Interactive Data Analysis

Black Friday Project Report

By: Albert Pierce

Introduction

This particular project was created based on a free competition hosted on a website. The goal of the project was to take the given dataset, which was a list of Black friday customer sales transactions from a retail company called “ABC Private Limited”. Using this dataset the eventual goal was to take the list of sales transactions and predict a customer’s purchase amount. To predict this amount, I implemented a General Linear Regression Model using the Scikit learn package provided in Python.

Finding the Right Data Source

The topic of Machine Learning has many different types of projects that one can do. Searching around the internet, I wanted to find a project that was not already completed because this would allow me to challenge myself with my machine learning skills. I came across a website “www.analyticsvidhya.com” that hosts paid and unpaid online competitions revolving around Data Science. For this project, I wanted to keep it in the scope of a student and go with the unpaid competitions. I was able to find and enlist myself in a Machine Learning competition and put my skills to the test. The overall challenge of this competition was to help a retail company “ABC Private Limited” predict the purchase price of various products purchased by customers based on historical historical purchase pattern. Using Jupyter as my Interactive Development Environment, I implemented a General Linear Regression Model in Python.

Black Friday DataSet

This dataset in particular contains a list of sales transactions captured at a retail store. The black friday dataset contains 550,069 observations and 12 attributes. This data set also only captures a month's worth of purchased products. The attributes of the black friday dataset are customer demographics such as age, gender, marital status, city_type, stay_in_current_city, product_details, product_id, product category, and total_purchase_amount. From here my next step was to consider which attributes are my predictor and response variables. Taking into account that my overall goal is to predict the total purchase amount of a customer i used the attribute total_purchase_amount as my response variable and the rest of the attributes in the dataset as my predictor variables. Displayed in the Figure below is the Black Friday Dataset with labeled attributes.

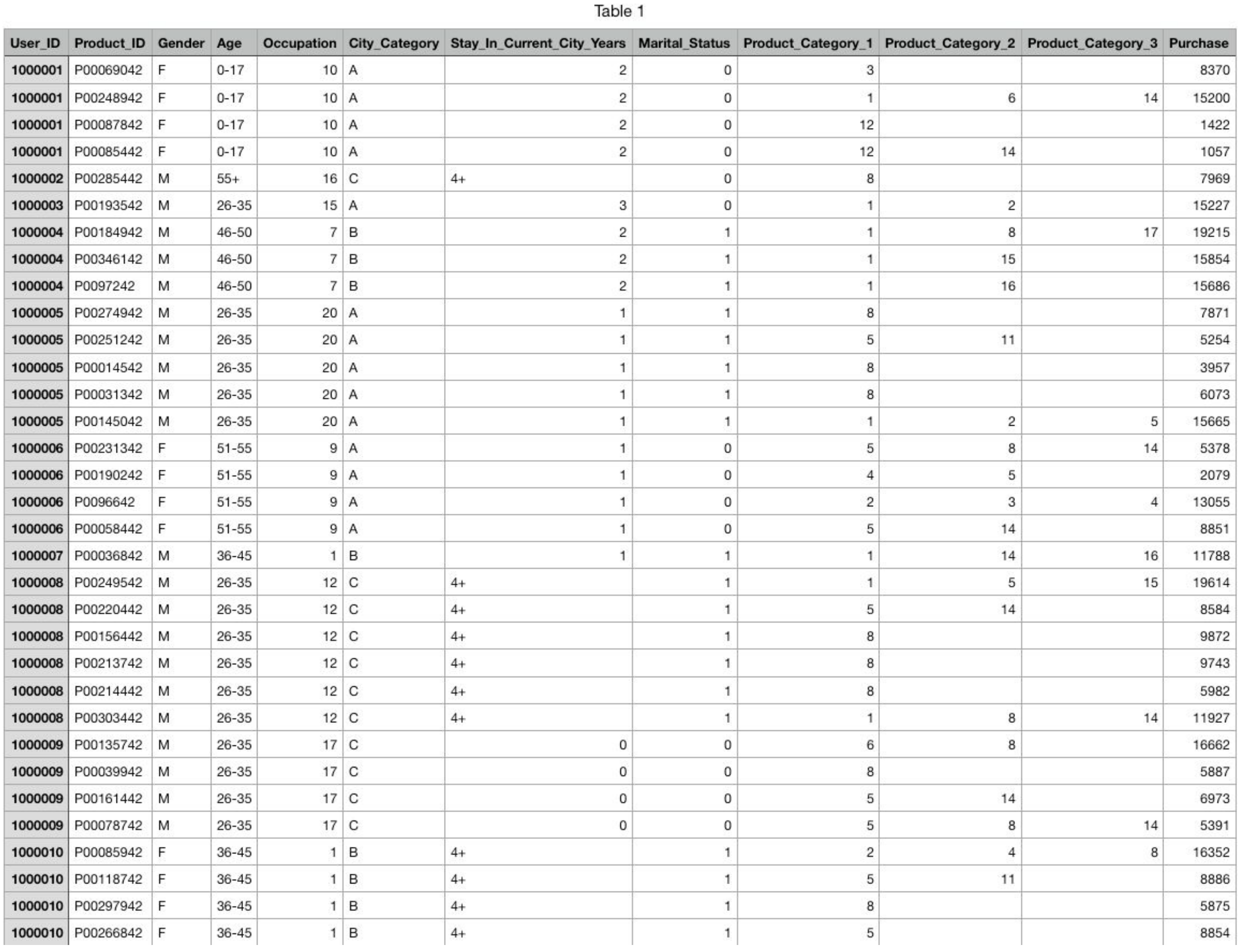


Table 1

User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
1000001	P00069042	F	0-17	10	A	2	0	3			8370
1000001	P00248942	F	0-17	10	A	2	0	1	6	14	15200
1000001	P00087842	F	0-17	10	A	2	0	12			1422
1000001	P00085442	F	0-17	10	A	2	0	12	14		1057
1000002	P00285442	M	55+	16	C	4+	0	8			7969
1000003	P00193542	M	26-35	15	A	3	0	1	2		15227
1000004	P00184942	M	46-50	7	B	2	1	1	8	17	19215
1000004	P00346142	M	46-50	7	B	2	1	1	15		15854
1000004	P0097242	M	46-50	7	B	2	1	1	16		15686
1000005	P00274942	M	26-35	20	A	1	1	8			7871
1000005	P00251242	M	26-35	20	A	1	1	5	11		5254
1000005	P00014542	M	26-35	20	A	1	1	8			3957
1000005	P00031342	M	26-35	20	A	1	1	8			6073
1000005	P00145042	M	26-35	20	A	1	1	1	2	5	15665
1000006	P00231342	F	51-55	9	A	1	0	5	8	14	5378
1000006	P00190242	F	51-55	9	A	1	0	4	5		2079
1000006	P0096642	F	51-55	9	A	1	0	2	3	4	13055
1000006	P00058442	F	51-55	9	A	1	0	5	14		8851
1000007	P00036842	M	36-45	1	B	1	1	1	14	16	11788
1000008	P00249542	M	26-35	12	C	4+	1	1	5	15	19614
1000008	P00220442	M	26-35	12	C	4+	1	5	14		8584
1000008	P00156442	M	26-35	12	C	4+	1	8			9872
1000008	P00213742	M	26-35	12	C	4+	1	8			9743
1000008	P00214442	M	26-35	12	C	4+	1	8			5982
1000008	P00303442	M	26-35	12	C	4+	1	1	8	14	11927
1000009	P00135742	M	26-35	17	C	0	0	6	8		16662
1000009	P00039942	M	26-35	17	C	0	0	8			5887
1000009	P00161442	M	26-35	17	C	0	0	5	14		6973
1000009	P00078742	M	26-35	17	C	0	0	5	8	14	5391
1000010	P00085942	F	36-45	1	B	4+	1	2	4	8	16352
1000010	P00118742	F	36-45	1	B	4+	1	5	11		8886
1000010	P00297942	F	36-45	1	B	4+	1	8			5875
1000010	P00266842	F	36-45	1	B	4+	1	5			8854

Looking at the dataset here we can already draw a couple of conclusions that would help fit the data into the Linear Regression Model. The attributes “Gender”, “Age”,”City_Category”,”Stay_In_Current_City”, and ”Product_Category(1-3)” are all categoricals with different amounts of levels. Taking these attributes into account, I knew that I needed to create dummy variables in order for Linear Regression to accept the data and make its predictions. These are also the attributes that are being inputted into the model. The response variable on the other hand, represented in dollars, is of type numeric and does not need to be factorized in the dataset and can be fed directly into the linear model.

Parsing the Data

After I figured out which attributes from the dataset will corresponds to predictor or response variables I used Python to load the contents of the ".csv" file into my program. Using Python libraries "numpy", "pandas" and "csv" I stored the csv contents into a dataframe with pandas and used that dataframe to create train and test datasets with the numpy library.

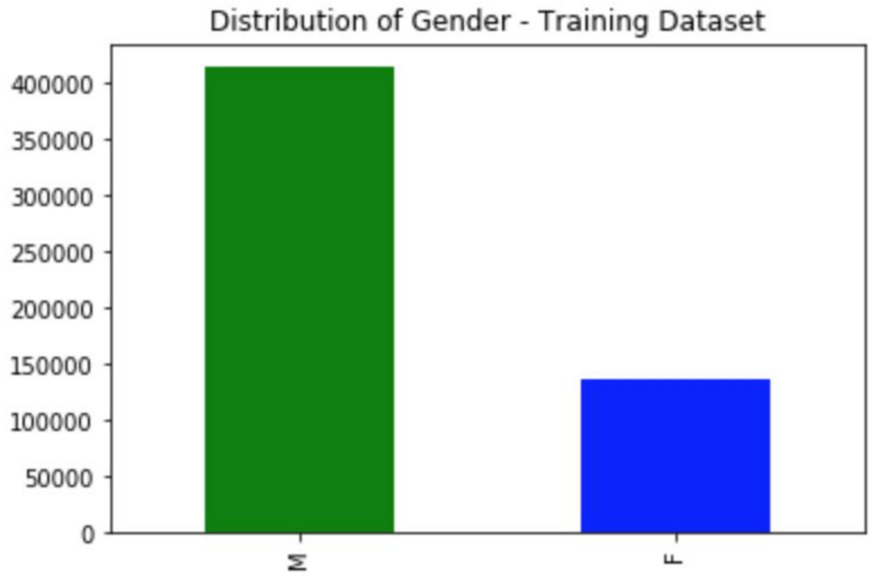
To not skew my Linear Regression Model's accuracy, I checked to see if any of the columns in the train and test datasets had null values. Many times, machine learning algorithms compute on null data and predict inaccurate results. To avoid for that we can replace the null values with a "0" to account that it is null value.

Many of the attributes in the database were also of type categorical. Linear Regression Models compute on floating point values and cannot compute on values that have a category weight to them. To account for this I changed the types of the attributes to as type 'category'. In doing this, I would create a new attribute column for every level in that specific attribute labeled with 1's and 0's that would satisfy my models constraints.

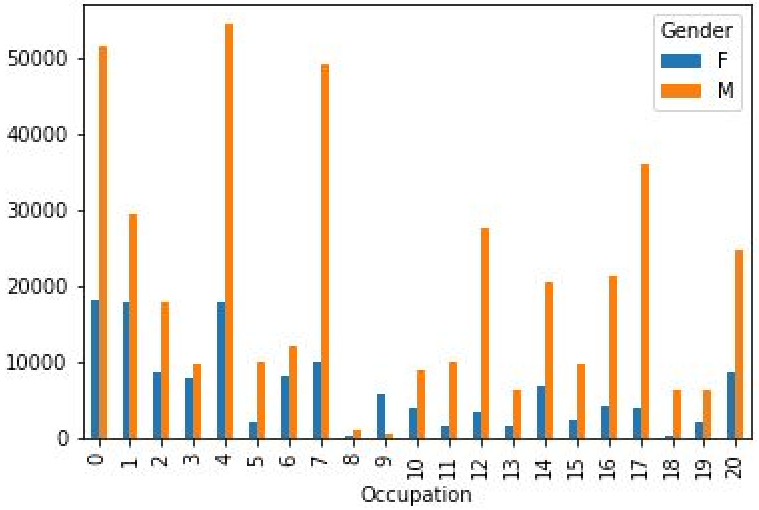
After parsing the data from the original ".csv" file, I created a single dataframe so that I can visualize the attributes of the Black Friday Dataset.

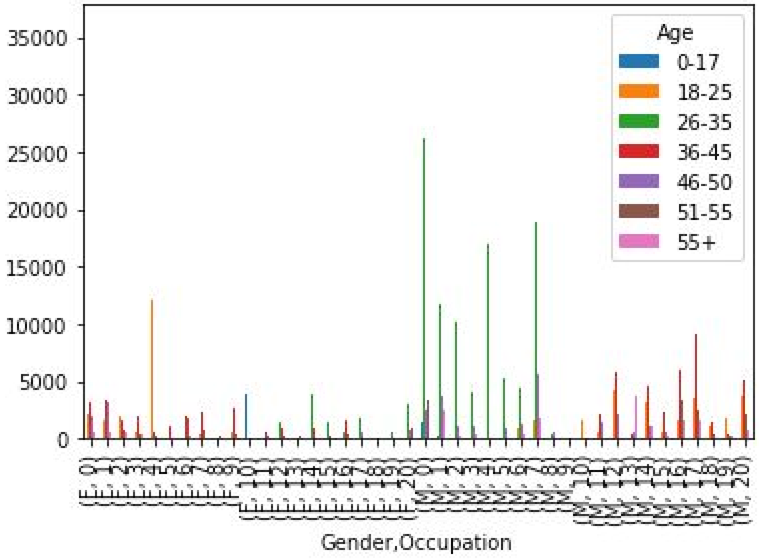
Visualizing the Data

To really analyze what kind of data I am dealing with, it is best to plot the data and recognize patterns throughout the set. One of the attributes that I wanted to see the distribution of is Gender. This would at least tell me who are my main customers within the dataset. The bar graph below displays the distribution of male to female within the dataset.



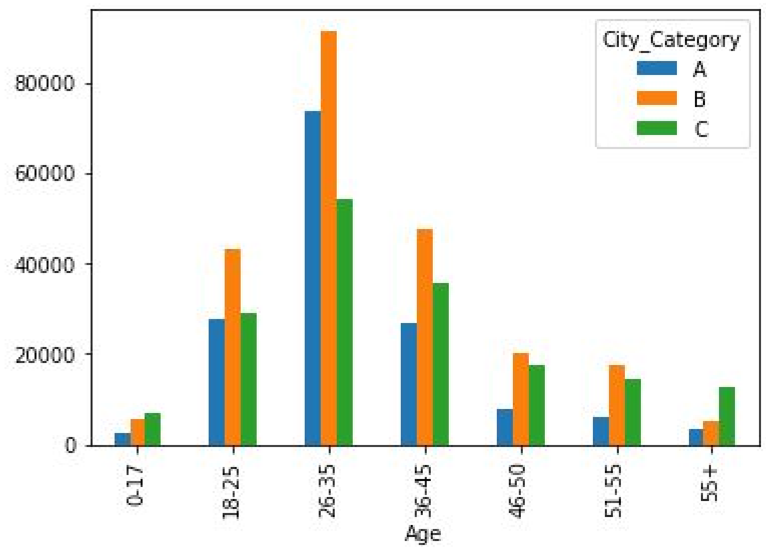
Here I was able to see that most of the customers that were buying these products were male and a significant portion of them. With that in mind, I was also trying to think of what type of products these might be that men predominately buy over women. Here I was interested in studying the male category because there were so many data points. I then wanted to plot more in depth plots that took into consideration multiple atributes. The figure below shows the distribution among Gender vs Occupation vs Age. This is very intersting to see because I can view a distribtuion of male and females in certain occupations and see roughly the age of those workers.



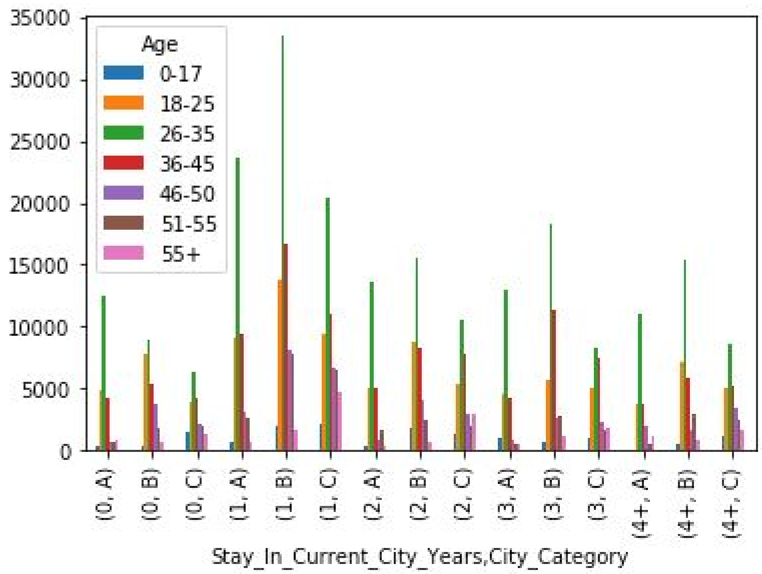


Here two bar plots are shown that show the distribution of Gender vs Occupation vs Age. Visualizing these plots with python these provide the ability to zoom in on the graph which helped subset the data. Looking at the second graph we can see that most of the males in occupation 0 were around the ages of 26-35. These are good attributes to visualize because I can now see by job who would buy more products. Depending on the person's occupation that can either give him freedom to purchase products or limitations with how many they can buy.

I then wanted to visualize if the age of a person and the city they lived in had any correlations. Using a bar graph I plotted the two attributes and saw that most of the people who were age 26-35 lived in city B. After plotting some of these attributes, I was able to see that most of the people throughout the dataset fall into the 26-35 category. Below shows a bar graph with Ave vs City and the highest peak of the graph is at 26-35 at City B.



I then moved on to see if there were any other correlations between the attributes of the dataset and saw that the stay in current city can also have a correlation with the city variable and the age variable. This is nice to visualize because now I would be able to see if older people were possibly paying for their dues so they don’t tend to spend more. Creating these types of relations before visualizing the data and seeing if the attributes actually follow that pattern can help see the useful information that is packed within a big data set. Below is a plot of the stay in city vs City vs Age. After looking at the plot, I was able to conclude that in city B mostly the young people that lived there were pretty recent. This might add to why that age group appears in a lot of these attribute correlations because if they are recent movers then they might buy more house related goods.



Linear Regression Model

After visualizing the data, I was able to correctly see trends throughout the dataset. This is always a good step to do before implementing a machine learning algorithm because we can kind of get a hint to as what the possible outcome might lean towards. Doing so I used the scikit learn library from Python to implement the General Linear Regression Model. Parsing the data was the preprocessing of the data so all I needed to do was create the instance and apply the model to the training and testing dataset. This is generally done by fitting the model the target values and the training data so it can accurately fit the data. Once the model was trained on the data the model was used to predict the truth values of the testing data set for which we can get can retrieve a score. This score is a value between 0 and 1 that tells me how well it was able to predict on the testing data. If the value is closer to 1 then it predicted with a higher accuracy and vice versa. Unfortunately, implementing a Linear Regression model on the Black Friday data set was not the best idea because the r-squared value that was returned was very low. The model returned a 12.5% accuracy which is less than the probability of just flipping a coin. I realized that linear regression is best used when the dataset is numeric so the attribute columns can actually help with the prediction factor. Most of the attributes that I dealt with were categorical values which is hard for linear regression to accurately fit to the data.

Overall, this dataset could have been improved before and provide more detailed data so Linear Regression could work properly. Knowing that, If I had the opportunity to expand on this dataset, i would expand to classify with different types of machine learning algorithms and compare their score values. This would then show me which algorithm is the best to use to predict the purchase amount of a customer.

Conclusion

Summarizing, linear regression was not the best machine learning algorithm to use to predict the purchase outcome of various product lines. Most machine learning algorithms can be generally classified to a wide range of datasets, but Linear regression does not do a good job of fitting this data set. Even though i got a 12.5% prediction, If i were to extend this project to incorporate multiple types of machine learning classifiers then this dataset could potentially be predicted with accurate results.