Diego Gamboa

Ahora que ha creado un entorno de programación local, se comenzará el trabajo preparando y explorando los datos.

Prepare and Explore the Data

Leasson Learned

# Lecciones Aprendidas

* El tiempo invertido y el esfuerzo concienzudo de lograr la verdadera pregunta a responder, permite
  + Definir los atributos realmente importantes a incluir en el estudio
  + Definir los métodos a utilizar para dar solución al problema planteado
  + Es una forma de identificar individuos con conocimiento experto que puedan aportar valor e interpretación a los datos que se poseen.
* Los datos históricos mostrados dan valor al negocio y su necesidad de definir perfiles de clientes con gran potencial de incumplimiento en sus créditos.
* El valor de la experiencia adquirida con el Lenguaje de Programación R y la metodología seguida durante tareas previas, permitieron un decremento sustancial en la curva de aprendizaje y aplicación de conceptos, ahora aplicados con el lenguaje Python.
* Los conceptos ya aprendidos sobre Preparación de Datos, Limpieza de Datos, reducción de datos, Transformación de Datos e interpretación de visualizaciones, en lugar de perder vigencia se demostró fueron valiosos, aplicados y totalmente utilizados para la ejecución del nuevo caso presentado en la tarea.
* Se ahondaron en la aplicación y diversas posibilidades del uso de la graficación que permite Python, logrando gráficas con mayor valor, como por ejemplo histogramas que comparaban varios atributos o scatter con inclusión de leyendas , formas y colores que mostraban y simplificaban la lectura de los datos.
* La Matrix de Correlación mostró nuevamente su gran utilidad en la discriminación de atributos, logrando remover elementos que desviaban la veracidad de los resultados futuros.
* Aunque la forma por defecto de la Matriz de correlación da muchos datos, se consultaron nuevas formas de dar otros tipos de salida a la información como por ejemplo
  + ordenar los valores de forma columnar en donde se observaron mejor los atributos con mayor y menor correlación.
  + Restringir los valores a mostrar (Mayor a 0.85 de correlación por ejemplo)
  + Eliminar duplicados (A-B es igual a B-A)

Todo esto logró una rápida y veraz forma de tomar decisiones con respecto a atributos que no generaban valor.

* Los Datos obtenidos por la Covarianza[[1]](#footnote-1) o medida de dispersión, serán un gran soporte para determinar variables con relaciones directas (cuando uno variable crece la otra variable también) y relaciones inversas (cuando uno variable crece la otra variable decrece ), lo cual significará información sobre el comportamiento de la selección que hagamos.
* Los Datos sobre históricos de pagos realizados (PAY\_XX) y cantidades de pagos previos (PAY\_AMT\_XX) resultaron los atributos estadísticamente más significativos.

1. https://conceptosclaros.com/que-es-la-covarianza-y-como-se-calcula-estadistica-descriptiva/ [↑](#footnote-ref-1)