
SMARTBRIDGE
Let's Bridge the Gap

Machine Learning project report on

Prediction of Sales of a Super Store

Acknowledgement

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to Mr. Mohan for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.
I would like to express my gratitude towards my parents & member of SmartBridge for their kind co-operation and encouragement which help me in completion of this project.
I would like to express my special gratitude and thanks to industry persons for giving me such attention and time.
My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

TABLE OF CONTENTS

Abstract

Marketing data is one of the renowned data set when it comes to research of how public relation, location and other features effect the companies' turnover out of the given dataset. We are trying to predict the Profits that is dependent in nature. By involving necessary independent variables like location, etc., In this Project we are going to predict the nature of sales in a Super Store.

Introduction

The exchange of a commodity for money like clothes, furniture for money, the action of selling something is called Sales and data is the quantities, characters or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic etc., In other words simply we can say that things known or assumed as facts, making the basis of reasoning.

There are mainly two types of data

1**. Categorical data**: Data that is collected can be either categorical or numerical data. Numbers often don't make sense unless you assign meaning to those numbers. Categorical data helps you do that Categorical is when numbers are collected in groups or categories. It is further divided into two types

- Dichotomous: Dichotomous variables are nominal variables which have only two categories (Binary) or levels. For example, if we were looking at gender, we would most probably categorize somebody as either “male” or “female”.
- Non Dichotomous: Non Dichotomous are nominal variables which have more than two categories or levels. For example, if we were looking at grade in a semester, we would categorize many grades (A,B,C,D….).

2**. Numeric data**: Numeric data is a data that is measurable, such as time, height, weight, amount, and so on. You can help yourself identify numeric data by seeing if you can average or order the data in either ascending or descending order.

The Sales data we can assume the facts for making few computations and predictions, also reduce the relation between feature and parameters that leads to higher turnover of the Super Store.

Our data is consistent of both Categorical and Numerical. Here our aim is to drop all the unnecessary features and assign numerical equivalence to categorical features and to perform linear regression to predict the Profits of the company.

Optimizing Data

Independent variables:

An independent variable is a variable that is manipulated to determine the value of a dependent variable s. The dependent variable is what is being measured in an experiment or evaluated in a mathematical equation and the independent variables are the inputs to that measurement.

Dependent variables:

A dependent variable is what you measure in the experiment and what is affected during the experiment. The dependent variable responds to the independent variable. It is called dependent because it "depends" on the independent variable.

From the given dataset we observed that few features doesn't affect our independent variable.

Such features in "16Super_Stores.csv" were ROW_ID, Order ID, order date, shipping date, ... Are dropped using drop command.

REPLACING all the categorical data with equivalent numerical data.

In the table attributes of Shipping mode and Segment were categorical in nature, we replaced them with 'replace' command.

In Shipping mode we replaced

- Second Class with 0
- Standard Class with 1
- First Class with 2
- Same Day with 3

And in Segment feature we replaced

- Consumer with 0
- Cooperate with 1
- Home Office with 2

Shape

The **shape** attribute for numpy arrays returns the dimensions of the array.

in shape we were having observations and columns

observations: these are having features

columns: these known to be a set of data

Command : df.shape

We got (2121,21)

Here 2121 are the observations in the dataset and 9 is columns in the data set

SIZE OF THE DATA SET:

Features:

These are the input values of given data set

Row ID: Database server identifies each data row in the table with a unique int ernal row id. Row id identifies the location of row.

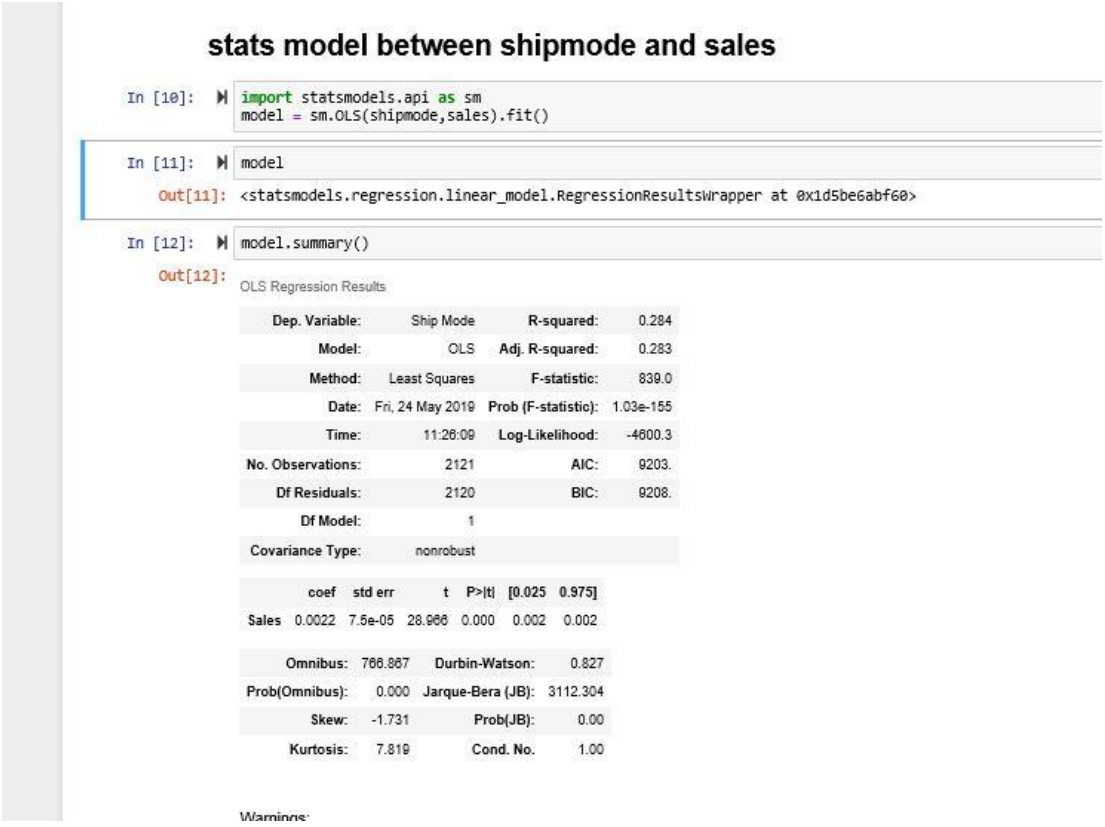
- Order ID: Order id is a number system, used to identify your transaction.
- Order Date: A customers last order date is the data on which their last order wi th your store was placed.
- Ship Date: The date on which goods are sent out to a customer.
- Ship Mode: The shipping mode is a way of shipping goods
  - Same day:
  - Frist class:
  - Second class:
  - Standard class:
- Customer ID: A customer id is assigned to each of our accounts for purposes of order Processing, tracking and customer account militance
- Customer Name: A customer is an individual or business that purchases anothe r Company's goods or services
- Segment: each of the parts into which something is divided

- Country:
- 'City'
- 'State'
- 'Postal Code'
- 'Region',
- 'Product ID'
- 'Category'
- 'Sub-Category'
- 'Product Name'
- 'Sales'
- 'Quantity'
- 'Discount'
- 'Profit'

Exploratory Data Analysis

EDA for Ship Mode VS Sales

From the boxplot of the given data we can say that neither there are many outliers in the data so it decreases the Accuracy. From the Distribution graph its neither evenly distributed nor the graph resembles a bell.



stats model between shipmode and sales

```
In [10]: import statsmodels.api as sm
         model = sm.OLS(shipmode,sales).fit()

In [11]: model
Out[11]: <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x1d5be6abf60>

In [12]: model.summary()
```

Out[12]:

OLS Regression Results

Dep. Variable:	Ship Mode	R-squared:	0.284
Model:	OLS	Adj. R-squared:	0.283
Method:	Least Squares	F-statistic:	839.0
Date:	Fri, 24 May 2019	Prob (F-statistic):	1.03e-155
Time:	11:26:09	Log-Likelihood:	-4600.3
No. Observations:	2121	AIC:	9203.
Df Residuals:	2120	BIC:	9208.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Sales	0.0022	7.5e-05	28.966	0.000	0.002	0.002

Omnibus:	766.867	Durbin-Watson:	0.827
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3112.304
Skew:	-1.731	Prob(JB):	0.00
Kurtosis:	7.819	Cond. No.	1.00

Warnings:

EDA for Product ID and Sales

Comparing the product ID and Sales we can get to know that there are about 364 unique product Id's and the Product ID 10001095 has the number of repetitions. It has appeared 21 times in the data set.

The mean of the profit set is 8.699327 and the profit for about 32 products is 0.00.

EDA for PC and profit

From the scatter plot between profit and postal code we can say that all the profit and loses counts are almost similar. From the boxplot of postal code there are no outliers.

postalcode and profit

In [28]:
```
P =df['Postal Code']
q = df['Profit']
m3 = sm.OLS(q,P).fit()
m3.summary()
```

Out[28]:

OLS Regression Results

Dep. Variable:	Profit	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	3.760
Date:	Fri, 24 May 2019	Prob (F-statistic):	0.0526
Time:	04:05:35	Log-Likelihood:	-13432.
No. Observations:	2121	AIC:	2.687e+04
Df Residuals:	2120	BIC:	2.687e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Postal Code	8.907e-05	4.59e-05	1.939	0.053	-1.01e-06	0.000

Omnibus:	1385.765	Durbin-Watson:	1.941
Prob(Omnibus):	0.000	Jarque-Bera (JB):	120428.094
Skew:	-2.258	Prob(JB):	0.00
Kurtosis:	39.638	Cond. No.	1.00

EDA for CID and profit

From the customer id and the regular customers who visited the store contributed randomly to the profits as well as loses. So we can say that no fixed customer to generate more profit even if the sales are more.

Customer ID and profit

In [30]:
```
P =df['Customer ID']
q = df['Profit']
m3 = sm.OLS(q,P).fit()
m3.summary()
```

Out[30]:

OLS Regression Results

Dep. Variable:	Profit	R-squared:	0.004
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	7.481
Date:	Fri, 24 May 2019	Prob (F-statistic):	0.00629
Time:	04:07:55	Log-Likelihood:	-13430.
No. Observations:	2121	AIC:	2.686e+04
Df Residuals:	2120	BIC:	2.687e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Customer ID	0.0005	0.000	2.735	0.006	0.000	0.001

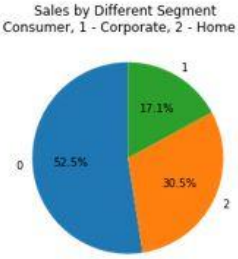
Omnibus:	1398.210	Durbin-Watson:	1.944
Prob(Omnibus):	0.000	Jarque-Bera (JB):	122503.246
Skew:	-2.288	Prob(JB):	0.00
Kurtosis:	39.949	Cond. No.	1.00

EDA for Segment and Sales

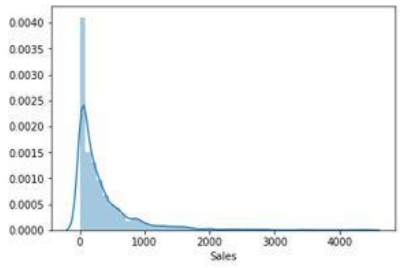
More number of sales were driven by the segments which are having the observations “Consumers” and followed by “Corporates” and finally by “Home Office” contributing 52.5%,17.1%,30.5% respectively. Even though the sales distribution curve resembles the bell curve the distribution curve of segment does not resemble bell curve even after normalization.

	Sales	Segment
Sales	1.000000	0.001425
Segment	0.001425	1.000000

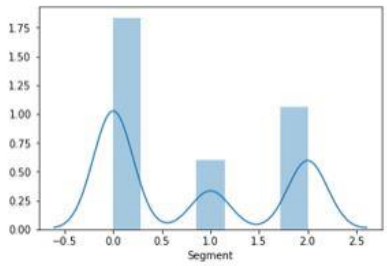
```
In [10]: plt.pie(seg.value_counts(), labels=seg.unique(), startangle=90, autopct='%.1f%%')
         plt.title('Sales by Different Segment \n 0 - Consumer, 1 - Corporate, 2 - Home Office')
         plt.show()
```




```
In [11]: sns.distplot(sales)
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x21b26fa24a8>
```

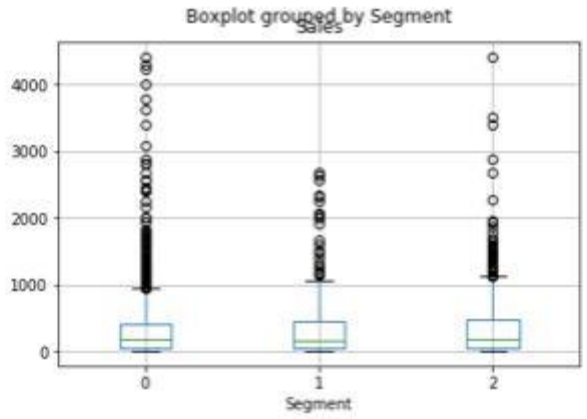



```
In [12]: sns.distplot(seg)
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x21b270989b0>
```



Segment

```
In [13]: data_SS.boxplot(column = 'Sales' ,by='Segment')
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x21b24ef70b8>
```



EDA for Model1

Reason for choosing this model is because this model shows accuracy of 24.03% low RMSE value compared to other train split models and high R2 values. From this scatter plot between the test value and predicted value. We can say that it is almost linear but not perfectly accurate because all the values are not uniquely mapped. By drawing the mean curve we can say that both profit and loses are almost numerically equal.

Accuracy = 26

RMSE = 117

R2 = 0.2651

EDA for Model2

Reason for choosing this model is because this model shows accuracy of 24.03% low RMSE value compared to other tran_split models and high R2 values. From this scatter plot between the test value and predicted value. We can say that it is almost linear but not perfectly accurate because all the values are not uniquely mapped.

Accuracy = 0.24

RMSE = 116.652

R2 = 0.208

From the model 1 and model2 passing on same inputs which is already provided in the given data set.

Model1 predicted = -111 as profits.

Model2 predicted = 83632 as profits which is an outlier value.

Which is nearer to the actual value 43333. Hence making model2 much more feasible.

Conclusion

From the above inferences we conclude that most of the values are in negative i.e the company's profits are negative which means the company is running in losses due to orders from specific postal codes and type of shipping modes as well as the discounts offered on the products are high.

The super store should stop providing high discounts at specific postal codes and also in should increase the charges of the shipping modes.

If the company keeps performing like this then very soon the company would be shut down.

Bibliography

- www.google.co.in
- www.quora.com