**Restaurant Sales Prediction**

*A report submitted in partial fulfillment of the requirements for the Award Degree of*

**BACHELOR OF TECHNOLOGY**

in

**COMPUTER SCIENCE AND ENGINEERING**

By

**CHANDAKA BHARATH KUMAR Reg. No. :22B91A0539**

Submitted to

**Sri R Shiva Shankar**

(Assistant Professor)

# &Dr

# DNSB Kavitha

(Assistant Professor)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING S.R.K.R. ENGINEERING COLLEGE(A)**

SRKR MARG, CHINNA AMIRAM, BHIMAVARAM-534204, A.P

(Recognized by A.I.C.T.E New Delhi) (Accredited by NBA & NAAC) (Affiliated to JNTU, KAKINADA)

**SAGI RAMA KRISHNAM RAJU ENGINEERING COLLEGE**

(Autonomous)

## **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**



# CERTIFICATE

This is to certify that the **ML-CONNECT** Report titled **“RESTAURANT SALES PREDICTION”** is the bonafide work done by **Mr/Mrs. CHANDAKA BHARATH KUMAR** bearing **Register Number: 22B91A0539** in the second year first semester at **SRKR Engineering College, Bhimavaram** from 25th Novemeber 2023 to 26th November 2023 in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering.

**Dr V Chandra Sekhar Sri R Shiva Shankar Dr DNSB Kavitha**

**(Professor and Head) (Assistant Professor) (Assistant Professor)**

**ABSTRACT:**

We imported Pandas ,numpy,seaborn and matplotlib libraries and warnings module.Pandas is a Python library used for working with data sets.It has functions for analyzing, cleaning, exploring, and manipulating data.Pandas allows us to analyze big data and make conclusions based on statistical theories.numpy is a Python library used for working with arrays.matplotlib is a python library used to create 2D graphs and plots by using python scripts and It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc.warnings module is a built in python library that allows you to control the display of warning messages,To turn off warning messages,we use filterwarnings() function.Python Seaborn library is a widely popular [data visualization](https://www.simplilearn.com/data-visualization-article) library. You build it on top of the matplotlib data visualization library and can perform exploratory analysis.We mount the drive and then load the datasets using the read\_csv function into a dataframe naming it sales\_df to read comma separated values files and copy it to a backup just in case and displayed using head().After examining the data set we decided to remove an unnecessary column using del backup['Restaurant ID'] and since there are only continuous integer values and no strings there is no need of encoding and Scaling is needed to speed up the model which is not a necessary feature .We derive the correlation matrix to plot a figure using pyplot by specifying the figsize which is a heatmap using sns so that we can analyse the correlation between features and we set the fmt value to .2f and cmaps =”greens” so that the cells are green and the annot=True so that the cells are give numbers instead of being just a coloured cell,and we fixed the linewidths.Then we analyse all our existing data using describe(),shape(this gives out length and width),info(),dtypes..which give out description,shape,information,datatypes of the dataframe.All the following are done to backup dataframe and then we find the duplicate vales and place them first later we drop(remove them) and then reset the indexes.using isnull().sum() we find total number of null values in each feature ,since we have no null values there is no need of any imputing NULL values. We separate target and features by assigning them to y and x from sklearn.model\_selection import train\_test\_split as tts and then It splits into x\_test,x\_train,y\_test,y\_train and then we display the shape.After importing all the regression models from sklearn ,based on the R2 score of the 3 regression models we selected linear regression model.

**TABLE OF CONTENTS:**

* **Loading data set**

* **Data visualisation and correlation models**

* **Imputing null values**

* **Encoding**

* **Scaling**

* **Data analysis**

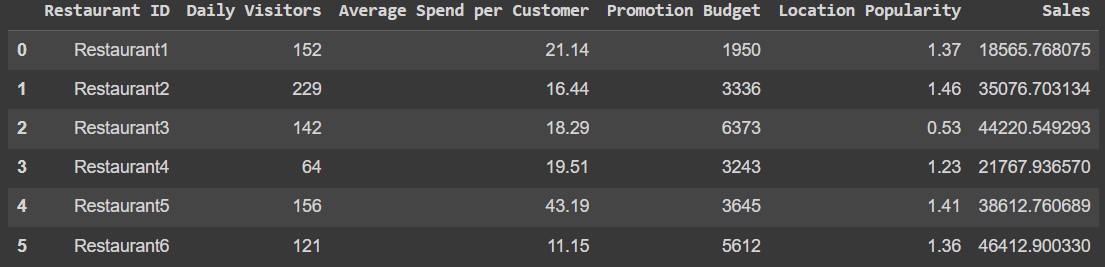
* **Preprocessing**

* **Splitting**

* **Building ML models and evaluation**

**LOADING THE DATASET:**

* We imported pandas as pd since it has functions for analyzing, cleaning, exploring, and manipulating data
* Then we imported numpy as np which is a Python library used for working with arrays.
* matplotlib is a python library used to create 2D graphs and plots by using python scripts.
* pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc
* The warnings module allows you to control the display of warning message and python Seaborn library is imported for [data visualization](https://www.simplilearn.com/data-visualization-article)
* Then we display the data frame by using head() function OUTPUT:

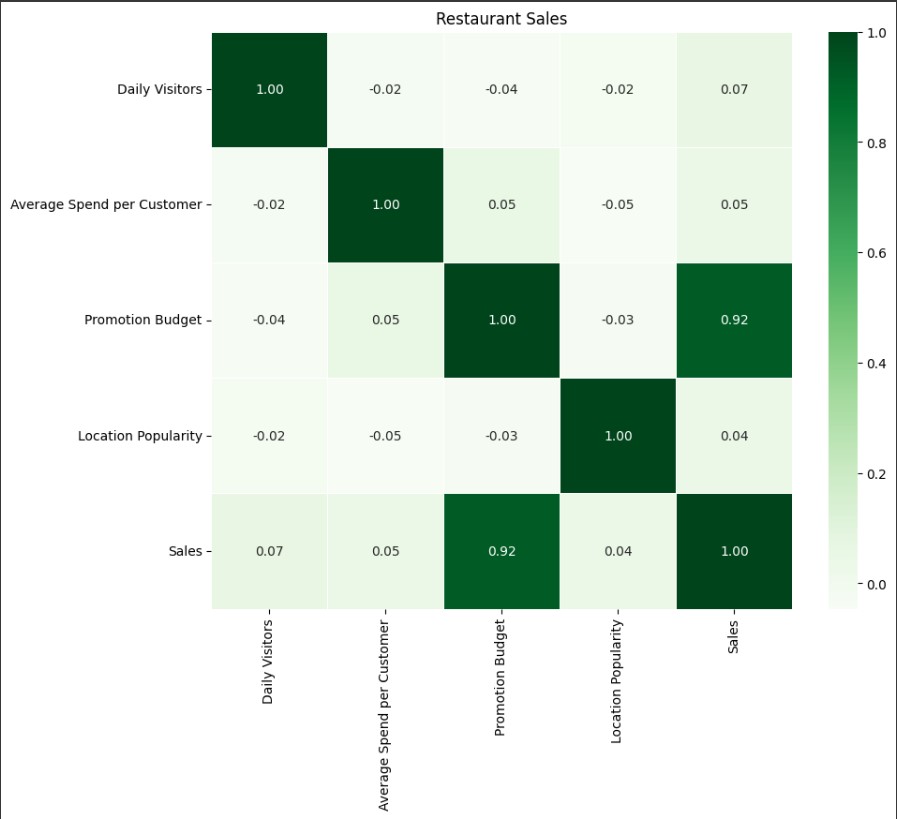


**DATA VISUALIZATION AND CORRELATION ANALYSIS:**

We derived correlation matrix between the features of the dataset using corr() function and then plotted a figure by specifying a figsize.

we pulled up a heatmap from sns(seaborn lib) since it is used for data visualization and modified it using colormaps and annot=True and set title to “Restaurant Sales”

Output:



**IMPUTING NULL VALUES:**

There are no NULL values in the dataset

**ENCODING:**

No need of encoding since the dataset contains data which are of numerical values

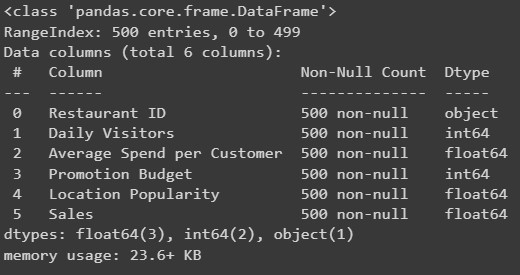
**SCALING:**

Scaling can help speed up convergence and improve model performance but it is not strictly necessary in case of regression .

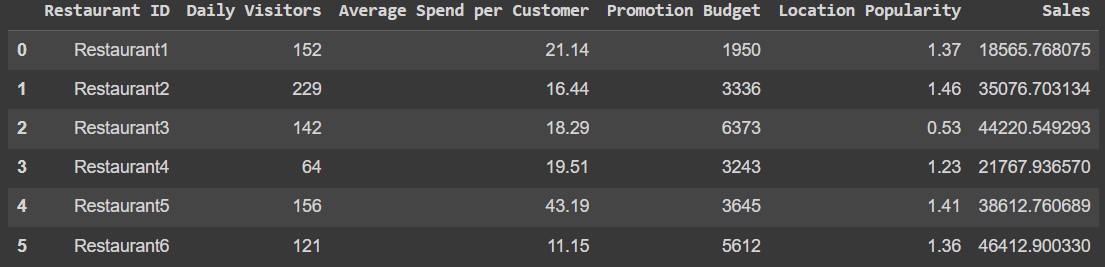
**DATA ANALYSIS:**

We analyse our data by printing the information ,datatypes,description and shape(length,width) of our dataset

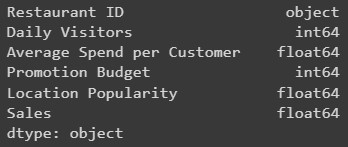
Output:(for info())



For describe()



For dtypes output:



For shape output:



PREPROCESSING:

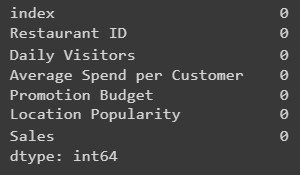
* we place any existing duplicate values at ‘first’,and then reset index after deleting the dupicates.

output:



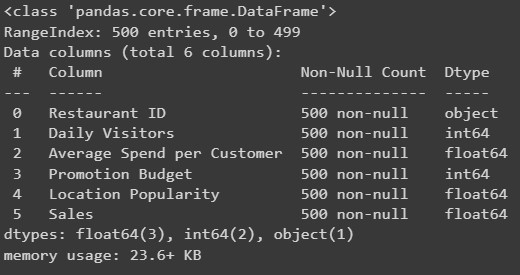
(this deletes duplicates and resets the index)

* then to find out the number of null values in each feature of dataset
* using isnull().sum() we check for any NULL values output:



* we delete the unnecessary feature using del keyword
* then we check the dataset again using info() to see if feature got deleted.

output:



Later we separate the target and features using a for loop for the columns and features by assigning them to y and x respectively we do this by creating a list for the features and keep adding the features unless it is equal to ‘Sales’

**SPLITTING:**

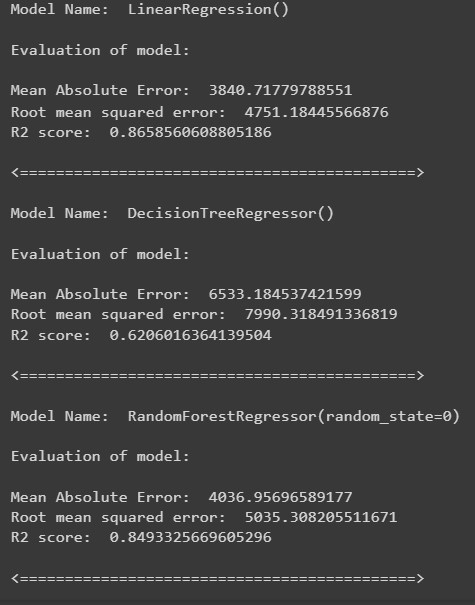
We divide the dataset into 2 sections/parts for testing and training,we declared test\_size=0.3 that is 30% is used for testing and other 70% for training,and we display the shape of all 4 x\_train,x\_test,y\_train,y\_test.

**Output:**

**BUILDING ML MODELS AND EVALUATION:**

We import LinerRegression,Decision Regression,RandomForestRegressor from sklearn.ensemble ,slearn.tree ,sklearn.linear\_model respectively .Here we test the accuracy of all 3 models to select the one with higher accuracy and then we predict the mean\_squared\_error,mean\_absolute\_error and R2 score by importing them from sklearn.metrics here R2 score is the accuracy of the model

**Output:**



**CONCLUSION:**

Here the R2 score (i.e.. accuracy) is more for linear regression hence this is the best model which suits our dataframe .