universidad
cenfotec_

Certificación Internacional Data Analytics y Big Data

Prof. Gustavo A Rojas

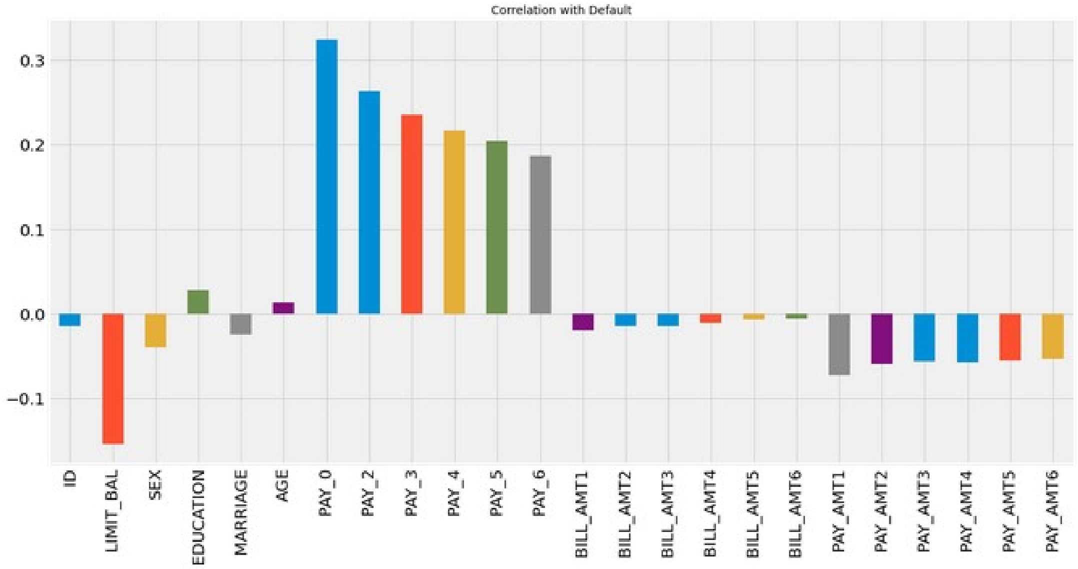
por: Federico Ruilova A

San José, Costa Rica 31 de Marzo , 2020

Task 4 : Complete a Data Science Capstone Project

LESSONS LEARNED REPORT

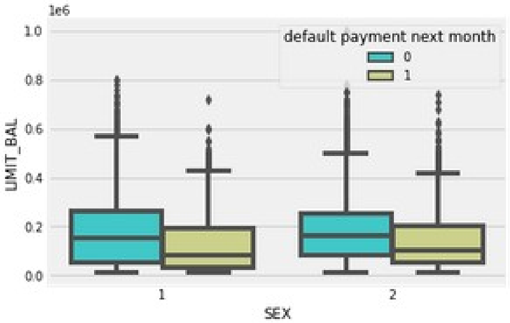
SUMMARY OF FINDINGS AND POTENTIAL

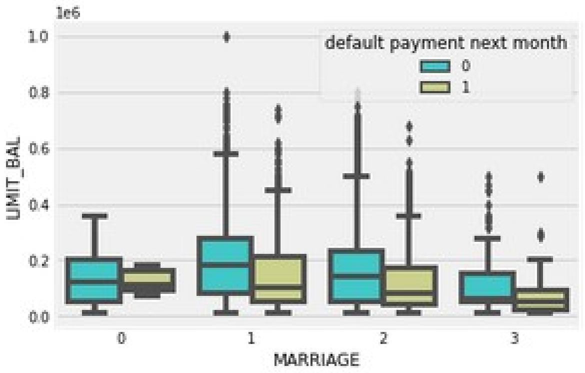


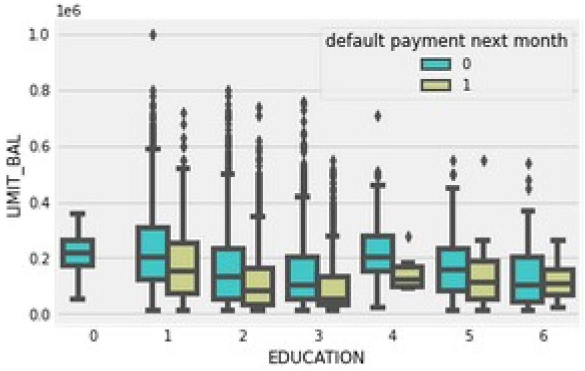
1. The correlation between the PAYX status in the observations and the dependant variable is basically the most important finding. High correlation leads to higher acuray when a model is develop .

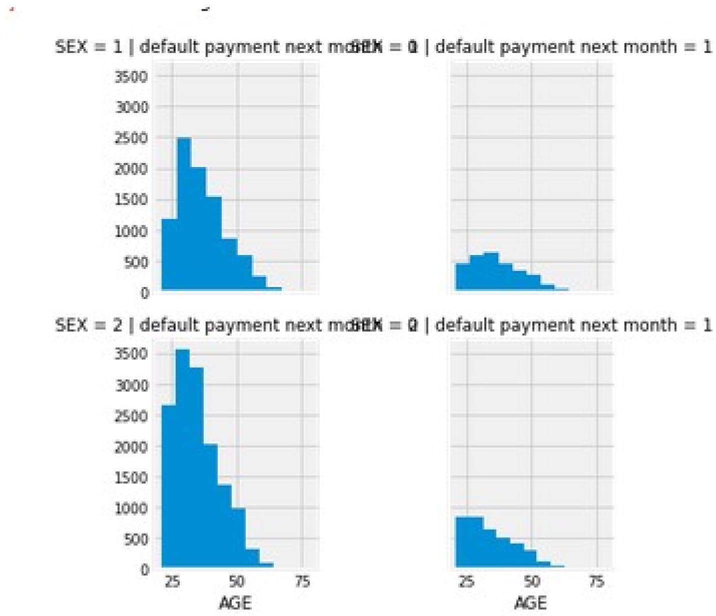
2. In the correlation Matrix data showed high correlation between PAYx and PAYx beeing x different when comparing , meaning that tha the repayment status leads to predicatable results.

3. Sex and age are relevant predicting for certain populations, same as marriage and education.









<u>LESSONS LEARNED</u>

1. working with python / jupyter could be very handy to handle resources and data in an easier way , language was easy to learn and adopt .

2. With the correlations found, cleaning the data again could be a better way to develop models

3. Asumptions are fine, but theories must be proven with explratory data analysis .

RECOMENDATIONS

1. To have a more acurate models , and based on the correlations, a multidimensional model could be develop, meaning that separating data into age groups and gender could lead to different findings when testing models.

2. Some other features or considerations for predicting default status could be based on what are the spent cateories of each individual. For example a theorie could go like this : Individuals that eat more fast food tend to default more than individuals with higher grocery spending .