What is Exploratory Data Analysis

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

Exploratory Data Analysis was coined by John W. Tukey, who wrote the book Exploratory Data Analysis in 1977. It has been widely known and accepted that it is an essential part of any data science project. Methodologies such as the CRISP-DM, SEMMA , JTA, and others highlight the necessity of EDA in any data science project. This paper offers a brief description of the process of data exploration and its role in a data science project. It will present a use case example of a data mining exercise that aims to evaluate the efficacy of independent variables in the designated model.

Keywords: CRISP-DM,SEMMA,JTA,EDA

What is Exploratory Data Analysis

Introduction

In the discussion of data science, the terms Machine Learning, and Data Mining are used interchangeably and often in the same sentence. The words probability & statistics, algorithms, and various algorithm techniques are also widely used in data science discussions. The bottom line for any data science project is that it should produce and answer to a question. The answer should add value to the questioner who posed the question in the first place. The questions are typically asked in a meeting room filled with executives, managers and directors. These people monitor the heartbeat of their organization by reviewing data and coming up with new ways of collecting data. Data collection and data analysis are heart and sole of a data science project. It is mostly these two activities that inspire questions such as how we can improve customer retention at our bank. Sometimes questions are asked without having looked at data. These questions are based on anecdotal evidence-a hearsay. An example given by Allen Downey is that first babies tend to arrive late, or my favorite one down here in Texas is that Sage bushes blossom before rainy days.

With the advent of the computer technology combined with a rich history of mathematical advancements, we now live in an era that we can create and utilize massive amount of data that we can pose and answer questions related to improving anything in any domain as long as we collect the correct data, and analyze it in a way that yield accurate results. This makes the role of data analysis and more specifically exploratory data analysis of paramount importance in the field of data science-at least for the time being.

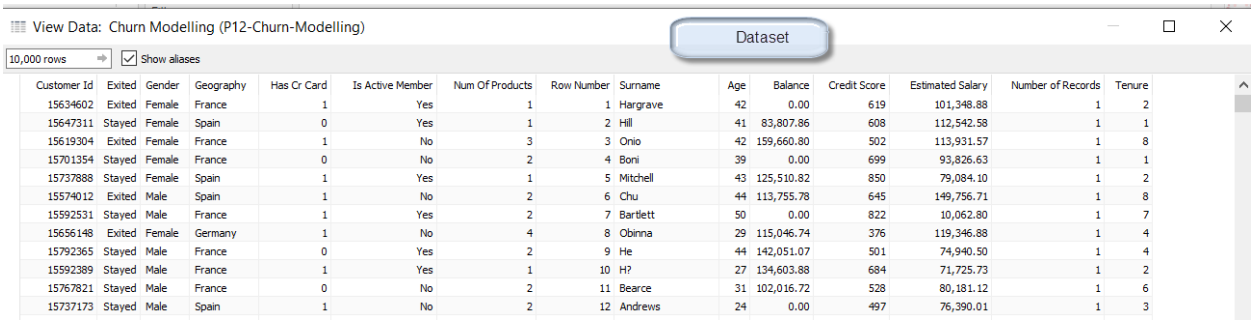
The analysis of data or the exploration of data involves statistics at all levels of exploration process. The visualization of data comes at the very beginning of the process and

may be repeated a few times. It allows the person with “trained eyes” and the person with expertise in the domain of data to summarize, compare and interpret the data. The person with domain knowledge can suggest which variables to include or exclude from the algorithm, and the “trained eye” person can create a visual effect such as a bar chart showing the percentage of customer who left vs those who stayed.

In the next two sections, we present two activities in the EDA process, visualization and quantitative analysis of data pertaining to a bank’s churn rate.

Data Visualization

This use case explores a data set composed of 10,000 records. The goal is to determine what impacts a customer’s decision to stay with the bank. Is it their sex, age, number of product they have with the bank, or is it geography or salary or whether they have a credit card. As shown below, the ‘exited’ column in this dataset is regarded as the dependent variable which is the subject of this analysis, and the rest of the variables are the regressors or independent variables.



Customer Id	Exited	Gender	Geography	Has Cr Card	Is Active Member	Num Of Products	Row Number	Surname	Age	Balance	Credit Score	Estimated Salary	Number of Records	Tenure
15634602	Exited	Female	France	1	Yes	1	1	Hargrave	42	0.00	619	101,348.88	1	2
15647311	Stayed	Female	Spain	0	Yes	1	2	Hill	41	83,807.86	608	112,542.58	1	1
15619304	Exited	Female	France	1	No	3	3	Onio	42	159,660.80	502	113,931.57	1	8
15701354	Stayed	Female	France	0	No	2	4	Boni	39	0.00	699	93,826.63	1	1
15737888	Stayed	Female	Spain	1	Yes	1	5	Mitchell	43	125,510.82	850	79,084.10	1	2
15574012	Exited	Male	Spain	1	No	2	6	Chu	44	113,755.78	645	149,756.71	1	8
15592531	Stayed	Male	France	1	Yes	2	7	Bartlett	50	0.00	822	10,062.80	1	7
15656148	Exited	Female	Germany	1	No	4	8	Obinna	29	115,046.74	376	119,346.88	1	4
15792365	Stayed	Male	France	0	Yes	2	9	He	44	142,051.07	501	74,940.50	1	4
15592389	Stayed	Male	France	1	Yes	1	10	H?	27	134,603.88	684	71,725.73	1	2
15767821	Stayed	Male	France	0	No	2	11	Bearce	31	102,016.72	528	80,181.12	1	6
15737173	Stayed	Male	Spain	1	No	2	12	Andrews	24	0.00	497	76,390.01	1	3

Table 1- Dataset

The data visualization process will attempt to find correlation between independent variable. In the picture below, Tableau shows the bar chart of male and female customers with the percentage of those who stayed and exited. The graph shows that more females exited than males. The chi-

squared analysis from http://www.evanmiller.org/ab-testing/chi-squared.html that Gender in this analysis pertinent. In Figure 2, replacing Gender with 'Has Credit Card' shows that this column does not contribute to the analysis. This analysis is preliminary. Perhaps "Has Credit Card" in combination with another independent variable(s) would be better fit for the model. This is the exercise that keeps the data analysis pinned to their desks!

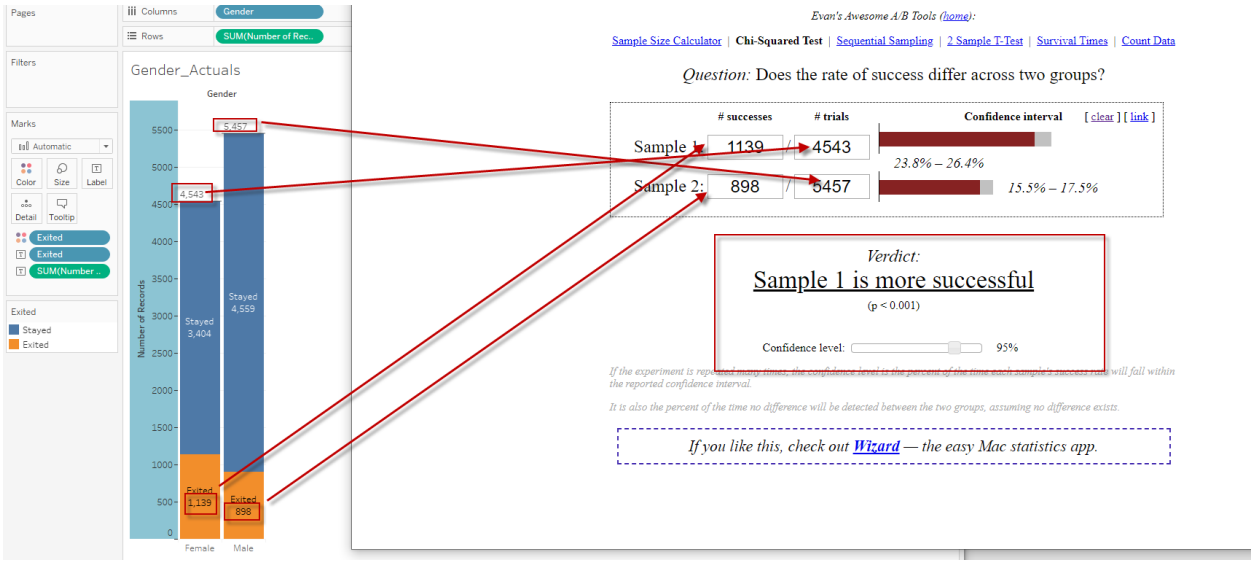


Figure 1- Data Visualization I

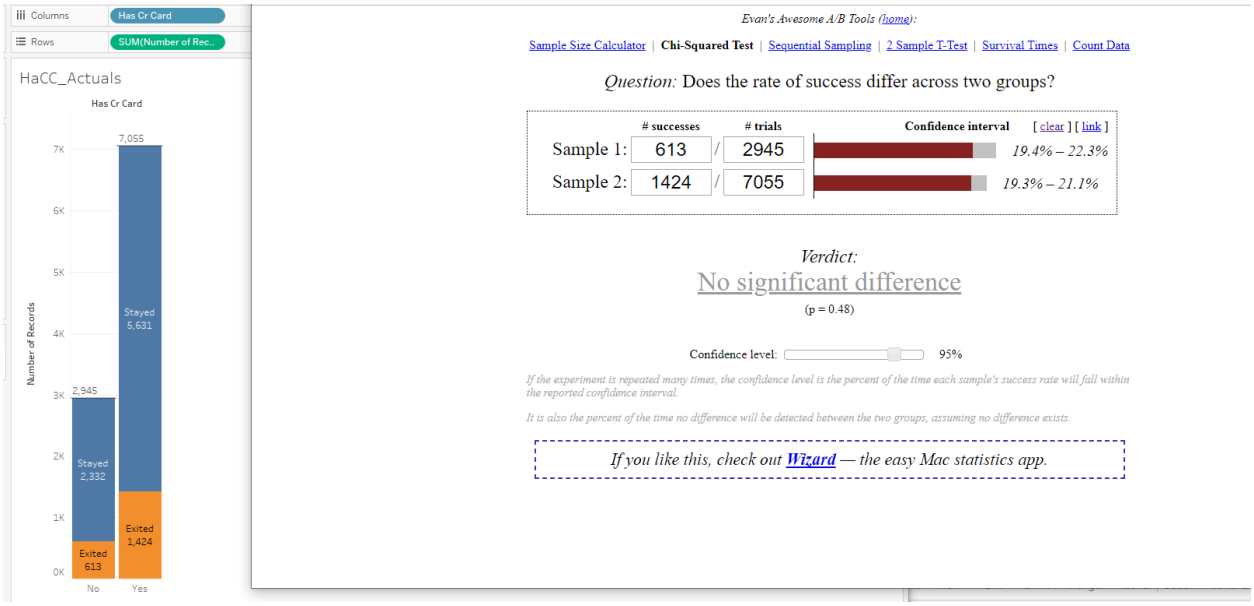


Figure 2 - Data visualization II

Quantitative Analysis

This process involves making decisions which of the independent variables to include in the model. Where in the visualization part, we needed a domain expert to view the data with us, in this phase, we rely on the mathematical results from our choices.

In this use case, we selected logistic regression model in Gretl and performed 5 backward eliminations. We made dummy variables ‘Spain’, Germany’ and ‘France’ from the ‘Geography’ variable and ‘Male’ and ‘Female’ variables from ‘Gender’ variable. We included ‘Female’, ‘Spain’ and ‘Germany’ in the model along with the other independent variables in the dataset. In each run of the model, Gretl recommended removal of a variable. Table below shows the summary of each elimination. The main criteria for keeping a variable in the model was that the p-value to be below out threshold of 0.5 for the variables and the Adjusted R-Squared increasing for each model.

BWElimination Number	Variable eliminated	Variable P-Value	Model's Adjusted R-squared before/after removal	Adjusted R-Squared Difference
1	Spain	0.6181	0.150787/ 0.150961	0.000174
2	HasCrCard	0.4489	0.150961/ 0.151102	0.000141
3	EstimatedSalary	0. 3091	0.151102/ 0.151197	0.000095
4	Tenure	0.0873	0.151197/ 0.151106	-0.000091

Table 2 - Backward Elimination Analysis

As shown in elimination 4 'Tenure' was removed, but not by recommendation from Gretl, but because we wanted to see the impact of removal to test the p-value threshold. It shows that the Adjusted R-Squared was not impacted by much, so we reincluded 'Tenure' in the model. After transforming the 'Balance' variable to Log10(Balance +1) for better uniformity, we got the result shown below.

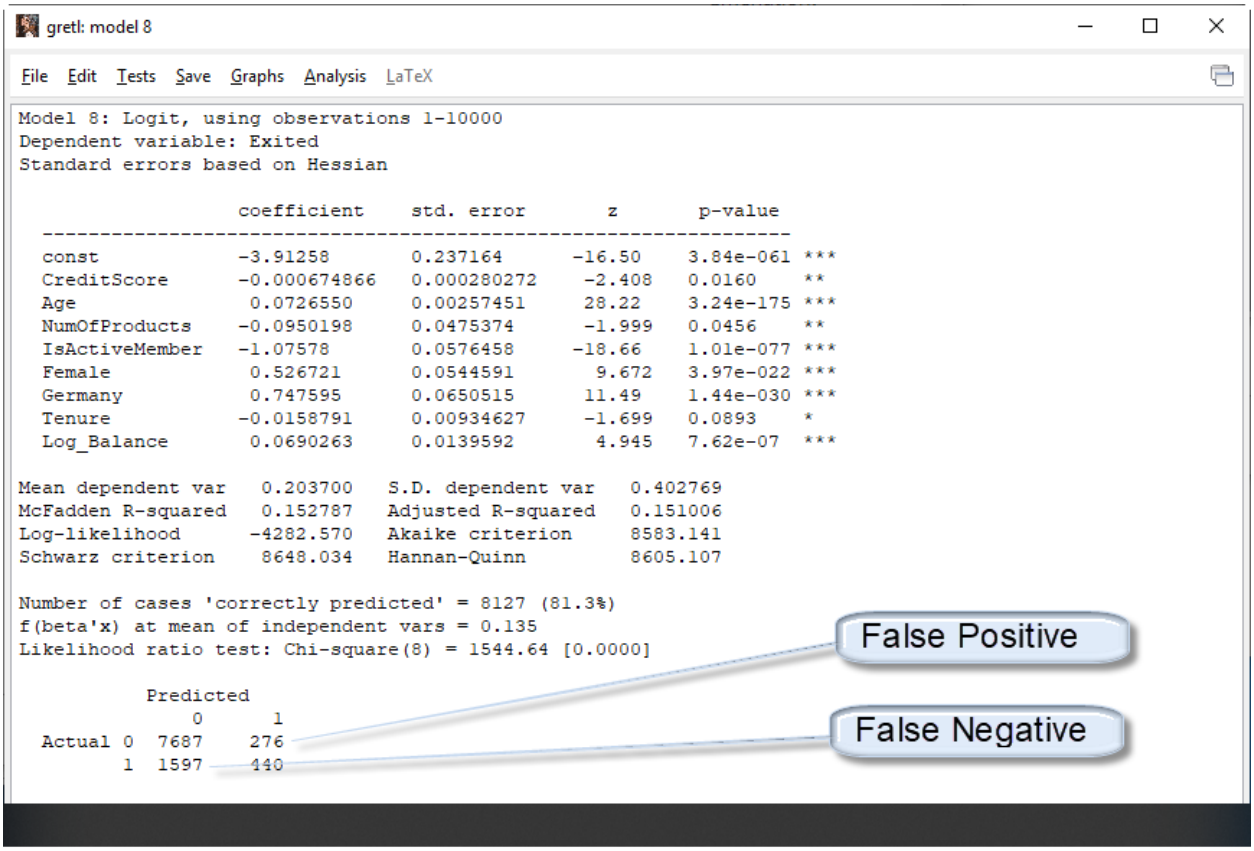


Figure 3 - Confusion Matrix

From the confusion Matrix, the accuracy and error rates were calculated as shown below.

Accuracy Rate = Correct/Total = (7687+440)/10000 = 81.27 %

Error Rate = Wrong/Total = (276+1597)/10000 = 18.73 %

So this analysis showed that the accuracy rate based on the chosen independent variable and using the logistic regression algorithm gives us an accuracy rate of 81.27. This could be acceptable for a bank, because the accuracy or lack thereof might hist the banks bottom line, but it would not cause any harm to anyone (physical or financial). We would however have to reconsider our approach if the accuracy was detrimental to health and safety of people, animals

or the environment. In those cases, we must deem 81.27% unacceptable and go back to data collection and analysis phase.

Conclusion

Data analysis is an essential part of data science project. It involves deep understanding of the data, the domain of the data and all the statistical knowledge and know-how we can through at it. Without it, the outcome of the project will be inaccurate. Advances in technology will help shorten this cycle and lead to better and more beneficial outcome to the overall project. However for now and for the foreseeable future, this step will be among the first steps data scientists will have to take to begin the data science project.

References

1. Rahul Gupta,Michael Koffie,Brandon May,Tushar Muley,Edris Safari –"As The World Churns: Customer Data, Business Models, and Predicting Customer Trends and Behaviors". Final project DSC500 Bellevue University
2. https://www.kdnuggets.com/2017/04/value-exploratory-data-analysis.html
3. Kelleher, John D. Data Science (MIT Press Essential Knowledge series) (p. 1). The MIT Press.