Restaurant revenue prediction using deep learning.

Mokhutli Letsae

Abstract

Predictive analytics is a powerful tool that enables businesses to forecast metrics such as future revenue and profitability by analyzing historical data and market trends. This capability allows businesses to anticipate upcoming financial performance and proactively address potential profit shortfalls. For instance, they can launch targeted marketing campaigns to boost revenue and mitigate losses. Ultimately, predictive analytics helps businesses optimize their operations and make informed decisions to maximize profitability.

This write-up documents a data science project aimed at predicting monthly revenue for restaurants using the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework. The project utilized a deep learning model, specifically a multi-layer perceptron neural network, to build the predictor. The training data for the neural network model was sourced from Kaggle's "Restaurants Revenue Prediction" dataset. The project demonstrates how advanced machine learning techniques can be effectively applied to real-world business problems, providing actionable insights that enhance decision-making and operational efficiency.

By following the CRISP-DM methodology, the project involved several key phases:

Business Understanding: Defining the project objectives and requirements from a business perspective.
Data Understanding: Collecting and exploring the dataset to gain insights and identify patterns.
Data Preparation: Cleaning and transforming the data to make it suitable for modeling.
Modeling: Building and training the multi-layer perceptron neural network using the prepared data.
Evaluation: Assessing the model's performance to ensure it meets the business objectives.
Deployment: Implementing the model in a real-world setting to start generating predictions.

The results of this project highlight the effectiveness of deep learning models in predicting restaurant revenue, demonstrating significant potential for improving financial forecasting and strategic planning in the food-selling industry.

Introduction

This project implemented in python 3.8 on Jupyter notebooks using CRISP-DM aims to predict monthly restaurant revenue using a variety of predictors such as review count, marketing spend, and average menu price, with the goal of helping restaurant owners anticipate and mitigate potential losses through targeted campaigns hence maximizing profit. Utilizing a dataset from Kaggle, a Multi-Layer Perceptron (MLP) neural network was implemented due to its robustness and superior accuracy compared to other models like ensemble methods, XGBoost, and Random Forests. Through careful exploratory data analysis (EDA), data preprocessing, model training, and evaluation, the MLP demonstrated its efficacy in forecasting revenue, making it the preferred choice for this regression task. The project's findings underscore the potential of data science to enable proactive business strategies in the restaurant industry.

1.Methodology

1. Dataset Overview

The dataset with usability of 10 was downloaded in a **csv** format from Kaggle (https://www.kaggle.com/datasets/mrsimple07/restaurants-revenue-prediction).

The file had 1,000 rows and 8 columns and no missing values. The 8 columns' roles, data types, value count, and definitions are listed on the below table.

Attribute	ML-model data role	data type	value count	Attribute description
Number_of_Customers	Predictor	int	1000	The count of customers visiting the restaurant
Menu_Price	Predictor	float	1000	Average menu prices at the restaurant
Marketing_Spend	Predictor	float	1000	Expenditure on marketing activities
Cuisine_Type	Predictor	string	1000	The type of cuisine offered (Italian, Mexican, Japanese, American).
Average_Customer_Spending	Predictor	float	1000	Average spending per customer
Promotions	Predictor	float	1000	Binary indicator (0 or 1) denoting whether promotions were conducted

Reviews	Predictor	float	1000	Number of reviews received by the restaurant
Monthly_Revenue	Target	float	1000	Simulated monthly revenue, the target variable for prediction

2. CRISP-DM

Crisp-DM framework was used as a methodology for this project. The framework acted as a guideline for optimal attempt of the project.

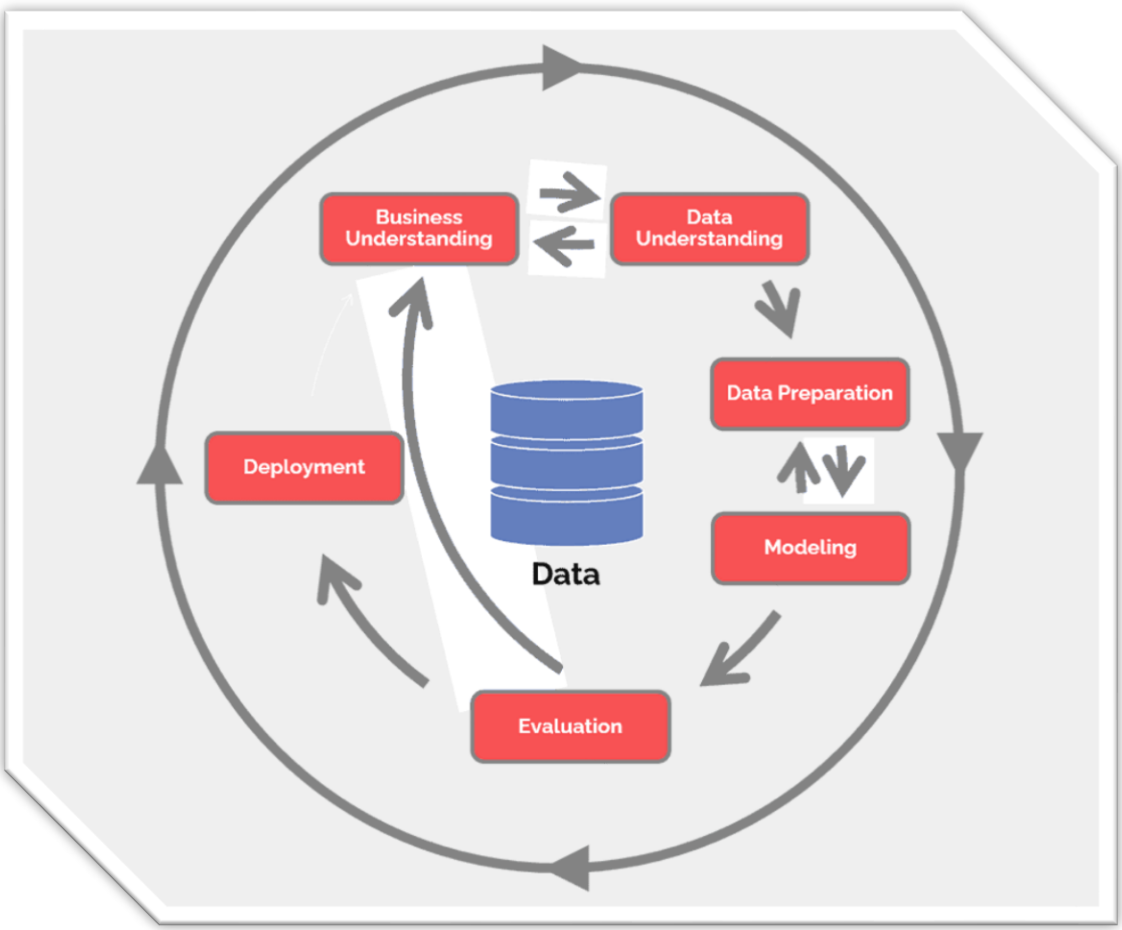


Figure 1

Figure 1 shows the stages of the framework, from the figure you can see that the framework follows a cyclic flow implying agility. The project was built in an agile manner.

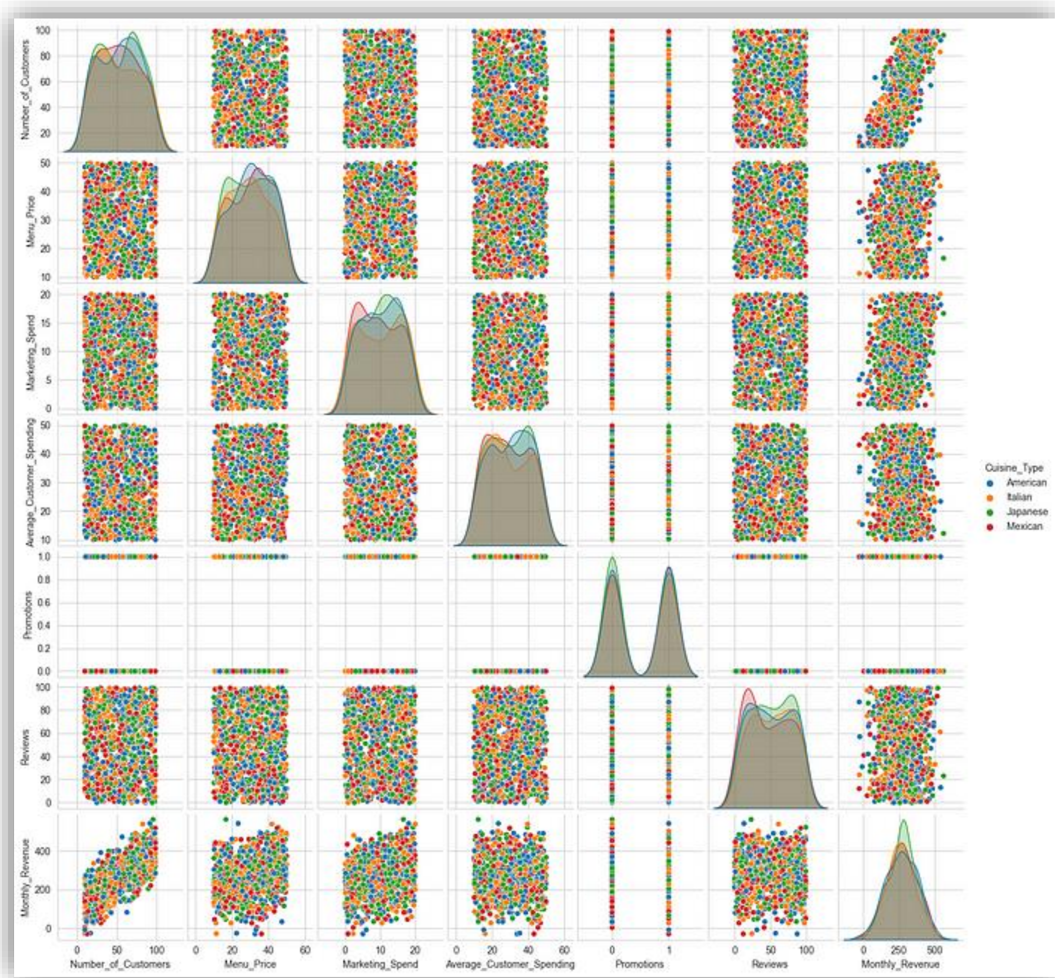
a. **Business understanding:** The business problem in this project is predicting monthly revenue for a restaurant. The value of forecasting revenue is that the restaurant owner will be able to avoid losses and maximize profits by putting in place targeted campaigns in seeing that the revenue forecast will produce a loss.

b. **Data understanding (EDA):** While establishing facts and insight, below is what was extracted from the data.

   1. From the table below, we observe that on average, 53 customers visit the restaurant per cuisine. Our dataset includes various types of cuisines, and we have expressed all statistical metrics per cuisine. This means that the metrics in this table represent averages across all cuisines.

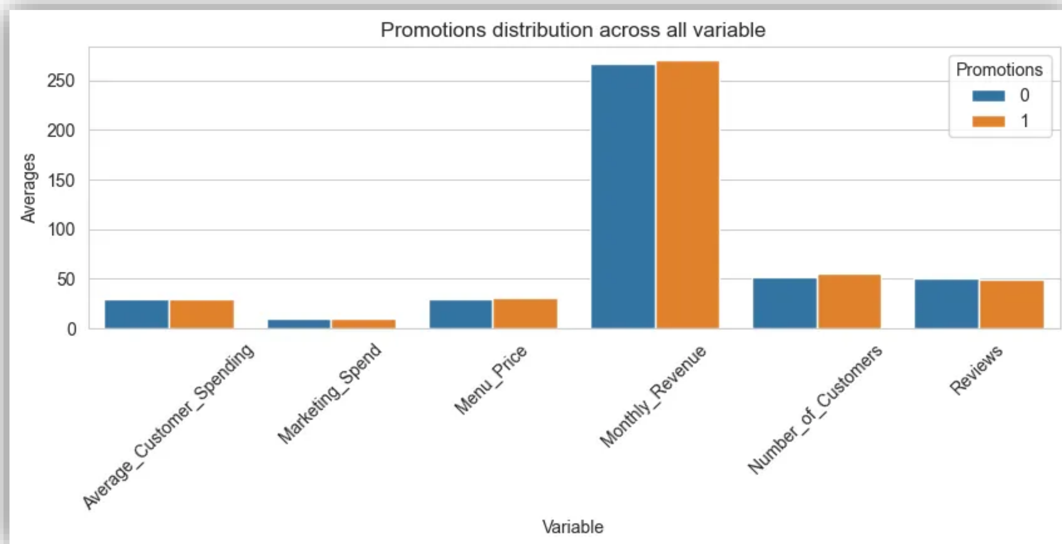
	Number_of_Customers	Menu_Price	Marketing_Spend	Average_Customer_Spending	Promotions	Reviews	Monthly_Revenue
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	53.271000	30.219120	9.958726	29.477085	0.497000	49.837000	268.724172
std	26.364914	11.278760	5.845586	11.471686	0.500241	29.226334	103.982950
min	10.000000	10.009501	0.003768	10.037177	0.000000	0.000000	-28.977809
25%	30.000000	20.396828	4.690724	19.603041	0.000000	24.000000	197.103642
50%	54.000000	30.860614	10.092047	29.251365	0.000000	50.000000	270.213964
75%	74.000000	39.843868	14.992436	39.553220	1.000000	76.000000	343.395793
max	99.000000	49.974140	19.994276	49.900725	1.000000	99.000000	563.381332

2. Another important aspect of understanding our data would be checking the relationships of our variables and their distributions, to achieve this we can use a pair-plot.

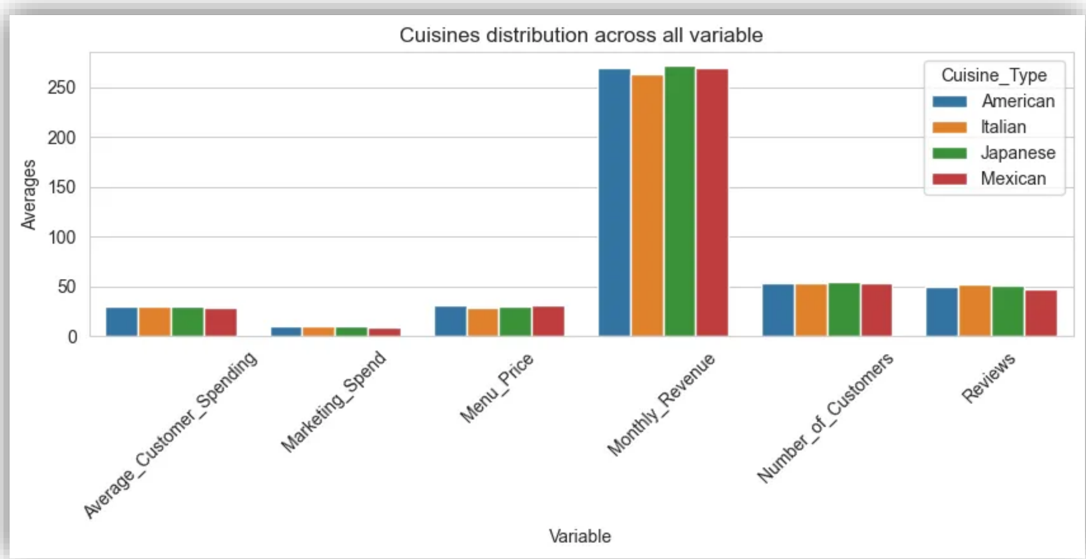


What we see from the above pair-plot is that no relationships exhibit linear relationships except for number of customers and monthly revenue of which is also not totally linear but forms a linear-like. The relationships are cluster relationships.

3. With the below histogram, we check if promotions have any impact on monthly revenue. We can extract that promotions seem to have a very little impact on monthly revenue as seen from the histogram, there's a very little difference between revenue made when items are promoted and not. To make promotion have impact we can increase marketing spend on promotions as marketing is the heart of promotions.



4. Lastly, there seems to be very small differences in pricings of cuisines hence the near the same monthly revenues. Japanese and Italian cuisines should have a higher pricing compared to the other 2 as they known to be high-end and expensive.



c. **Data Preparation:** For this stage, basic processes for data processing and preparation and cleaning are performed:

Handling missing values - There were no missing values from the dataset hence there was no imputations or dropping of values performed.

Addressing outliers - The data did not have may outliers, but IQR was performed on monthly revenue to remove outliers, and there were rows with negative revenue which were dropped.

Feature engineering – One-hot encoding was used on nominal variables in which this case we had one variable being cuisine type.

Feature selection - Variance and correlation threshold techniques were used to select features **Number_of_Customers, 'Menu_Price, Marketing_Spend**.

d. **Model Building:** Two models were developed(scikit-learn) for comparison: a Deep Learning model (multi-layer perceptron) and XGBoost regressor. The multi-layer perceptron regressor outperformed the XGBoost regressor and it was used for predictions.

e. **Evaluation:** Evaluation of the model's performance metrics.

The below performance metrics are for the MLP (multi-layer perceptron) neural network. They show that the model does not overfit as there's very little difference on all the training and test metrics.

The R-squared is 0.67 which defines how well the predictors well predict our target variable, the monthly revenue. The average mean absolute error is 44 which is acceptable but not optimal and our error percentage is 35%. All this metrics are acceptable to use and echo good performance of the model.

	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	Mean Absolute Percentage Error	R-squared	Adjusted R-squared	Explained Variance Score	Median Absolute Error
Test	3,101.19	55.69	44.67	34.62	0.67	0.66	0.67	38.25
Train	3,343.92	57.83	46.17	26.93	0.69	0.69	0.69	38.99

f. **Deployment:** The model artifacts can be deployed on on-prem or cloud platforms, and it can be utilized using a web pass, desktop app or just an environment it is used in.

2.Conclusion

The project was undertaken as a capstone project for completing a nanodegree from Udacity. It was undertaken by Mokhutli Letsae(https://www.linkedin.com/in/mokhutliletsae/), has the code repo on GitHub(https://github.com/letsaemokhutli/capstone_project.git). The MLP neural network model demonstrates good performance with an acceptable level of accuracy. The metrics suggest that the model does not overfit, as evidenced by the minimal difference between training and test metrics. While the error percentage indicates there is room for improvement, the model's predictions are reliable enough for practical use in decision-making processes for the restaurant.

3.Future Work

Augmenting data (features and row count) will be the next step to improve the model performance metrics. Using a cloud environment is also part of future intentions.

4.Acknowledgements

I would like to acknowledge my boss, Itumeleng Senekane (Executive Head: CVM and Big data), Vodacom Lesotho for giving me a chance to take this Udacity course. Chatgpt has been a very useful resource in the completion of the project hence it's owners I would like to acknowledge. I would also like to finally Acknowledge the whole Vodacom group for giving us a chance into a more than a glimpse of data science.