Universidad Adolfo Ibañez
Facultad de Ingeniería y Ciencias
Segundo Semestre 2022

Minería de Datos

Profesor: Andrés Medina
Ayudante: Leonel Vega
Aplicación: Modelamiento Predictivo en Credit Scoring.

Consideraciones:

a) Tarea grupal (máximo 5 alumnos).

b) Se debe adjuntar el notebook de python y un informe autocontenido de no más de 3 páginas presentando los resultados más interesantes de acuerdo al contexto del problema. Asegurarse que el código funcione, de lo contrario no se evaluará. Enviar el contenido en formato zip o rar con nombre del grupo y sección (ejemplo: HW02_G01_SEC01.zip donde la nomenclatura hace referencia a Trabajo 2 (HW02), grupo 01 (G01), sección 01 (SEC01).

c) **Plazo: 26 de Noviembre 23:59**

Ustedes trabajan en el área de riesgo de una conocida empresa crediticia chilena, que entrega créditos a personas naturales. Dentro de sus tareas se cuenta la construcción de modelos de minería de datos, lo que es de su entera responsabilidad. La empresa se encuentra en proceso de rehacer sus modelos de Behavorial Scoring, es decir, predecir si un cliente será bueno o malo en términos de la devolución del crédito y la información que ya se posee con respecto a los créditos que tuvo anteriormente. Para ello dispone de una serie de variables asociadas a los clientes antiguos, incluyendo el resultado del crédito si el cliente fue *bueno* o *malo*.

La empresa confía plenamente en su política de crédito, por lo que NO necesita hacer inferencia sobre las solicitudes rechazadas, pues la mantendrá a futuro. La empresa es supervisada por la CMF (Comisión del Mercado Financiero), así que recibe la instrucción de utilizar modelos de regresión logísticas y crear un score logarítmico clásico a partir de sus resultados. Sin embargo, el área interna de riesgo crediticio maneja otro tipo de modelos por lo cual, el manager le encarece que ajuste todo el abanico de posibilidades que ud maneja.

Para esta labor, esta contra el tiempo, y su manager también por lo que éste, le pide la planificación del proyecto en 5 etapas las cuales se presentan a continuación.

1. Análisis Exploratorio de Datos

- Generar los principales estadísticos descriptivos de la distribución, gráfico de la misma y evaluar su concentración.
- Análisis de Coherencia: Generar reglas para detectar datos anómalos, como por ejemplo; edad sobre 100 o menor a 18, ingreso menor al mínimo, médico de 20 años, etc.
- Depuración de los Datos. Evaluar permanencia de variables y/o datos, técnicas de imputación, transformaciones, etc.

2. Creación de Variables

A partir de la informacion disponible depurada, genere nuevas variables que puedan ser de interés, la idea de esta etapa, es que estas nuevas varibales tengan sentido económico.

- Ratio de endeudamiento: Deuda de Consumo / Ingreso
- Ratio Línea de Crédito: Línea de Crédito / (Línea de Crédito + Deuda Consumo)
- Edad-Estado Civil
- Proxy Patrimonio (Activos – Pasivos)
- Genero-Estado Civil
- Variables dicotómicas: tiene auto?

3. Análisis Bivariantes

Evaluar variables candidatas para modelar. Para esto, utilice la matriz de correlaciones. Sin embargo, también se les solicita que indage en otro criterio para seleccionar variables en relación al poder predictivo llamado *information value*. Busque información de cómo se implementa ésta metodología en python y aplíquela en su problema.

4. Entrenamiento de Modelos Predictivos

Construya un modelo de Credit Scoring que permita resolver la problemática del banco utilizando todos los modelos vistos en clases: Regresión Logística, Árboles de Decisión, KNN, SVM, Naive Bayes y Ensamblados.

1. Para cada modelo, aplique `cross-validation` de $k = 10$
2. Para cada modelo, aplique un entrenamiento exhaustivo de los hiperparámetros utilizando `GridSearchCV()` o `RandomizedSearchCV()`. Para ello, revise la documentación oficial de `sklearn`.
3. Los modelos entrenados en los pasos 1 y 2, debe hacerlos con `Pipelines`. Para ello, revise implementaciones en la página oficial de `sklearn`.

4. Genere una tabla comparativa de los 6 tipos de modelos con las siguientes métricas y encuentre el mejor modelo.

Modelo	**F1-Score**	**Recall**	**Precision**	**Accuracy**	**ROC**
Regresión Logística					
Arbol de Decisión					
KNN					
SVM					
Naive Bayes					
Ensamblados					

Observación Notar que los modelos que están en la tabla, son los modelos entrenados con los hiperpárametros optimizados.

5. Generación de Predicciones

Aplique el modelo seleccionado de la sección anterior sobre los datos del archivo Sample_OTT.csv. Genere la predicción en este nuevo set de datos en una columna nueva. Si su archivo no cumple con estos requerimientos, su nota de evaluación será un 1.0. Para evaluar este punto, se competirá entre todos los grupos del curso y se asignará un puntaje en base a los resultados obtenidos por cada grupo. El mejor grupo obtendrá un 7.0, y se irá bajando según los resultados de cada grupo. La medida de evaluación de cada grupo estará basada en el `F1-score`.

6. Anexo

La base contiene 52.394 clientes. La variable a modelar es la variable `marca_malo` donde el 1 significa que el cliente no pago el crédito solicitado en un plazo de 12 meses, y 0 si pagó.

- Cliente_Id: Identificador de cliente
- Fecha de nacimiento
- IngresoLiquido: mensual
- Comuna: informada por cliente
- Ciudad.
- Region.
- GSE: estimación modelo externo
- Prof: Profesión o Cargo
- Genero (1=Hombre, 2= Mujer)
- Estado_Civil (1=Soltero, 2= Casado, 3=Separado)
- Antigüedad_Cliente
- Auto_Marca
- Auto_ANIO_FABRICACION
- Auto_TASACION
- Auto_Total_TASACION
- Auto_Cuenta_De_Automovil
- Prop_AVALUO_Total
- marca_malo: Marca morosidad con 12 meses de observación.
- Deu_DConsumo: Deuda total en consumo
- Deu_DLinCred: Disponible en líneas de crédito y tarjetas de crédito
- Deu_DHipotecario: Deuda hipotecaria
- Deu_NumInstAcr: Numero de Instituciones acreedoras