UNIVERSIDAD CENFOTEC

Data Analytics / Big Data 2018c-3 (Feb 25, 2020 session)

Program 5 Task 1

Data Science with Python

Get Started with Data Science and Python

PRESENTA:

Alberto Madrigal Jiménez

ASESOR:

Gustavo Rojas

San José, CRC. 2020

Framework One – Selected: *Zumel and Mount.*

Reason:

Based on the business needs, and the parts of the each framework, the framework of Zumel and Mount is almost perfect to explain to the stakeholders every single step, before, during, and after of the new solution. The way to ask for each subject creates an easy and fast way to achieve the main idea and draw quickly the path that we will be working.

Define the goal

- Why do the stakeholders want to do the project?
  - Decrease the number of customers who have defaulted on loans.
- What do they need from it?
  - Don't risk losing business if the problem with the customers is not solved.
- Why is their current solution inadequate?
  - First of all, they are not using a framework to understand the problem, and analyze the possible solutions; also, they are not considering behaviors of the data.
- What resources do you need?
  - I will need a consolidated plan of attack and Python with all its libraries, to manipulate the data and find the best option to solve this issue.
- How will the result of your project be deployed?
  - With the correct process, this project is high level for the business. Due to that, we need to validate each result against the current values and also, check them with the SME's of the project.

Collect and manage data

- What data is available?
  - The necessary fields to understand the behavior of the last 6 months, includes history of past payments, consumptions, Amounts, etc.
- Will it help to solve the problem? Is it enough?
  - For now, we believe the answer is yes, cause based on the previous experiences with other datasets, we achieve good results with similar data.

- Is the data quality good enough?
    - Yes, and also, the data dictionary provided in the case, with the explanation for each field can help to understand quickly each value.

Build the model

- Which techniques might I apply to build the model?
    - The list of the techniques should be applied to build a strong model are:
        - Data Extraction: In this case, we have the dataset.
        - Data Clean: Checking the current dataset, we need to improve the headers of the file, because, it is not in the layout and correct format to be loaded.
        - Diving Deep: Validating the excel file, we need to understand if to improve the models, we need to change the numerical values for string, or maybe, create new cluster to reduce the count of groups by type.
        - Evaluate and deploy: The last and probably the most important step, lunch the data results and evaluate them is very important to mark as achieved the project.
- How many techniques should I apply?
    - The correct answer for this answer should be the necessary, but, it depends of the outcomes obtained. Previously we mentioned some techniques, but when we define the new models, according to the values caught by the final variables, we will define the number of techniques to reach the expected numbers.

Evaluate and critique the model

- Is the model accurate enough to meet the stakeholders' needs?
    - Yes, it achieved all the needed results.
- Does it perform better than "the obvious guess" and any techniques being used currently?
    - Yes, because it is more accurate.
- Do the results of the model make sense in the context of the real-world problem domain?
    - I compare manually some obtained results against real cases and it shows a good expectative.

Present results and document

Once you have a model that meets your criteria, you will present your results to your project sponsor and other stakeholders.

- How should stakeholders interpret the model?
  - We will provide all the necessary data to interpret each model, I mean, it should be very accurate, because, is very important to avoid different interpretation with the data results obtained. As we worked in the past, we will provide summaries, Pivot Tables, KPI's, comparative chart and other tools to visualize the models, their features, and results.

- How confident should they be in its predictions?
  - As high as we can, without damage the analysis, a model over fitted with 95% or higher, does not mean that is a good model, maybe, the dataset should be checked. The best way to understand and answer this question is to apply the correlation matrix and then validate the summary of the false and positive values.

- When should they potentially overrule the model's predictions?
  - As we explained previously, when the data is very inaccurate, or with the data is so accurate that looks like a necessary the application of the model.

Deploy and maintain the model

- How is the model to be handed off to "production"?
  - It is a strong model with a good QA process applied. If we need to modify it, is really important join a call with one SME in Python and the Business, additionally, a second QA process to validate the new results.

- How often, and under which circumstances, should the model be revised?
  - In many cases this requires enhancement of the requirements based on customer feedback or in some cases fixing bugs.

Data source:

The data source was provided in a file named: default of credit card clients.csv. We can see the following list of the fields.

- LIMIT_BAL
- SEX
- EDUCATION
- MARRIAGE
- AGE
- PAY_0
- PAY_2
- PAY_3
- PAY_4
- PAY_5
- PAY_6
- BILL_AMT1
- BILL_AMT2
- BILL_AMT3
- BILL_AMT4
- BILL_AMT5
- BILL_AMT6
- PAY_AMT1
- PAY_AMT2
- PAY_AMT3
- PAY_AMT4
- PAY_AMT5
- PAY_AMT6
- default payment next month

And we can find the explanation of each row in the document provided in the case called: DataSourceUpdated5.18.docx

Data Management:

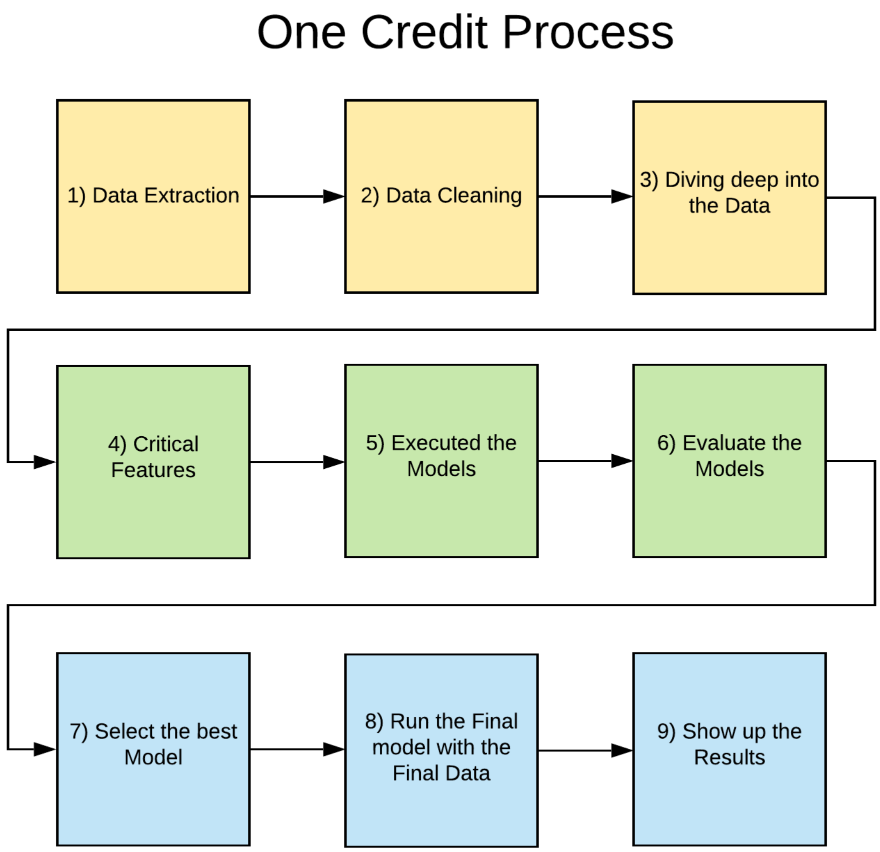
As all the dataset, we should understand every single column, its values, and define if you can use a numerical value or it requires a cast to a new range, a non-numerical value or something else, but it depends of the business case.

The data management will be applied as the rest of the previous project, trying to figure out and find the best route to get the expected values.

Issue and how to drive it:

This specific dataset has a good quantity of columns and important data to reach out the answers required by the business side, however, we need to improve the headers of the file, due to That, it cannot work as a table to be loaded in the system. Probably if we want to improve the model, we will need to change some numeric values to string, and if we want to reduce the final options, we need to create a new cluster with flags as: good, intermediate, low; and other similar cases.

Flow Chart Visualization:



Addition:

Was it straightforward to install Python and all of the libraries?

I did not face any issue installing Python, I used the attack plan, and applied the guideline provided in the resources tab, due to that, the installation was easy and fast.

Was the tutorial useful? Would you recommend it to others?

Yes, I am a system engineer, but I did not have any experience with Python, in my personal case, the use of the tutorial was very useful and obviously I will recommend this one to other professionals.

What are the main lessons you've learned from this experience?

The main lesson learned in this experience was understand deeply the way to work of the Python, compare that with R, and start the use of the basic command in the command line of Anaconda.

End