Universidad Cenfotec.

Big Data & Analytics.[3]

Data Science and Big Data.

Task 3: Final Report - Credit One.

Mentor: Diego Alfaro Bergueiro.

Estudiante:

Giovanna Francesa Alfaro.

24 de marzo de 2020

universidad cenfotec_

Este proyecto se ha enfocado en establecer una forma de identificar a los clientes bancarios que no pagarán un crédito, de manera que se puedan filtrar al momento de que el banco tome la decisión de otorgarlo o no y así no tener pérdidas económicas.

De proyectos y experiencias anteriores, la compañía ha aprendido las siguientes lecciones, las cuales, deben tenerse claras y ser contempladas como hechos en este problema:

- No se puede controlar los hábitos de gasto de los clientes.
- No siempre se puede pasar de lo que encontramos en nuestro análisis al "por qué"subyacente.
- Debemos centrarnos en los problemas que podemos resolver: ¿Qué atributos en los datos podemos considerar estadísticamente significativos para el problema en cuestión? ¿Qué información concreta podemos derivar de los datos que tenemos? ¿Qué métodos probados podemos usar para descubrir más información y por qué?

Todas estas características fueron tomadas en consideración, de modo que se estableció unas serie de procedimientos que ayudara a mantener el orden y a mantenerse enfocados en responder la pregunta al problema: cómo identificar a clientes que no pagarán adecuadamente los créditos.

Como primer paso, se eligió el framework Badin, como referencia para establecer una solución en conjunto con las partes interesadas, este es un framework sencillo de aprender y de aplicar para resolver problemas, además de que el proceso que sigue ayuda a garantizar la satisfacción de los clientes y acoplarse a sus necesidades y requerimientos. El mismo fue elaborado y entregado a las partes interesadas en etapas más tempranas de este proyecto.

Como segundo paso, se procedió a realizar un análisis exploratorio de los datos, este paso es bastante importante ya que permite identificar y visualizar información relevante cuando se quiere analizar y encontrar patrones en un conjunto de datos o problema bajo estudio, por lo cual es valioso realizarlos siempre que se empiece algún proyecto de *data science*.

A pesar de que en este problema nos enfocamos en encontrar una manera más precisa de identificar clientes que no cumplen con el pago de sus créditos, cabe resaltar que de este análisis exploratorio se encontraron algunos patrones interesantes que podrían ayudar a nuestros clientes, a potenciar, mejorar y promocionar sus negocios, dichos hallazgos pueden ser discutidos en futuras iteraciones de este proyecto.

A continuación se resumen algunos de los hallazgos encontrados, los cuales brindan alguna información acerca de atributos que pueden ser estadísticamente significativos además de información concreta que se puede derivar de los datos.

- De la totalidad de los créditos otorgados, el 22 % presentaron problemas en el pago.
- Con respecto a la edad de los clientes, a partir de aproximadamente los 45 años y en adelante, se tiene la mayor proporción en créditos en estado *default* o no pago, esta característica también se observó en personas entre los 18 y 25 años aproximadamente. Por otro lado, personas en edades entre los 25 y 35 años, muestran la mayor proporción en créditos pagados correctamente. Lo anterior se puede observar en la siguiente gráfica:

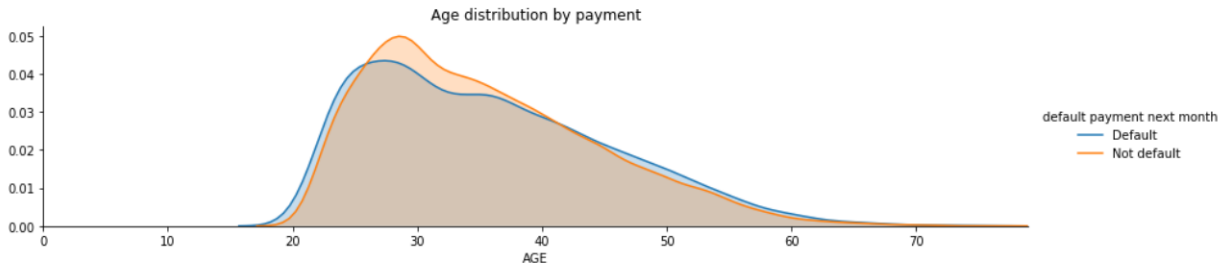

Figura 1: Distribución de edades según condición de pago.

- Con respecto al nivel de educación, se observa que a mayor nivel nivel de educación, mayor cantidad de créditos son otorgados. Un 35 % de la

totalidad de los créditos fueron otorgados a *graduate school* y un 47 % a *university*. Para la totalidad de créditos otorgados a *graduate school*, el 20 % presenta la condición *default*, mientras que para *university* corresponde a un 24 %. La cantidad de créditos otorgados a personas en *high school*, es considerablemente menor en comparación con las dos categorías mencionadas y en este nivel, el 25 % presenta *default*. Esta información se resume en el siguiente gráfico:

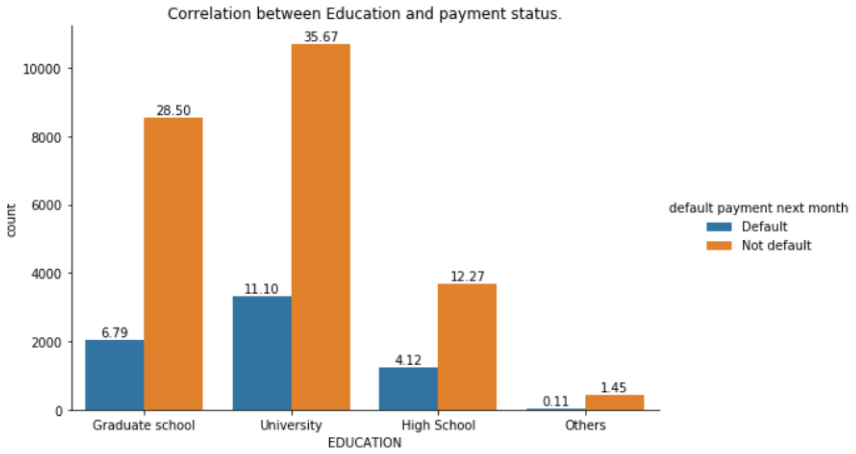

Figura 2: Niveles de educación según condición de pago.

- Con respecto al estado civil, el 98.75 % de los préstamos son otorgados a personas solteras y casadas. De la totalidad de créditos otorgados a personas solteras, el 21 % no pagaron, mientras que de la totalidad de créditos otorgados a personas casadas, un 23.4 % no lo hicieron.

  La proporción de personas divorciadas a las que se les fue otorgado un crédito es practicamente nula en comparación con personas que registran el estado civil casado o soltero, siendo su tasa de no pago un 26 %. La gráfica se muestra en la figura No.3.

- Con respecto a género, el 60 % de los créditos son otorgados a mujeres mientras que el 40 % a hombres.

  De la totalidad de creditos otorgados a mujeres, el 20.8 % reporta estado *default*, y en el caso de créditos otorgados a hombres, el 24.2 % reporta

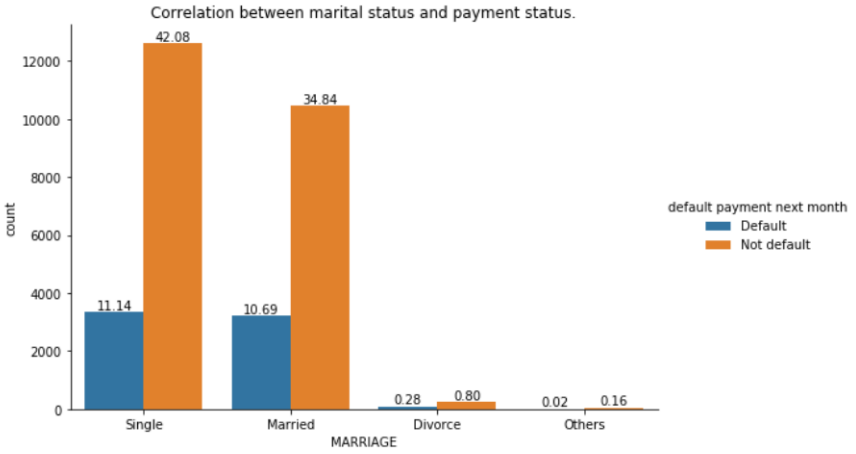

Figura 3: Estado civil según condición de pago.

default, lo cual se puede observar en la gráfica No.4.

Figura 4: Género según condición de pago.

Los métodos de análisis de exploración de datos, en conjunto con los modelos de predicción descritos a continuación, fueron de suma utilidad para

descubrir características y relaciones entre los datos, así como para lograr encontrar una manera más precisa de predecir e identificar que clientes no pagarán a tiempo sus créditos.

Dichos métodos de predicción, fueron creados utilizando diferentes técnicas conocidas como métodos de selección de variables o *feature selection*, este es un proceso mediante el cual automaticamente se selecionan aquellas variables que más contribuyen a la adecuada predicción de la variable dependiente o variable a predecir, esto se realiza debido a que tener variables irrelevantes en los datos puede entregar una menor precisión en los modelos. Entre los métodos empleados se encuentran: RFE *Recursive Feature Elimination* y PCA *Principal Component Analysis.*

También se aplicaron al menos tres modelos de *machine learning* distintos, conocidos como *Random Forest*, *Decision Tree* y *K-Nearest-Neighbor.*

El modelo más óptimo generado, corresponde a *Random Forest*, utilizando el conjunto de datos original, es decir, sin aplicar ningún método de selección de variables. Este entrega una precisión de aproximadamente 82 %, lo cual significa que 82 clientes de cada 100 se predecirán correctamente.

En los métodos de predicción actuales, se tiene una precisión de aproximadamente 78 %, es decir, de cada 100 créditos, 78 son pagados adecuamente y 22 no lo son. Esta información se muestra en la siguiente gráfica:

Con el nuevo modelo, se logró incrementar la precisión en aproximadamente 4 %. En etapas iniciales, se obtendrá aproximadamente el 82 % de precisión en nuestras predicciones, sin embargo, conforme se genere datos nuevos se podrá realimentar los modelos creados con la nueva información y eventualmente podremos incrementar la precisión de los mismos.

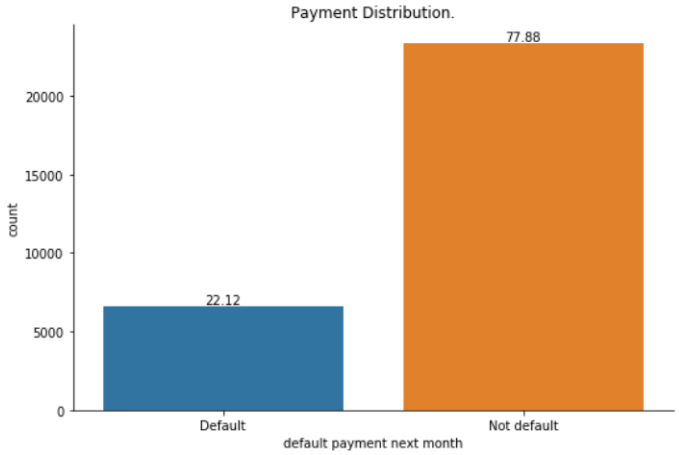

Figura 5: Distribución de créditos pagados y no pagados.