|  |
| --- |
| Data Analyst Report |
|  |

|  |
| --- |
| **April 2022 | Richard Valades** |

****

Blackwell Electronics

|  |
| --- |
| Lessons Learned: |
|  |
| Profiling the Data: Task 1 When beginning a data analysis project, it is essential to “profile” the data, so that the Data Analyst can begin to understand what the data “means.” This exercise is foundational to the process as it is likely that the analyst will modify some aspects of the data to better prepare it for more advanced analytical techniques. This would be extremely challenging without a thorough understanding of the features of the data.  In the process, I employed the following techniques to understand and prepare the data for analysis:   * Import – Bringing the data into the Data Science software platform * Evaluate – Using Python to give us some basic information about the quality of the data * Preprocess – Using Python to perform data “cleanup” (remove duplicates, null values) * Discretize – convert continuous variables such as “sales” into discrete “bins” * Analyze – Using statistical tests to “describe” the data   This exercise is critically important to perform because more advanced analysis is dependent on how the data is prepared in this process. In addition to preparing the data, the Data Analyst may also develop a greater understanding of some properties and relationships between the variables, even at this stage.  Beginning a data analysis project from scratch can be daunting. Based on my own experience, here are some recommendations I would like to offer to any new analyst taking on this task:  **Environment Setup:** Be sure to follow the platform installation steps precisely and have any documentation readily available. Sometimes there are issues in the setup process, so reviewing the documentation prior to setup will make it easier to troubleshoot any installation errors. It will also help you to know what questions to ask if you search for a solution on-line or from another resource.  **Reviewing documentation:** For any Python libraries you may use in your analysis, it is a good idea to review this documentation as well. The algorithms used in this process are complex and are easily prone to error if an input or formatting mistake is made. Having familiarity with how these libraries function will make it easier to understand errors and warnings and will provide guidance on what to do if there is a problem.  **Keeping Organized:** Practicing good folder and notebook organization is essential. As the analyses become more complex, and as you develop more pipelines, knowing where things are will make your job much easier. |

# Creating and interpreting visualizations: Task 1

Rarely in Data Science will an analyst simply look at data in a table and by inspection alone be able to derive keen insights. Data visualization is an extremely useful tool for the analyst to “tell the story with data.” For consumers of the analysis, it can be much easier to understand relationships, comparisons, measurements, etc. if the data is presented visually, versus in written form or even verbally. This process also assists the Data Analyst in determining which features are most important for analysis and which are less important.

The visualization techniques I used in the analysis were:

* Plotting Histograms for different features of the data
* Creating Scatter Plots to see the distribution of the points of intersection
* Box Plots to understand the shape and distribution of the data
* Stacked Bar Charts to compare the values for each variable
* Correlation Matrix to show correlation coefficients between variables

These exercises proved incredibly useful in determining if I could answer the business questions with this data. The visualizations also proved to be essential in understanding the relationships between the features.

For future data analysis projects, I would encourage the Data Analyst to create multiple charts and graphs, and to experiment with different variables. The technique is far more effective at elucidating relationships between the different points than by table data alone.

# Exploratory Data Analysis: Task 2

A critical step in any data analysis project begins with an Exploratory Data Analysis. Virtually identical to the steps taken in Task 1, this time I approached the project with a different set of questions in mind. A best practice in Data Science is to thoroughly understand the business question and to have a clear idea of what we are trying to find in our analysis. Lack of understanding will likely result in an outcome that might be thorough but fails to meet the needs of the decision maker or stakeholder. What’s worse, the Data Analyst may come to the wrong conclusions and provide unhelpful or confusing information.

My advice to the Data Analyst that is taking on a project for the first time is, before performing any Exploratory Data Analysis process, ensure that you are very clear on what the Business is trying to determine. As you move through the analysis and generate charts and graphs, revisit the question at each step to determine whether you are still answering the question.

# Training and fitting a model: Task 2

The goal for task 2 was to create a machine learning model that would determine if we could predict customer behavior with our data. I was also tasked with testing an assumption related to whether two features had any correlation. The previous steps were to prepare the data for modeling. The next steps were to train and fit the machine learning model. The techniques employed in this process are as follows:

* Split the datasets using Python and a machine learning library
* Create training and test datasets to “teach” and test the model
* Perform cross validation checks to determine the reliability of the model
* Evaluate the performance of different ML models

Based on the results, I observed that the model functioned properly and could make predictions. Unfortunately, the model was not 100% predictive with the available data. It is likely that by employing feature engineering techniques, the accuracy of the model could be improved.

For the Data Analyst, I recommend being very fastidious about naming. In the test cases, multiple test and train datasets were created with different dependent and independent variables. As your pipeline becomes more complex, it is very easy to get confused. Creating naming conventions that help clearly differentiate what dataset you are working with will save a lot of time and confusion as you move through the process.

# Decision Tree and Cross-Validation scores: Task 2

Interpreting the outputs from the Decision trees and the Cross-validation scores may have been the most challenging tasks for this project. I found in the analysis that depending on which DV’s and IV’s we used, the model was more or less predictive. Fortunately, I had good success running the models and was able to make determinations, even if the results were not definitive.

The decision tree provided insight into the predictive value of some variables, and I was able to infer a great deal to help answer the questions. Even if I could not determine what was especially helpful, if it’s possible to determine what is not helpful, then it’s possible to look for answers in other ways. The process gave me many insights not only into the data, but to what steps I might take next to explore the business case further.

For the Data Analyst, I would encourage them to explore feature engineering methods, parameter tuning of the ML model and exploring the possibility of obtaining additional data.

# Recommendations for the Future:

Iwould recommend that Blackwell continue to pursue data analytics efforts to better understand its market and what opportunities may exist. Indeed, in the analysis I found multiple areas of opportunity where certain customer populations and geographic regions might be under-served. I was also able to challenge some assumptions from the data that might have had negative consequences had those assumptions been turned into actions. There is a great deal more exploratory work needed to fully support the Executive leadership as they consider strategies for growth based on data-driven decision making. I would encourage you to consider the value that these data analytics projects bring to the company, and how these efforts translate into tangible results for Blackwell Electronics.