LABORATORIO DE INTELIGENCIA COMPUTACIONAL
15 DE MARZO DE 2018

Modelo de Clasificación Predictivo de Credit Scoring

DESCRIPCIÓN DEL PROBLEMA

La globalización y la competencia entre países se han intensificado a nivel mundial, observándose que el financiamiento y acceso a la tecnología han sido dos ejes que han contribuido al crecimiento de ciertas industrias. La mayor competitividad ha afectado de sobremanera a diversas empresas, siendo especialmente afectadas las pequeñas y medianas (PyME's). En la práctica, este tipo de empresa no cuenta con flujos permanentes de efectivo y poseen un acceso limitado al financiamiento.

Para resolver el problema del financiamiento en las PyME's algunos gobiernos han creado organismos gubernamentales que administran fondos concursables para financiar las diferentes iniciativas de inversión de las PyME's. Ahora bien, una tarea que ha sido muy difícil de realizar en la práctica corresponde a la definición objetiva de las características que deben tener las solicitudes de crédito para adjudicarse estos fondos concursables. La importancia de esta tarea recae en que una mala asignación de estos fondos aumenta drásticamente la probabilidad de quiebra de las empresas, que a pesar de merecerlo, no se lo adjudican (error tipo I). Además, si consideramos que estos fondos concursables son obtenidos de las arcas fiscales existe un interés político de que estos fondos sean bien utilizados por el gobierno. Si una solicitud de crédito es aceptada y no recuperada afecta negativamente los indicadores de gestión, ya que en la práctica disminuye el fondo total concursables. Por esta razón, nace la necesidad de incorporar nuevas tecnologías y enfoques analíticos innovadores, que permitan dar una mayor objetividad en la decisión de aceptación de una solicitud de crédito.

Suponga que una institución financiera de PyME's lo ha contratado a usted para que desarrolle un modelo de Credit Scoring que le permita diferenciar si un cliente va a pagar o no el crédito que solicita. Para estos fines esta institución le ha dado acceso a su Datawarehouse con información de sus transacciones, además de las características personales. En general, las instituciones financieras utilizan Árboles de Decisión para tramificar las variables y Regresiones Logísticas para construir estos modelos de clasificación. Sin embargo, dado que saben de su conocimiento en Inteligencia Computacional le han pedido que construya el modelo también probando técnicas tales como Support Vector Machines y Redes Neuronales Artificiales, y comparando sus

resultados de predicción con respecto a la combinación Árbol de Decisión – Regresión Logística.

El objetivo de su trabajo entonces será apoyar y dar solución a la problemática del banco incorporando algunos conceptos de Machine Learning en un ambiente real y de negocio. Un objetivo secundario, pero muy importante, es el desarrollo de habilidades comunicacionales para explicar con mayor detalle la problemática de negocio, qué desafíos presenta hoy por ejemplo en las instituciones financieras chilenas este asunto, entender los datos disponibles a través de técnicas de aprendizaje no-supervisado, obtención de variables calculadas más adecuadas para modelar la respuesta, modificación de variables fuentes de modo de maximizar las capacidades predictivas de cada algoritmo a usar, identificar variables que de acuerdo a sus valores son inconsistentes con la realidad del problema y la generación de la predicción del modelo para el conjunto de puntuación que se describe más abajo. En aquellas variables donde considere que su descripción es limitada puede crearse su propio contexto en base al entendimiento del problema.

INFORMACIÓN DISPONIBLE

Se cuenta con un archivo llamado **CREDITRISK_RAW**, el cual contiene 2.294 clientes y, para cada uno de ellos, se cuenta con las siguientes variables descriptivas:

VARIABLES	DESCRIPCIÓN
ID	Identificador del cliente
GENERO	Genero del cliente
RENTA	Renta en pesos
EDAD	Edad en años
NIV_EDUC	Nivel Educacional
E_CIVIL	Estado Civil
COD_OFI	Código de la Oficina donde se realiza la solicitud
COD_COM	Código de la comuna donde está la Oficina
CIUDAD	Ciudad donde se realiza la solicitud
Crédito_1	Monto crédito 1
Crédito_2	Monto crédito 2
Crédito_3	Monto crédito 3
Crédito_4	Monto crédito 4
Monto solicitado	Monto actual solicitado
Días de Mora	Número de días que ha estado en mora (histórico)
Monto Deuda Promedio	Deuda promedio Anual
Número de meses inactivo	Número de meses en que no tiene el negocio activo
Número de cuotas	Número de cuotas que solicita el crédito actual
Aval	Con o sin aval
PAGA	Target

Tabla 1: Descripción de las variables

El segundo archivo llamado **CREDITRISK_SCORE** provee de otro conjunto con 1.200 clientes descritos con las mismas variables. La diferencia central de este nuevo conjunto de clientes es que no se conoce si pagarán o no. Usted debe determinar a qué clientes de le debe aplicar políticas de retención.

ACTIVIDADES A DESARROLLAR

Los resultados esperados de su trabajo son los siguientes:

1. Desarrollo de un modelo predictivo de pago basado en la información histórica de la institución financiera.
2. Definir el patrón característico de los clientes pagadores y de los clientes no pagadores utilizando sólo la información que los caracteriza.
3. Definir tres acciones a emprender con los clientes que el modelo indique como clientes no pagadores y tres acciones para los clientes que sean pagadores. Se valorará la sensatez comercial, la evaluación económica de la acción y resultados esperados, como lo efectivas que éstas puedan ser para evitar la el no pago de los créditos.
4. Predecir la clase para el conjunto **CREDITRISK_SCORE**, determinando para cada cliente si será pagador o no.

PAUTA INFORME TAREA COMPUTACIONAL

El informe desarrollado debe contener al menos los siguientes puntos:

1. **Índice de contenidos**
2. **Introducción**: Se debe dar una descripción general del problema a resolver, una breve descripción del informe y una clara visualización de objetivos, tanto generales como específicos. No debe ser una copia del enunciado.
3. **Análisis estadístico y exploratorio de los datos**: En este punto deben calcular las principales métricas estadísticas para una mejor descripción de los datos, por ejemplo estadísticas descriptivas, tablas de frecuencia, histogramas, kurtosis, asimetría, etc. Adicionalmente, cada resultado debe ser analizado en términos de los perfiles encontrados. En esta sección deben definir un primer perfil, tanto para clientes pagadores como para los que no son pagadores, basado en el análisis de las variables que describen a un cliente.
4. **Algoritmos de Machine Learning**: en esta sección se deben comparar en términos de error de clasificación el modelo desarrollado. Se debe realizar una pequeña sensibilización de los parámetros de ajuste que definen a los modelos respecto a la medida de performance (error de clasificación).
5. **Determinar el mejor modelo** de clasificación utilizando como criterio el mínimo error de clasificación. Como optativo inventar una matriz de pesos que castigue los errores y/o premie los aciertos.
6. **Políticas comerciales.**
7. **Conclusiones y discusiones**.
8. **Anexos (no en exceso)**.