International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 12 | December - 2022 Impact Factor: 7.185 ISSN: 2582-3930

APPLICATION OF REVENUE MANAGEMENT USING MACHINE LEARNING

Y. Sai Pranay, V.Varun, B. Athirath
Mr. N. Venkata Subba Reddy (Guide)
Sreenidhi Institute of Science and Technology
Hyderabad

ABSTRACT:

It is important to every Business to understand its revenue so as to make as much as possible from the customers. It is also important to view the Business in every aspect that exists and evolve according to the changing environments. Our project involves the dataset consisting of all the transactions of a Business from which we analyze the revenue aspects of it and provide the overview of the revenue generated by the business using multi-linear regression technique of machine learning.

INTRODUCTION

In order to maximise earnings and optimise pricing, firms must use revenue management as a crucial strategy. Python's pandas module is a helpful tool for data analysis and manipulation to support revenue management choices in the field of Machine Learning.

Business Understanding Problem Statement

A sizable children's educational toy firm that sells gaming consoles and edutainment tablets online and in brick-and-mortar stores needed to evaluate the consumer information. They have been in business for a while and maintain all transactional data. The provided data is in csv form.This is an example of customer-level information that was taken from multiple sets of transactional files which can be used for analysis and prediction of the model.

Data Preprocessing:

Preparing raw data for analysis is a crucial part of the data science process, which is known as data preprocessing. It is required because unprocessed data must first be cleaned and converted before it can be properly evaluated and used to inform decisions. Unprocessed data is frequently inaccurate, noisy, and inconsistent.

The components which we use for data preprocessing are as follows

Missing values:

Missing values are frequently present in raw data, which can have an impact on the analysis's correctness and dependability. Data preparation involves locating and addressing missing values, for as by imputing estimates for missing values or removing rows with an excessive number of missing values.

Outliers:

An outlier is a data point that differs noticeably from the rest of the data and can have a disproportionately large impact on the study. Outliers must be recognised and handled during data preprocessing, possibly by being taken out of the dataset or being altered in some way.

Inconsistent data:

Inconsistent raw data, such as those with varying formats or units of measurement, might occur frequently. Standardizing and normalising data to make it consistent and simple to analyse is a part of data

WHAT IS REGRESSION?

To simulate the relationship between a dependent variable and one or more independent variables, regression is a statistical technique. In order to minimise the difference between the anticipated values and the actual values, it entails estimating the values of the coefficients in the linear equation that best fits the data.

Multiple Linear Regression:

Simple linear regression can easily be extended to include multiple features. This is called multiple linear regression:

$y=\beta 0+\beta 1x1+...+\beta nxn$

Each x represents a different feature, and each feature has its own coefficient.

$y=\beta 0+\beta 1\times NoOfGamesBought+\beta 2\times Frequency\ Purchase+\beta 3\times NoOfUnitsPurchased$

OLS Regression Model :

Regression analysis using OLS (Ordinary Least Squares) models the relationship between a dependent variable and one or more independent variables. In this type of linear regression, the difference between the predicted values and the observed values is minimised by predicting the coefficient values in a linear equation that best fit the data.

The OLS approach includes reducing the sum of the squared residuals, or the discrepancies between the anticipated and observed values. To ensure that positive and negative residuals are treated equally and to prevent cancellation of positive and negative residuals, the residuals are squared. The OLS estimator is the name of the derived equation.

In OLS regression, the goal is to model the relationship between a dependent variable (Y) and one or more independent variables (X1, X2, ..., Xp).

The model is represented by the following equation:

$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \betapXp + \varepsilon$

where β0 is the intercept term, β1, β2, ..., βp are the coefficients for the independent variables, and ε is the error term.

Coefficients :

The coefficients (β1, β2, ..., βp) represent the expected change in the dependent variable for a one unit change in the independent variable, holding all other variables constant.

Intercept :

The intercept term (β0) represents the expected value of the dependent variable when all independent variables are equal to zero.

Error :

The error term (ε) represents the difference between the observed response and the response predicted by the model. It is assumed to be normally distributed with a mean of zero.

Residual Sum of Squares:

To estimate the coefficients for the independent variables, the OLS regression model minimizes the sum of the squared errors between the observed responses and the responses predicted by the model. This is known as the residual sum of squares (RSS), and it is represented by the following equation:

$RSS = \Sigma(Yi - \hat{Y}i)^2$

where Yi is the observed response, and Ŷi is the response predicted by the model.

WORKING OF OLS Model

Working of OlS Model is divided into following steps :

Collect and prepare the data: The first step is to collect and prepare the data for analysis. This involves cleaning the data to handle missing values, outliers, and inconsistent data, and selecting the appropriate independent and dependent variables.

Split the data into training and test sets: It is generally a good idea to split the data into a training set, which will be used to fit the model, and a test set, which will be used to evaluate the model's performance.

Fit the model to the training data: The next step is to fit the OLS model to the training data. This involves estimating the values of the coefficients in the linear equation that best fit the data. There are several ways to do this, including using matrix algebra or an optimization algorithm.

Evaluate the model's performance: Once the model has been fit to the training data, it can be evaluated using the test set. This typically involves calculating the mean squared error (MSE), which is a measure of how well the model fits the data, and other metrics, such as the R-squared value, which measures the proportion of variance in the dependent variable that is explained by the independent variables.

Make predictions: Once the model has been fit and evaluated, it can be used to make predictions about the dependent variable based on new values of the independent variables.

ADVANTAGE OF OLS

OLS is a well-known and widely-used method, making it simple to obtain materials and information on it.

OLS has a number of advantageous statistical characteristics, including impartiality, consistency, and effectiveness. As a result, OLS estimates are the most accurate among all unbiased estimators, are unbiased, and have the lowest variance. They also become more accurate as the sample size increases.OLS is usable by a variety of people because it is reasonably easy to apply and comprehend.OLS is a versatile technique for estimating regression models since it can be used to a range of data kinds and distributions.In order to describe more

complex relationships, OLS can be expanded to include more characteristics including interaction terms, polynomial terms, and dummy variables.

STEPS FOR IMPLEMENTATION

- Preparing Data Set
- Importing Libraries
- Splitting the Dataset into Training and Testing Modules
- Training the dataset with OLS model
- Evaluating model performance

PREPARING DATASET

The dataset we use for this project is csv files which is a structured dataset. Structured data: Structured data is organized in a tabular format, with rows representing individual data points and columns representing features or attributes. Structured data is often stored in a spreadsheet or a database and can be used for tasks such as classification or regression.

CSV files are useful in machine learning because they are a simple, widely supported file format for storing and exchanging tabular data. They can be easily read and written using various software tools and libraries, and are often used to store datasets for training and evaluating machine learning models. CSV files are convenient for sharing data and importing it into machine learning frameworks.

IMPORTING LIBRARIES

In this project, we're utilizing a variety of libraries, including Matplotlib, NumPy, sklearn.metrics and statsmodels.api.

Matplotlib: It is a visualization library that is used to present the results in more understandable ways i.e., in graphs and plots.

Sklearn.metrics: It is a popular library for machine learning in Python that provides a wide range of tools for tasks such as classification, regression, clustering, and dimensionality reduction. The sklearn.metrics module contains a collection of evaluation metrics and scoring functions that can be used to evaluate the performance of machine learning models.

Statsmodels.api:It is a Python library for estimating, testing, and analyzing statistical models. It provides a range of functions for tasks such as linear regression, generalized linear models, and time series analysis.

Splitting the dataset into Training and Testing Modules

You can adjust the model's hyperparameters and make sure that it is not overfitting to the training data by dividing the dataset into training and testing sets. When a model is overly sophisticated and able to fit the training data exactly but is unable to generalise to new data, this is known as overfitting. You may make sure the model can generalise to new data and is not merely memorising the training data by utilising a testing set.

"train_test_split" is a function in the Python scikit-learn library that is used to split a dataset into two subsets: a training set and a testing set. The training set is used to fit a model, while the testing set is used to evaluate the performance of the model.

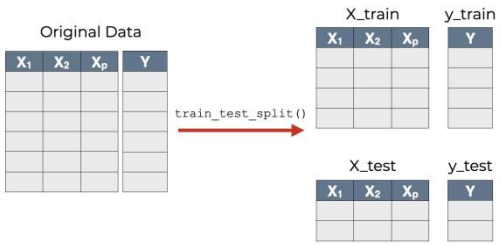

fig-1

Training the Dataset with OLS Model

OLS(Ordinary Least Squares) identifies the variables that reduce the total squared residuals between the predicted and observed values.

We must first prepare the data by separating it into input features (X) and a target variable before you can train an OLS model on it (y). The next step would be to develop a linear regression model and fit the training data to the model using the fit method.

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.947
Model:	OLS	Adj. R-squared (uncentered):	0.947
Method:	Least Squares	F-statistic:	3556.
Date:	Tue, 20 Dec 2022	Prob (F-statistic):	0.00
Time:	19:00:43	Log-Likelihood:	-12389.
No. Observations:	2391	AIC:	2.480e+04
Df Residuals:	2379	BIC:	2.487e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	7.1098	1.140	6.235	0.000	4.874	9.346
x2	4.9253	0.565	8.717	0.000	3.817	6.033
x3	-1.3348	0.411	-3.247	0.001	-2.141	-0.529
x4	-0.0011	0.009	-0.120	0.905	-0.019	0.017
x5	11.0663	0.516	21.426	0.000	10.053	12.079
x6	9.1829	0.409	22.451	0.000	8.381	9.985
x7	0.0030	0.001	3.696	0.000	0.001	0.005
x8	-0.0482	0.017	-2.861	0.004	-0.081	-0.015
x9	-11.9959	0.328	-36.572	0.000	-12.639	-11.353
x10	12.9611	2.935	4.416	0.000	7.205	18.717
x11	-14.4603	2.413	-5.993	0.000	-19.192	-9.729
x12	2.0585	3.650	0.564	0.573	-5.099	9.216

Omnibus:	655.460	Durbin-Watson:	1.926
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3699.657
Skew:	1.175	Prob(JB):	0.00
Kurtosis:	8.623	Cond. No.	9.62e+03

fig-2

Evaluating Model Performance

One method to check it is by looking at the R^2 value. This gives us the information about how much variance in our target could be explained by all independent variables. The closer R^2 to 1 the better the model is. This tells us about goodness of our model

Next we check how close our predictions are to the actual values. The difference between actual and predicted is called error. For a good model the error should be as small as possible. There are various kinds of error metrics. In this case we use mean squared error computed as 1/n*∑(y-(yhat))^2 where y are the actual values, what are the predicted values and n is total number of records

Residual vs Fitted plot

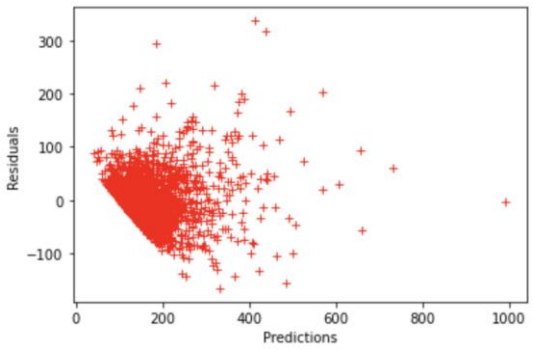

fig-3

Residuals and Leverage

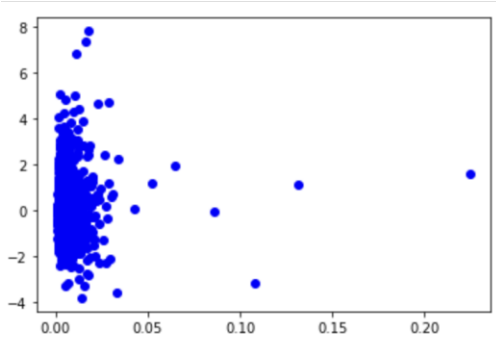

fig-4

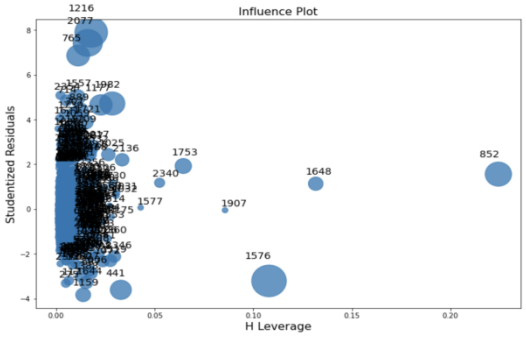

fig-5

RESULTS

After dividing the dataset into training and testing data and passing them to the OLS regression model all the coefficients of the independent variables can be attained which are used to predict the dependent variable.

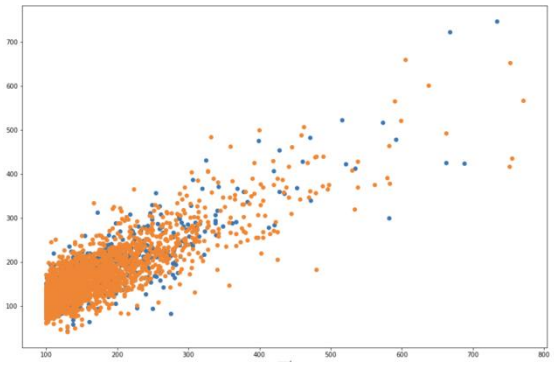

fig-6

The orange colour in the scatter plot is the training set data. Blue colour represents the testing set and data distribution between actual and predicted data .

MERITS

When it comes to merits, the higher accuracy of our project makes it more reliable for decision-making. It makes it easy for businesses to predict the amount of resources required and the amount of goods that are to be manufactured so as to maximize the profit .

LIMITATIONS

Forecasting client demand and real-time availability and price adjustments are all part of the data-driven business strategy known as revenue management. The goal is to maximise revenue. While revenue management has the potential to be a potent instrument for boosting earnings, it is not without its drawbacks.some of them are :

- Complexity
- Limited Flexibility
- Limited Visibility
- Customer Dissatisfaction
- Limited Scope

SOFTWARE AND HARDWARE REQUIREMENTS

To engage in machine learning, we will need certain hardware and software components. The hardware requirements include a processor with multiple cores, a sufficient amount of memory, and adequate storage space. The software requirements include a modern operating system, a Python distribution, and machine learning libraries. The specifics of these requirements may vary depending on the type and complexity of the machine learning we plan to do, and we may need to use specialized hardware or cloud services for some tasks.

CONCLUSION

Forecasting client demand and real-time availability and price adjustments are all part of the data-driven business strategy known as revenue management. The goal is to maximise revenue. It can be a potent instrument for boosting revenue and enhancing operational effectiveness. However, putting revenue management into practice may be difficult and time-consuming, and organisations need to be aware of the limitations of this approach. Businesses must have the appropriate software and hardware solutions, as well as the required information and resources, to support their efforts in revenue management. In order to leverage the advantages of this business approach, it is crucial for companies to carefully evaluate

the revenue management limits and develop solutions to these problems.

ACKNOWLEDGMENTS

We would like to express our special thanks to our mentor Mr.N Venkata Subbareddy who gave us a golden opportunity to do this wonderful project on this topic which also helped us in doing a lot of research and we came to know about so many new things. We are really thankful to them.

Secondly, We would also like to thank my friends who helped us a lot in finalizing this project within the limited time frame.