Understanding Sales Data by Regression Models

June 2022

SpringBoard Data Science Career Track Program

1. Summary

1.1 Problem statement

With the growing demand in the market, a Superstore giant likes to have a better understanding of what factors matter to their sales, what works best for them, and which products, regions, and categories they should target or avoid, so they would improve the forthcoming sales strategy.

1.2 Data

For this project, I obtained data from Kaggle. It contains the sales data collected between Jan. 2014 and Jan. 2018 across the 49 states in the US.

The dataset includes 9994 entries with 21 columns.

This capstone project contains four parts: data cleaning, exploratory data analysis, data preprocessing, final model evaluation, and model selection.

1.3 Questions of interest

By analyzing the sales data from the year 2014-2018 year, this project is trying to build a multiple linear regression model to predict sale prices.

I have trained both linear regression models and random forest models to find out which features have the most influence in determining the target variable – 'sales'.

The main interest in this project is to understand the key factors driving sales, such as:

- What are the most profitable cities/states/regions?
- What are the categories in which the products made the most contribution to sales?
- Which factors, discounts/postal code/quantities, significantly influence the sales?

The intended stakeholders of this project would be owners, suppliers, investors, and any sort of retail store which uses sales as a performance benchmark would find this project to be interesting and meaningful.

2. Exploratory Data Analysis (EDA)

All data is loaded and stored in a Pandas dataframe. It looks clean and has no missing values as it is taken from Kaggle.

Exploratory data analysis consists of analyzing the main characteristics of a data set using **visualization methods** and **summary statistics**.

Use **pandas.DataFrame.info()** method to print a summary of the data frame.

2.1 Data Type

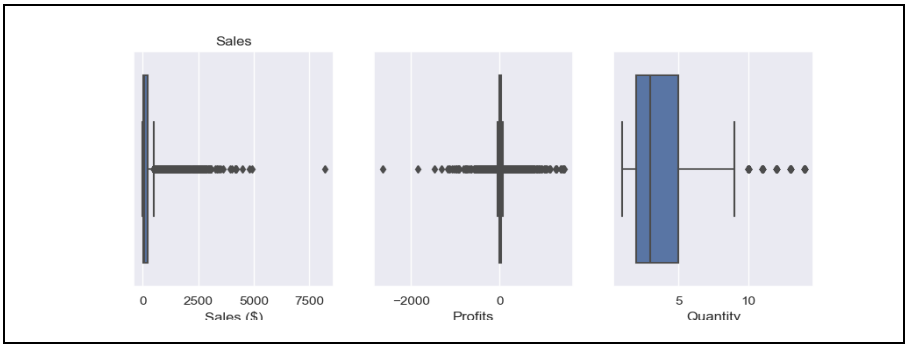
I change the datatypes of ‘Order date’ and ‘Ship Date’ from ‘object’ to ‘datetime’ so there would be no error when I plot the data later on.
A new feature’ month’ is extracted for plotting and visualization purposes.

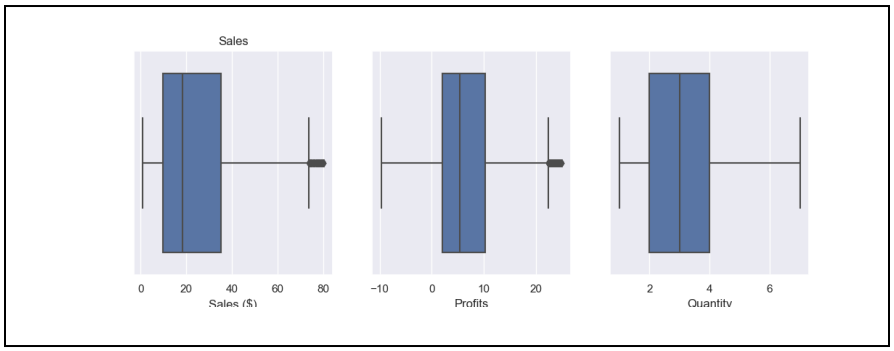
2.2 Drop the less useful columns

Columns of 'Order Date', 'Ship Date', 'Product Name', 'Order ID', 'Customer ID' were dropped as they have weak correlation with the target variable.

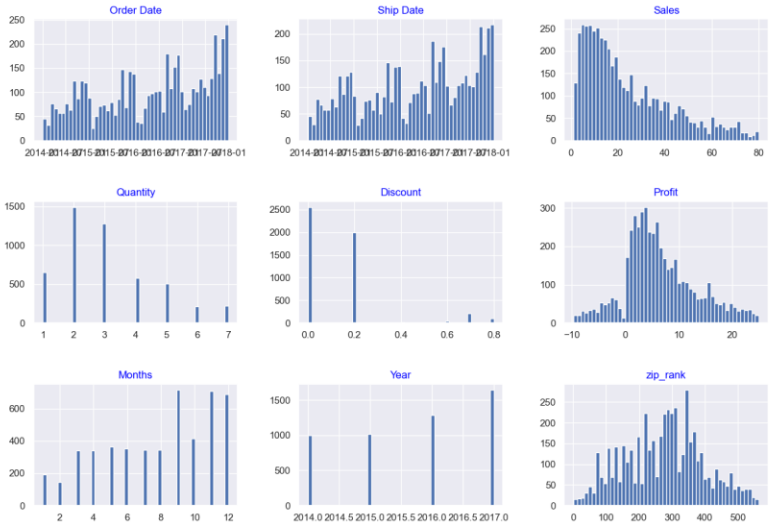
3. Data Processing

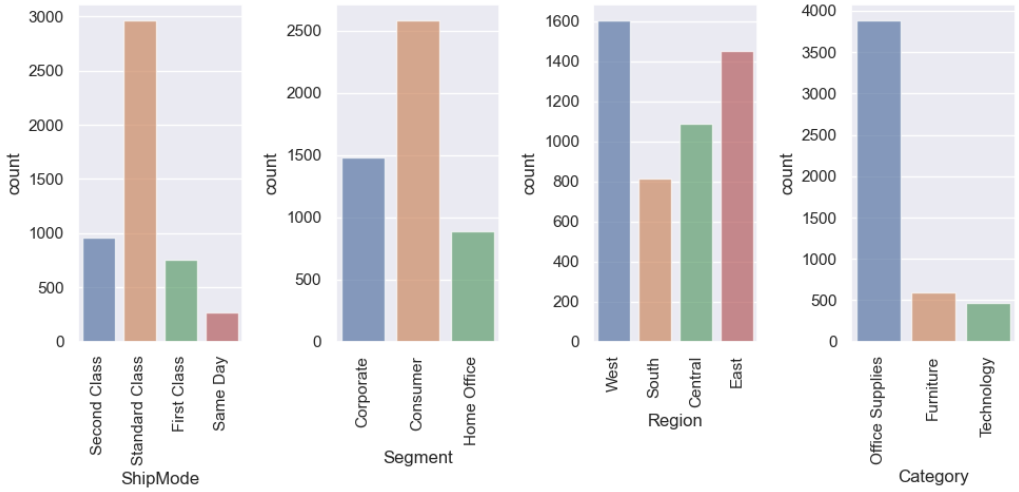
3.1 Remove outliers

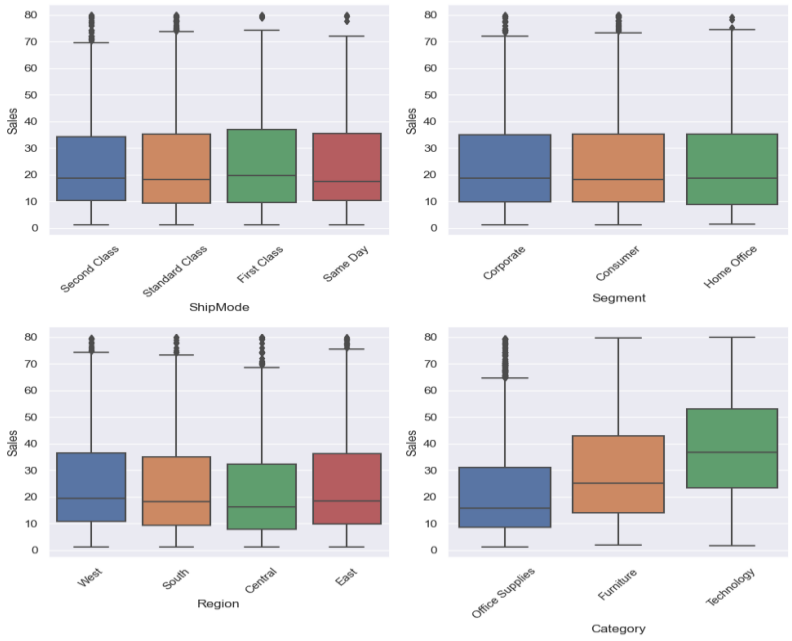




3.2 Data Distribution

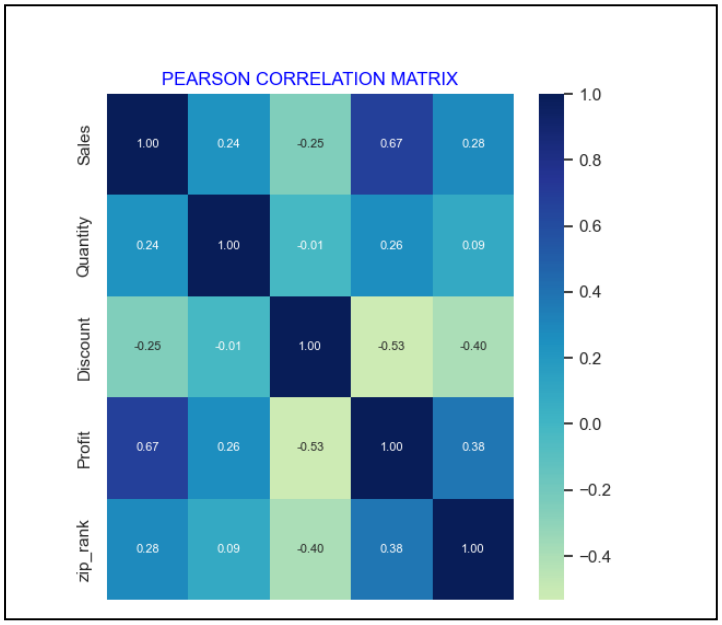

Order Date
Ship Date
Sales
Quantity
Discount
Profit
Months
Year
zip_rank


count
Second Class
Standard Class
First Class
Same Day
ShipMode
Corporate
Consumer
Home Office
Segment
West
South
Central
East
Region
Office Supplies
Furniture
Technology
Category

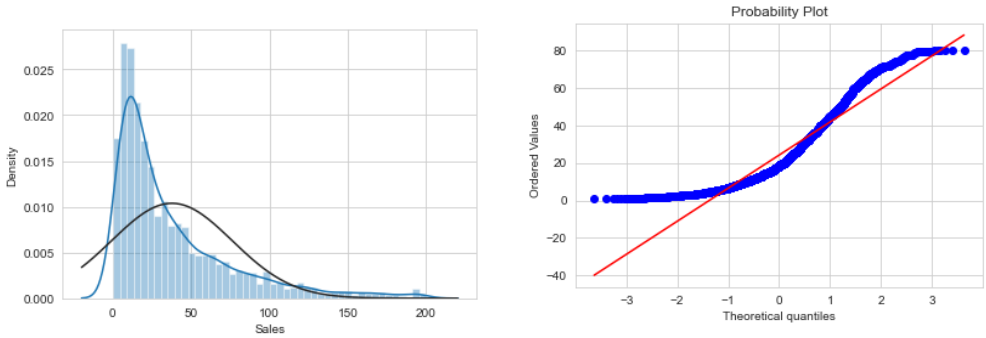


3.3 Check multicollinearity

I check for multicollinearity between the columns using a heatmap. It gives me a clear vision of which variables have correlation values higher than .75 and would be considered redundant in my dataset Then, I can drop the columns that would potentially cause my dataset to be unstable.



3.4 Target variable



I then plot a histogram and distribution for sales and look like we have a lot of low-value sales in our dataset which makes its distribution right-skewed.

3.5 Training /test dataset splitting

To fit the model on the data, I first ensured that all categorical features were converted to dummy variables or hot encoded. I then split the data into training (70%) and testing sets (30%). The model was then allowed to learn from the training set, make predictions, and report model performance on the test set.

4. Modeling

This step consists of building and tuning models using a python built-in library. In this case, I drop any insignificant variables that have p-values greater than 0.05.

4.1 Evaluation of the performance of the models

After comparing and testing different regressions with different features, the final results are:

Models	R2	MAE	RMSE
LinearRegression	63.26	7.53	10.78
Lasso	63.29	7.53	10.99
Ridge	62.84	7.48	11.01
RandomForestRegressor	82.96	4.02	7.28

- The results show that the best performances in terms of R squared (R^2) and Root Mean Square Error (RMSE) correspond to random forest Regressor.
- According to the result of R^2 and Root Mean Squared of these 4 models, the relationship between the features and the target variable is not clearly linear and shows the presence of non-linearity in the data.

4.2 Model selection

Based on the metrics of R2 and MAE, the random forest is selected to make a prediction of sales.

The reason I can think of is that the random forest is an ensemble method based on bagging of predictions of individual trees. It is more robust and not sensitive to small changes in the data.

5. The need for retraining

The model we built is learned from the data in the past. New data is generated and changing every day. If the statistics of new data don't change, the model is likely to remain stable and efficient. If the changes in data are related to the change in the distribution of new data, the correlation between the target and some features may introduce some new information that the model may not be able to handle correctly. Events like a new product introduced into the market or a new market campaign pushing the increase in sales, can change the distribution of the training data and affect the model that has been trained.

The retraining of the model may need to be scheduled once a month or once a quarter. This way the model is trained periodically, it will follow the change in the distribution of the features and the features that were not useful in the past may become useful in the new data. The general approach is to monitor the performance of the model. If the performance is acceptable with new data, there is no need to retrain. If the performance decreases, then it is time to schedule a retain, such as a drop in AUC_ROC of about 10%.

6. Summary and future work

Since the data was collected from 2014 - 2018, this model might not fully reflect all the price changes recently in the market. Moreover, features that were used in developing the models, might not be enough to sufficiently describe the sales. The sales range is mainly less than ~$100, which is a bit small in my opinion.

And lastly, the market of densely populated urban areas is definitely different from that of the rural area. This means more data needs to be collected from different regions across all 49 states in the US in order to increase the predictive power of models for the sales for any particular area.