Understanding Magazine Subscription Behaviour

Type: Marketing dataset

ALY6020 Predictive Analytics,
Fall 2023

Professor
Justin Grosz

Submitted by
Dhwani Patel

19th November 2023
College of Professional Studies
Northeastern University

I. Introduction

With the help of binary classification, the business problem will be addressed considering relevant variables. Both, SVM and logistic regression will be used to analyse precision rate, recall rate and accuracy. The dataset contains 'Marketing Campaign' data with 2240 entries. The provided data contains 29 variables including Birth year, Marital status, Children details, date of enrolment, amount spent on FMCG products, number of items bought and customer feedback.

Business Problem

A magazine company is trying to understand the mismatch in the sales expectations. With more people staying home, the company had expected a high increase in the sales of magazine. However, a decline in sales has been seen and wants to understand the causes.

The data analysis aims in knowing the spendings on various products of each customer with amount & quantity spend. Skipping through the data, it can also be seen that the data includes details of pas two years of each customer. Furthermore, details like Marital status, number of kids, birth year, and income are the major factors affecting the decisions of customers. Which will have to be correlated with the amount spending and the frequency of the purchase. Two of the factors indicting concern are the customer had reported or filed any complaints. Through which the suggestions and recommendations will be designed. Before suggesting it is important to know the relationship between the independent and dependent variables.

With the help of confusion matrix, the variable dynamics will be analysed for performance measurement. These type of models are useful in knowing the risk associated with the business. If any specific value or set of goals needs to be reworked for better results. In order to check the model accuracy, the confusion matrix will be build which will guide the decision- making process or possible approached to be tried by the company.

Here, in the mentioned business problem 'Response' is the dependent variable which will guide in observing the conversion rate of subscription. For the analysis, specific libraries in python will be used i.e. Numpy, Pandas and Sklearn on Jupyter Notebook IDE. And for the EDA and visualizations matplotlib will be utilized.

II. Data Cleansing

After importing the libraries and data, as a first step to keep the data clean I carefully reviewed the first 10 rows using '.head(10)'. As per the instructions provided, there were only 22 columns were to be used out of total 29 attributes.
To address the actions, I decided to go with the option of dropping the unwanted attributes. Initially, wanted to create a database with the necessary columns however decided to remove the unnecessary ones. Since, the number of unwanted columns were lesser than the necessary.

Checked the list of columns and after converting the data into data frame, dropped 7 columns which will not be considered for the analysis. The next step is to check the data types of each variable. After checking the data types of the data frame, it was a turn to check if the data has missing values or not.
Using '.isnull().mean()', the missing values were check where except for the income variable the data does not have missing values. Income has about ~1.07% of missing values which is not much comparing to the date.

Since ‘Income of an individual’ seems to be an important variable for the analysis, I am not thinking of dropping the column. I will go with the process of Imputing the data. Also because of the less quantity of missing value, replacing values would not have a huge impact. Hence, we don’t really have to lose any data by keeping values intact. As the datatype for the column is ‘float’ which is numeric value, it will not have a trouble replacing the values.

If in case of the selection of dropping the values, it will be tough to mention if the salary of people does have a real-time impact on their purchasing habit. Replaces the ~22 null values with the mean value as ‘52247.25’ which is the average of present income range. To check is there are any values mis-spelled or if the data has any other anomalies, checked unique values in each column.

III. EDA

For the exploratory data analysis, I used multiple methods of exploring the data along with the visualizations. Beginning from loading the data with useful libraries like Numpy, Pandas, sklearn, seaborn and matplotlib. I moved to understanding the data by reading through the variable descriptions and unique values with the help of .info, describe and head.

As described into the data cleansing, I moved towards cleaning the data keeping the necessary columns. The data types of some columns like Education, Marital status and Date of customer's enrolment with the company needed to be addressed. As the data type was object which later was converted to integer. Using pandas, changes the type to_ datetime variable to make it more precise and use three different variables as 'Day', 'Month' and 'Year'.

By using value mapping, I converted the values of Education and Marital status into numbers for the ease of analysis. After manually, reviewing the data I moved to creating visualization that would help having better understanding of the data.

In Appendix 1, 2 and 3 the visualization has been shared. Where the relation between various variables have been represented using graphs like histogram, scatterplot and barplot. The relation can really be seen between the parameters however, I believe the information of total expense would have added more value. There are expenses and spendings provided but the motto is to categorize them within the segment. For example Appendix 3 has four variables talking about number of purchases through various modes.

Where it is also visible that the monthly web visits have gone down. And in number of purchases made using catalog seems to be higher.

I. Data Analysis

The data analysis includes both Logistic Regression and Support Vector Machine as data modelling. The model accuracy can be seen in appendix 4 & 5 for both the models. Consumer's subscription behaviour depends on multiple factors which can be predicted based on the current customer preferences. 'Response' from customer being the dependent variable, we can see the logistic regression has given the result as ~86% accuracy. The SVM model has given accuracy of ~85%.

For the analysis I have picked all the factors except for the date of enrolment with the company since it was troughing value error. With the details like years of association and complain , more could be understood if over the years something has caused a decline.

The confusion matrix states, the majority events predicted are true with 348 count. Out of the 20% test data, 73 are false Negative which does not seem to be good as it indicates that the event has occurred but it shows false. On a contrary note, it is good to see the false positive to be low as 15. The number of true negative also is very low where it states that more precision is needed to find the real false.

Looking at the linear regression model (Appendix 6), it can be seen that the variable like Birth year, food purchases, luxury purchase, catalog purchases and store purchase have been major cause of money spending. Also, since the older generation would be more comfortable in store purchases also after pandemic (assuming is it recent) would cause people to go out. Which could possibly cause decline in online purchase.

II. Conclusion

The model can be further trained to identify true negative values in the prediction. Over all, since the model seems to be capturing true positive which shows a strong impact of certain independent variables having strong correlation in notifying that there has been a decline stated. Looking at the results, the customers have complained less which means nothing false has been reported. However, The major customer base is ranging from year 1960 to 2000. Which also states some redundancy due to the technological support. The density of store purchases also seem to be higher than virtual transactions. The type of the customer seems to be causing the decline into the sale it seems. Logistic regression seems to be providing better accuracy however it is more or less the same. Maybe after having hyperparameter, the accuracy can be improved. The linear regression model also provides with variables having higher impact on 'response'.

Recommendation

The company should focus more on the type of customers which are technologically advance to increase the sales. It is a good thing that the complains received from customers are less.

Responding to the assumption about people's reading preferences based on the variables provided for the analysis, it cannot really be states that people have been reading more or less. However, the activity based purchase for the FMCG products especially food seems. Not all the factors have a direct impact on customer response however, there are significant factors affecting the decline which could be other that the customer spending money one food or extreme luxury shopping. It the customer base that needs attentions and maybe targeting youth would help that.

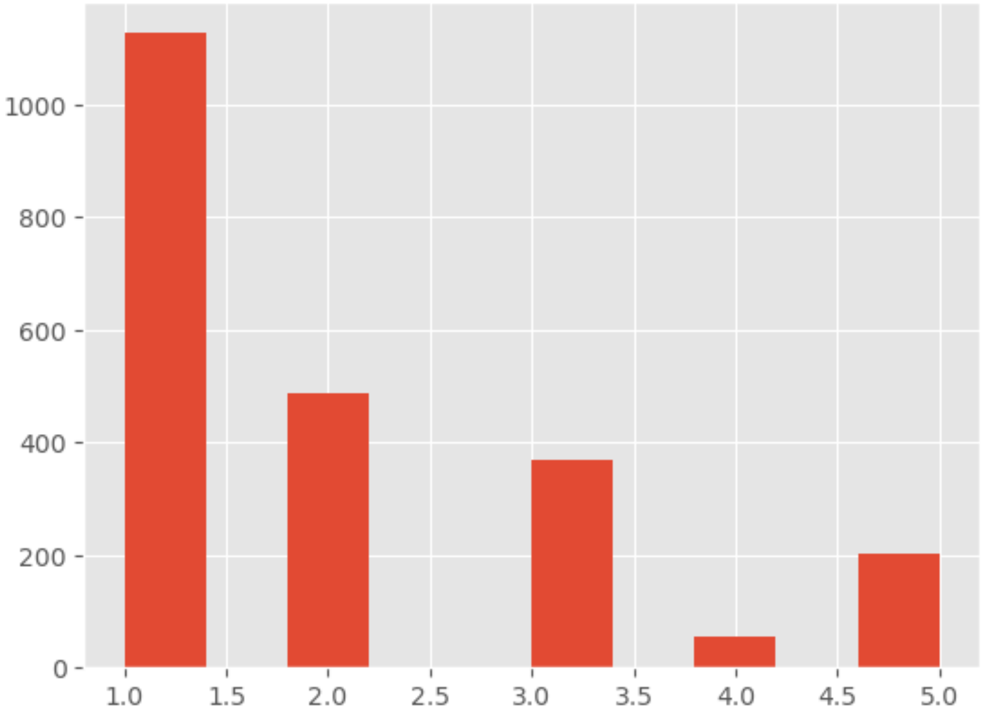
III. References

pandas.DataFrame.boxplot — pandas 2.1.3 documentation. (n.d.). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.boxplot.html

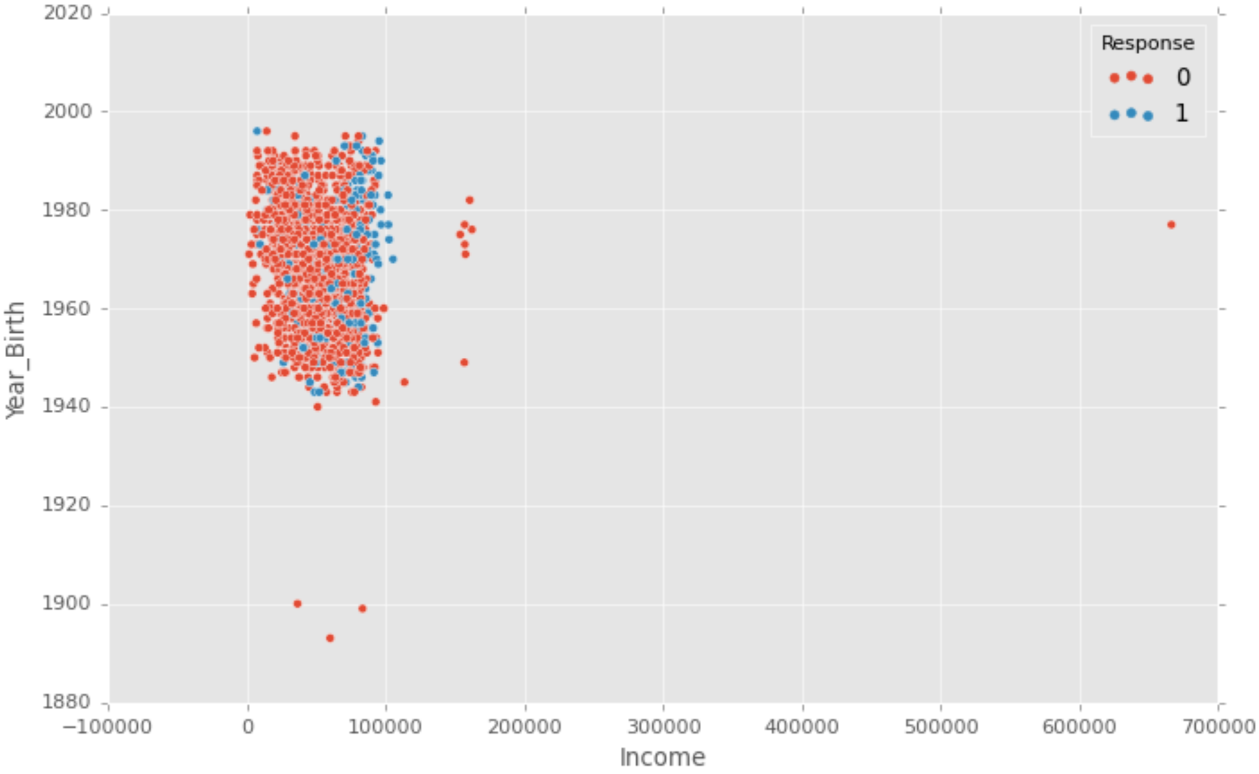
Kang, C. (2020, July 5). Applying logistic regression and SVM. *Chan`s Jupyter*. https://goodboychan.github.io/python/datacamp/machine_learning/2020/07/05/01-Applying-logistic-regression-and-SVM.html

IV. Appendix

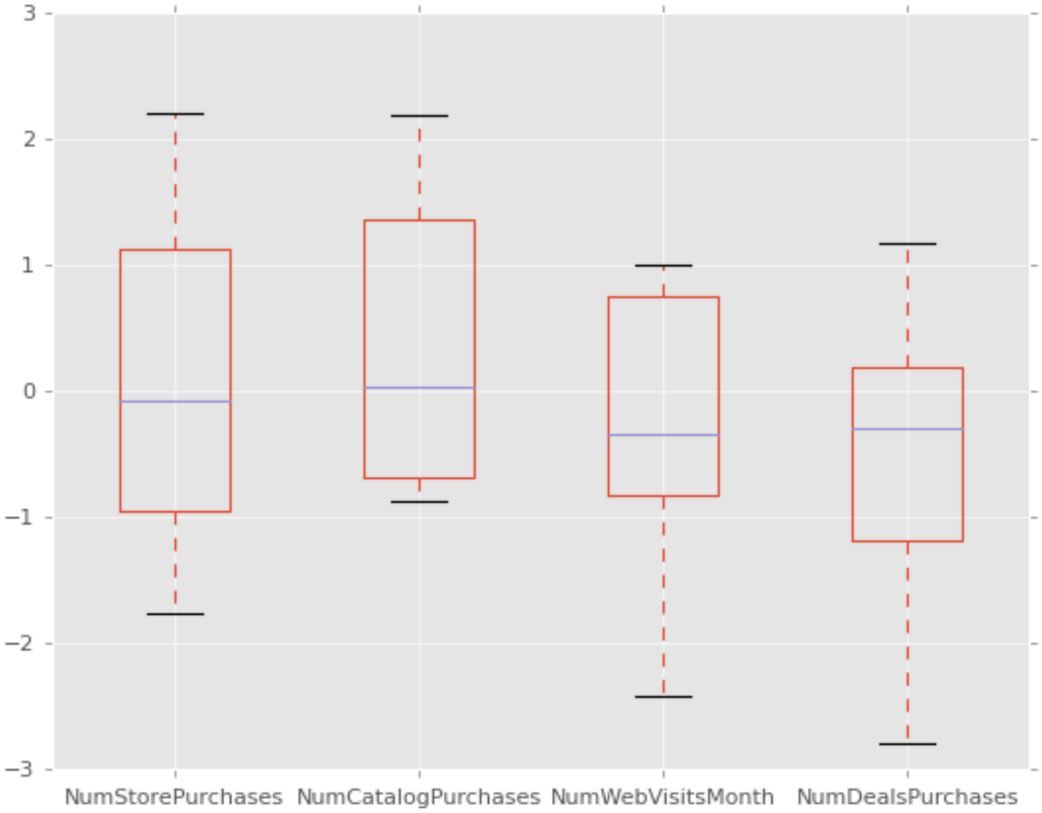
Appendix 1



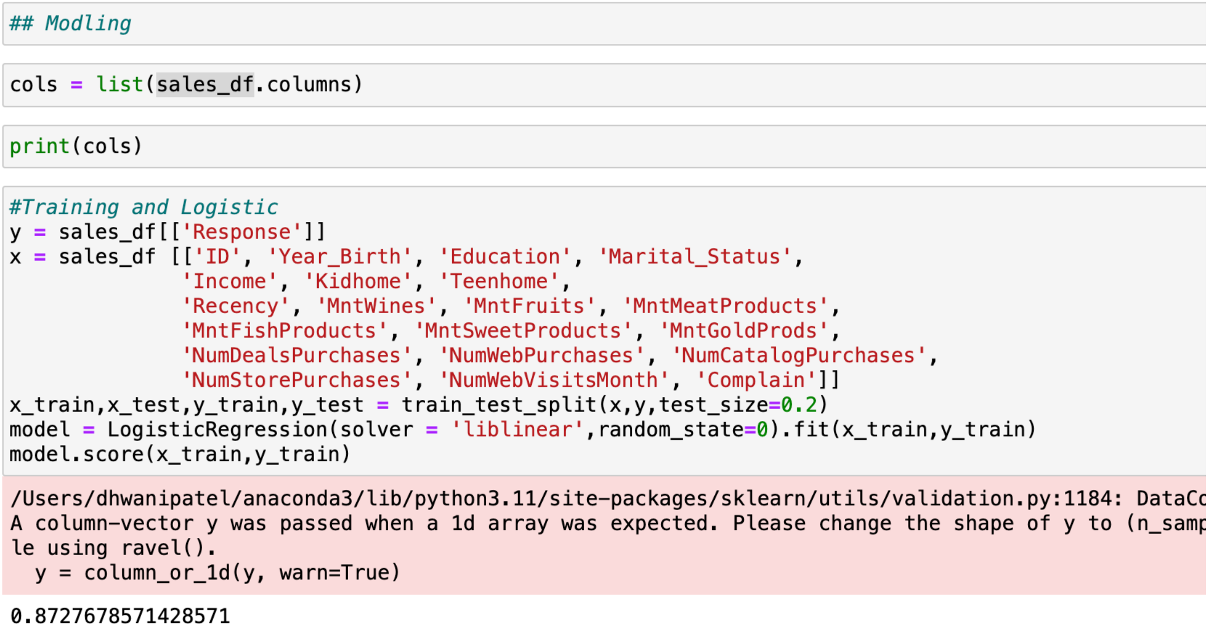
Appendix 2



Appendix 3



Appendix 4



```
## Modling
```

```
cols = list(sales_df.columns)
```

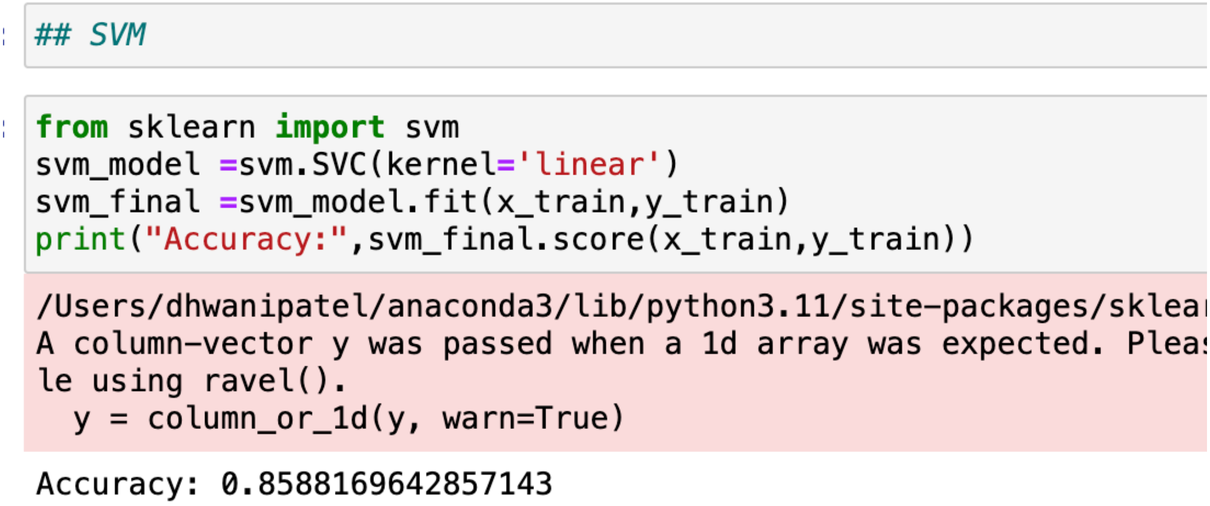
```
print(cols)
```

```
#Training and Logistic
y = sales_df[['Response']]
x = sales_df [['ID', 'Year_Birth', 'Education', 'Marital_Status',
              'Income', 'Kidhome', 'Teenhome',
              'Recency', 'MntWines', 'MntFruits', 'MntMeatProducts',
              'MntFishProducts', 'MntSweetProducts', 'MntGoldProds',
              'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases',
              'NumStorePurchases', 'NumWebVisitsMonth', 'Complain']]
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2)
model = LogisticRegression(solver = 'liblinear',random_state=0).fit(x_train,y_train)
model.score(x_train,y_train)
```

```
/Users/dhwanipatel/anaconda3/lib/python3.11/site-packages/sklearn/utils/validation.py:1184: DataCo
A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samp
le using ravel().
  y = column_or_1d(y, warn=True)
```

```
0.8727678571428571
```

Appendix 5



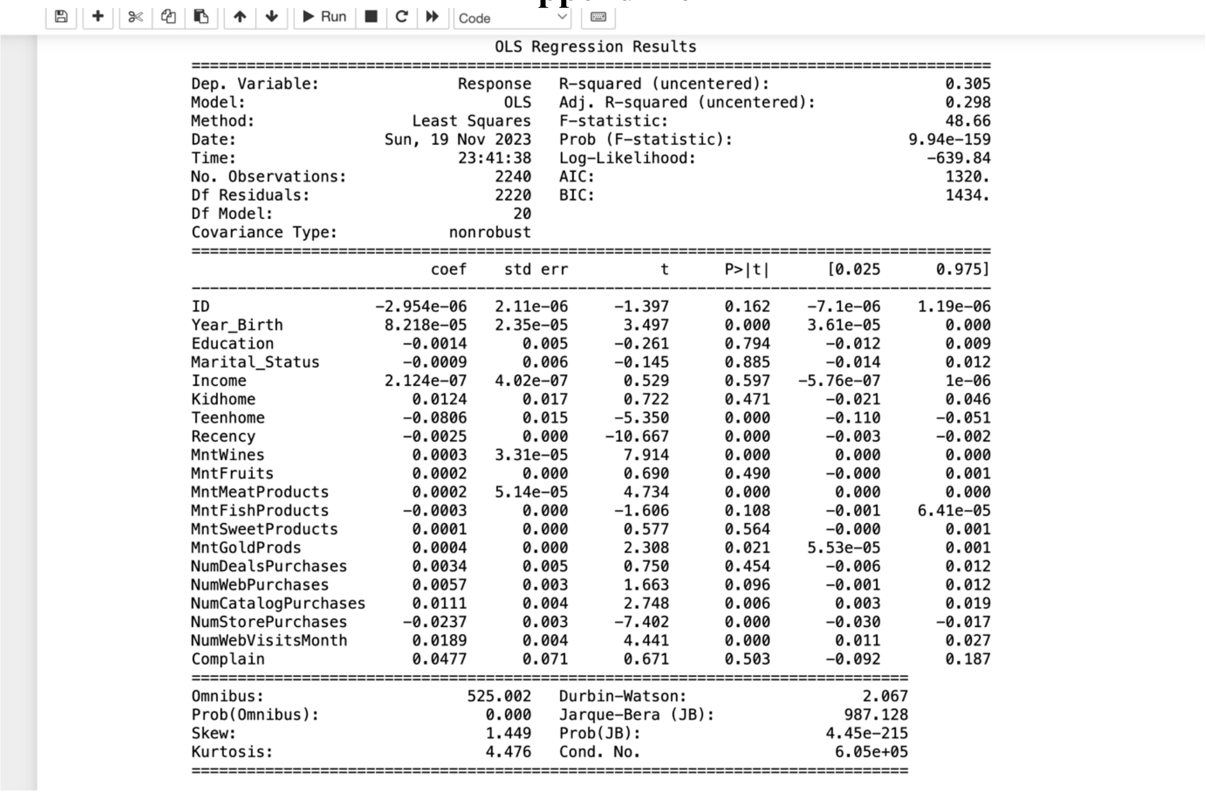
```
## SVM
```

```python
from sklearn import svm
svm_model =svm.SVC(kernel='linear')
svm_final =svm_model.fit(x_train,y_train)
print("Accuracy:",svm_final.score(x_train,y_train))
```

```
/Users/dhwanipatel/anaconda3/lib/python3.11/site-packages/sklea
A column-vector y was passed when a 1d array was expected. Pleas
le using ravel().
  y = column_or_1d(y, warn=True)

Accuracy: 0.8588169642857143
```

Appendix 6



```
                            OLS Regression Results
=======================================================================================
Dep. Variable:               Response   R-squared (uncentered):                   0.305
Model:                            OLS   Adj. R-squared (uncentered):              0.298
Method:                 Least Squares   F-statistic:                              48.66
Date:                Sun, 19 Nov 2023   Prob (F-statistic):                   9.94e-159
Time:                        23:41:38   Log-Likelihood:                         -639.84
No. Observations:                2240   AIC:                                      1320.
Df Residuals:                    2220   BIC:                                      1434.
Df Model:                          20
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
ID                  -2.954e-06   2.11e-06     -1.397      0.162     -7.1e-06    1.19e-06
Year_Birth           8.218e-05   2.35e-05      3.497      0.000     3.61e-05       0.000
Education              -0.0014      0.005     -0.261      0.794       -0.012       0.009
Marital_Status         -0.0009      0.006     -0.145      0.885       -0.014       0.012
Income               2.124e-07   4.02e-07      0.529      0.597    -5.76e-07       1e-06
Kidhome                 0.0124      0.017      0.722      0.471       -0.021       0.046
Teenhome               -0.0806      0.015     -5.350      0.000       -0.110      -0.051
Recency                -0.0025      0.000    -10.667      0.000       -0.003      -0.002
MntWines                0.0003   3.31e-05      7.914      0.000        0.000       0.000
MntFruits               0.0002      0.000      0.690      0.490       -0.000       0.001
MntMeatProducts         0.0002   5.14e-05      4.734      0.000        0.000       0.000
MntFishProducts        -0.0003      0.000     -1.606      0.108       -0.001    6.41e-05
MntSweetProducts        0.0001      0.000      0.577      0.564       -0.000       0.001
MntGoldProds            0.0004      0.000      2.308      0.021     5.53e-05       0.001
NumDealsPurchases       0.0034      0.005      0.750      0.454       -0.006       0.012
NumWebPurchases         0.0057      0.003      1.663      0.096       -0.001       0.012
NumCatalogPurchases     0.0111      0.004      2.748      0.006        0.003       0.019
NumStorePurchases      -0.0237      0.003     -7.402      0.000       -0.030      -0.017
NumWebVisitsMonth       0.0189      0.004      4.441      0.000        0.011       0.027
Complain                0.0477      0.071      0.671      0.503       -0.092       0.187
==============================================================================
Omnibus:                      525.002   Durbin-Watson:                   2.067
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              987.128
Skew:                           1.449   Prob(JB):                    4.45e-215
Kurtosis:                       4.476   Cond. No.                     6.05e+05
==============================================================================
```