance, One-Encoding technique was applied to *Market category*.

Label encoding

Contract Frequency, a categorical feature, that represents some kind of order, that also needs to be maintained. By using the label encoder in SciKit-Learn, the Contract frequency was assigned a unique number

sine-cosine transformation

This technique was used on the due month as I needed to maintain the time aspect to capture when Check Amount was received.

3.5. Model selection

This project is a regression machine learning problem, attempting predict a continuous numeric value - check amount in the comping Fiscal Year. Therefore, a potential solution is a predictive model, that can accurately predict or estimate the numeric target value.

I considered three (3) models

Linear Regression Model.

A very simple model. This is also a choice for a quick prototype.

Random Forest Model.

The reason for considering this model is its ability to handle outliers and its use of multiple decision trees to produce reliable predictions.

Gradient Boosting Regression Model

Like the Random Forest, this is another ensemble learning technique.

3.6. Model training

With all the data preparation tasks, my dataset is ready to be trained. Three different models were identified and in this step, I defined the features and the target variable and split the dataset to training and test.

Once the train and test sets are defined, fed the training data into the model and evaluated using various different metrics.

In this step, I also performed hyperparameter tuning and re-trained the two models, Random Forest and the Gradient Boosting Regression.

3.7. Model evaluation

This is the final step before I make the final selection of the model for further development.

Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared measures were used to compare the performance.

Cross validation was also applied to evaluate the performances of all three models

4. The results

Model	MSE	RMSE	MAE	R-squared	Cross-Validation Score
Linear Regression	3.23e+16	1.80e+08	1.15e+08	0.447	-3.41e+16
Random Forest	5.31e+15	7.29e+07	3.17e+07	0.909	-1.07e+16
Gradient Boost	4.12e+15	6.42e+07	2.64e+07	0.929	-1.21e+16

Based on the model performance results **Gradient Boost Regressor** seems to be the best-performing model, followed by Random Forest. Linear Regression seems to have higher prediction errors compared to the ensemble models.

5. Findings

- Based on the model performance, my findings are as follows:
- The most influential feature (predictor) is the Market Category. However, based on the feature important analysis, year and contract frequency which is how often the check payments are received seem highly influential.
- Of the various market categories, Services is the biggest contributor.
- The contracts that are set up to pay annually has the highest contribution, when compared to contracts that are setup to pay monthly and quarterly.

Future work

- The model performance in my opinion is moderate. I believe the performance can be improved and is proposed as future work.
- For the lack of experience and understanding I realized feature engineering was not successfully implemented when training the models.
- As future work I suggest feature engineering specifically scaling and imputation to better handle the huge variances that are shown as outliers.
- Additionally apply binning on the check amount to see Better and more feature engineering techniques – imputation, scaling, binning and bucketing