Maestría en ciencias de los datos y analítica

S2266-0166 Estadística Multivariada Avanzada

PREDICCIÓN DEL SCORE CREDITICIO MEDIANTE TÉCNICAS DE APRENDIZAJE SUPERVISADO

Presentado por: Paula Andrea Cárdenas López, Sebastián Giraldo González, Alejandro Rendón Jiménez, Alejandro Velásquez Arango

Profesores: Santiago Hernández y Tomás Olarte

Lugar: Medellín

Fecha: Noviembre de 2022

Contenido

1. Introducción

La evaluación precisa del riesgo financiero es un elemento importante para las entidades financieras, dado que el resultado de esta evaluación les permite hacer un uso más eficiente de su capital económico. Es por ello que la puntuación crediticia es definida como un proceso donde los clientes bancarios son reconocidos para concederles un crédito, basados en un conjunto de criterios predefinidos.

Como resultado de lo anterior, en la última década el uso de modelos de aprendizaje automático para la clasificación del puntaje crediticio ha incrementado notablemente, generando así, el uso de muchas combinaciones de modelos y técnicas que permiten mejorar cada vez más la precisión de esta predicción, usando algoritmos que van desde la ingeniería de características hasta el modelado de datos.

En el tiempo se han aplicado muchas técnicas para clasificar este puntaje, entre ellas, Logistic Regression, Random Forest, Support Vector Machines, entre otras; sin embargo, en diversos estudios se han interesado por usar técnicas como Gradient Boosting, Artificial Neural Network (ANN) y otros métodos de ensamble. Todas estas técnicas tomando como base principal unas características relevantes que ayudan a la precisión de los modelos.

Es importante resaltar que Light Gradient Boosted Machine (LGBM) y ANN son las técnicas con mejor precisión en las predicciones, alcanzando así porcentajes de precisión entre el 80% y 99%. Por tanto, para el desarrollo del trabajo en mención se generarán los modelos de Regresión Logística, Random Forest, LGBM y XGBoost, con el fin de evaluar el desempeño y seleccionar el mejor para este caso.

2. Marco teórico

En este punto se presenta una breve descripción de la metodología empleada para analítica de datos, con el objetivo de aportar al lector cuales son las características más relevantes de la misma.
Adicionalmente, se mencionan los algoritmos empleados en la revisión de literatura para la predicción del score crediticio, además de la prueba del mejor modelo a partir del error de generalización y métricas de evaluación.

Metodología CRISP-DM

Para el desarrollo de este proyecto de analítica se empleó la metodología de IBM, llamada Cross-Industry Standard Process for Data Mining (CRISP-DM); la cual consiste en brindar las fases de un proyecto de minería de datos, así como las tareas y resultados en cada una de ellas. El objetivo de seguir esta metodología es asegurar la calidad del resultado esperado del proyecto, además de tener claramente definidos los roles y participantes en cada una de ellas. Con miras a exponer las fases de la metodología, se presenta la siguiente figura:

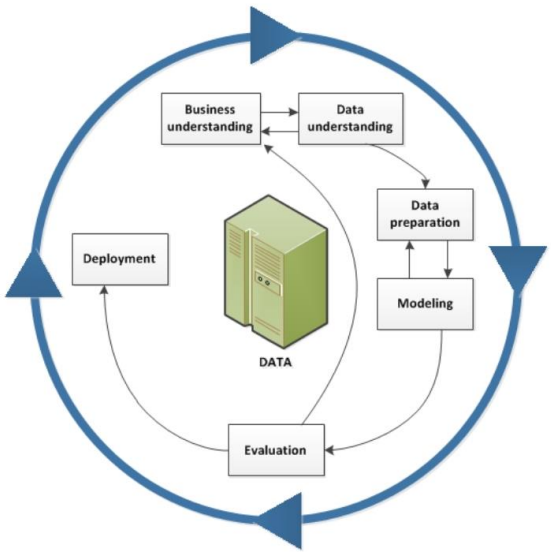


Figura 3: Fases de la metodología.

Es importante resaltar que esta metodología no es estrictamente secuencial, sino que permite adaptar las fases de acuerdo a las necesidades de cada tipo de proyecto en donde se vaya a implementar. Por recomendación de la metodología, dedicar el esfuerzo adecuado a las primeras fases, donde se comprende el negocio y los datos, conlleva a obtener mejores resultados en la fase de preparación de datos; pues esta representa aproximadamente el 50 - 70% del tiempo y esfuerzo del proyecto de analítica. Todo el esfuerzo anterior cobra sentido en la fase del modelado, pues es aquí donde se implementan herramientas analíticas para resolver el problema. (IBM, 2021)

Para el desarrollo de este proyecto no se llevará a cabo la última fase, la cual consiste en el despliegue del modelo, dado que no está incluida en el alcance.

Modelos estadísticos

A continuación, se mencionan los modelos más usados en la literatura para la clasificación del score crediticio.

- **Gradient boosting (GB) algorithm**

  Este modelo es usado para clasificación y regresión, el cual es potente para encontrar relaciones no lineales entre la variable objetivo y las características, esta ventaja se debe a la flexibilidad del modelo para trabajar con una gran variedad de problemas, dado que la función de perdida se puede sustituir de acuerdo al problema y se entrenan sucesivos weak-learners van minimizando la función de error iterativamente. Existen algunas librerías como XGBoost o LightGBM. (Masui, 2020)

- **Random Forest (RF) models**

  Este modelo es usado tanto para clasificación como regresión, uno de los usos en la literatura es para la clasificación del score crediticio, en donde las características de esta técnica ayudan a encontrar la mejor clasificación resultante de las salidas de los modelos de árbol de decisión, considerando así, que este es un método de ensamble y usa las técnicas de Boostrap Agreggation, que hace referencia al muestreo por reemplazo. (Machado & Karray, 2022)

- **Support Vector Machines (SVM) models**

  SVM es un clasificador lineal y su objetivo es encontrar el margen máximo sostenido por los vectores de soporte en un hiperplano separador, logrando así clasificar las etiquetas de los datos con base a sus características. En la figura se presenta la clasificación en el hiperplano:

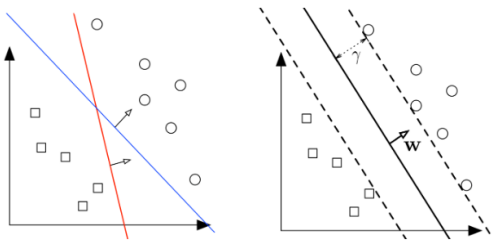


Figura 1: (Izquierda:) Dos hiperplanos de separación diferentes para el mismo conjunto de datos. (Derecha:) El hiperplano de margen máximo. El margen, γ, es la distancia desde el hiperplano a los puntos más cercanos de cualquiera de las clases (que tocan las líneas paralelas de puntos).

- **Artificial neural network (ANN) models**

  En la clasificación del score crediticio también se han usado las redes neuronales, en donde la red comienza pasando las características de cada cliente a la capa de entrada. Estas características son procesadas por las capas ocultas, y luego se llega a la capa de salida, que presenta la predicción final que se basa en los pesos. Estos últimos se definen para cada característica en función de su importancia relativa. Por último, usan la función sigmoidea como función de activación, la cual recoge todas las características ponderadas para producir las salidas. Este proceso se repite en varios bucles para reducir el error entre la clase predicha y la verdadera. (Machado & Karray, 2022)

Como conclusión, en los diversos estudios sobre esta temática, se han encontrado muchos experimentos, entre ellos, el uso de aprendizaje no supervisado para reclasificar la puntuación de crédito en la data histórica, para posteriormente correr los modelos de clasificación, encontrando así que las mejores métricas de precisión se las lleva el SVM y la aplicación de ANN–AdaBoost. Así mismo el enfoque en la selección de características como elemento clave para mejorar las precisiones en la predicción del score. (Koutanaeia, Sajedib, & Khanbabaeic, 2015)

Error de generalización y métricas de evaluación

Otro punto importante es el error de generalización o entrenamiento, el cual está compuesto por 3 elementos que permiten encontrar la mejor relación entre sesgo y varianza cuando se hace referencia a datos que el modelo no conoce. A continuación se presenta formalmente el error de generalización:

$$\underbrace{E_{\mathbf{x},y,D}\left[(h_D(\mathbf{x})-y)^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{\mathbf{x},D}\left[(h_D(\mathbf{x})-\bar{h}(\mathbf{x}))^2\right]}_{\text{Variance}} + \underbrace{E_{\mathbf{x},y}\left[(\bar{y}(\mathbf{x})-y)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))^2\right]}_{\text{Bias}^2}$$

Figura 2: función del error esperado.

La varianza está asociada a los datos tomados en el modelo, da cuenta del comportamiento de los datos al momento de entrenar con datos diferentes. El ruido es algo intrínseco de los datos y el sesgo está asociado a los modelos seleccionados. (Weinberger, 2018)

El error de generalización permite identificar qué acciones se deben tomar cuando los modelos tienen un mal comportamiento ante datos que no ha visto, es por ello

que la técnica de validación cruzada K-Fold, permite estimar este error dividendo el dataset en conjuntos de validación disjuntos.

Por último, las métricas para evaluar la precisión de las predicciones en los modelos de clasificación son:

- Matriz de confusión
- Accuracy
- Precision
- Recall
- F1-Score

3. Desarrollo metodológico CRISP-DM

3.1. Comprensión del negocio

3.1.1 Planteamiento del problema

Este proyecto utiliza como fuente de datos un reto actual de la plataforma Kaggle, el cual es llamado “Clasificación de puntuación de crédito”, en donde una compañía financiera a nivel mundial ha recolectado información bancaria básica de las personas, así como información relacionada con sus productos de crédito. (kaggle, 2022)

El objetivo principal es crear un modelo de aprendizaje automático supervisado que clasifique la puntuación crediticia de cada producto, para así, reducir tanto los esfuerzos manuales del personal como los riesgos a la hora de otorgar los créditos.

En los últimos años se ha incrementado el uso de algoritmos de aprendizaje automático en el sector financiero, en donde se han presentado diversas técnicas de Machine Learning como Random Forest, Support Vector Machine, Regresión Logística, en otros; con el objetivo de mitigar los riesgos crediticios que puedan presentar las compañías financieras al momento de aprobar o rechazar una solicitud de crédito.

La calidad de estos modelos de aprendizaje automático depende entre muchas cosas de los problemas de selección de la muestra, en donde diferentes autores recomiendan métodos de inferencia de rechazo estadístico, con el fin de subsanar el riesgo de rechazar o aprobar créditos de forma errónea. (Feng, Zhiyuan, Xingchao, & Dao, 2022)

Desde otra perspectiva es importante resaltar el uso de redes neuronales a través de las Maquinas de Aprendizaje Extremo (ELM), donde los autores Bequé y Lessmann realizan una comparación entre las técnicas de Machine

Learning de clasificación, como Random Forest, Support Vector Machine, KNN, Regresión Logística, entre otros; versus este tipo de red neuronal, proponiendo ANN; donde sus criterios de comparación fueron la facilidad de uso, complejidad computacional y precisión discriminativa. (Bequé & Lessmann, 2017)

De acuerdo al reto escogido en Kaggle y los documentos relacionados con respecto a este tipo de problemas, la pregunta de investigación está enfocada en evaluar las diferentes técnicas de clasificación de aprendizaje supervisado aplicadas a la clasificación del score crediticio, con el fin de crear un modelo predictivo que sirva de apoyo a la toma de decisiones de la entidad bancaria.

3.1.2. Objetivo General

Predecir el score crediticio de los productos de los clientes actuales de una entidad financiera usando técnicas de aprendizaje supervisado, con el fin de reducir los esfuerzos manuales y las apreciaciones subjetivas del personal, además de reducir los riesgos crediticios.

3.1.3. Objetivos específicos

- Realizar un análisis exploratorio de datos para comprender las variables y correlaciones que puedan existir entre ellas.
- Seleccionar las técnicas de aprendizaje supervisado más adecuadas para la clasificación del score crediticio.
- Evaluar los modelos de aprendizaje supervisado con las diferentes métricas de rendimiento acorde a los problemas de clasificación e identificar el mejor.

3.2. Entendimiento de datos

3.2.2. Recopilación de datos iniciales

El dataset tomado de la plataforma de Kaggle contiene datos bancarios básicos y mucha información crediticia de 12.500 clientes en una ventana temporal de 8 meses, por tanto, se tiene en total 100.000 registros crediticios (filas).
De cada registro se tienen 28 variables (columnas), en donde la variable objetivo “Y”, es “credit score”, una variable categórica de tres clases (Poor, Standard, Good).

3.2.3. Descripción de los datos

En esta sesión se hace una descripción del significado de cada variable, la cantidad de registros en cada una y la cantidad de variables categóricas y numéricas que tienen el dataset. Ver anexo 1.

Adicionalmente se hace una revisión del número y tipo de registros en cada variable, para así determinar que ingeniería de características se debe aplicar en cada variable.

A continuación, se mencionan algunas variables con datos incorrectos o vacíos que son tratados posteriormente, el detalle de toda la revisión de variables se encuentra en el notebook anexo:

- Monthly_Inhand_Salary: missing 15%
- Type_of_Loan: missing 11%
- Num_of_Delayed_Payment: missing 7%
- Payment_of_Min_Amount: missing 12%

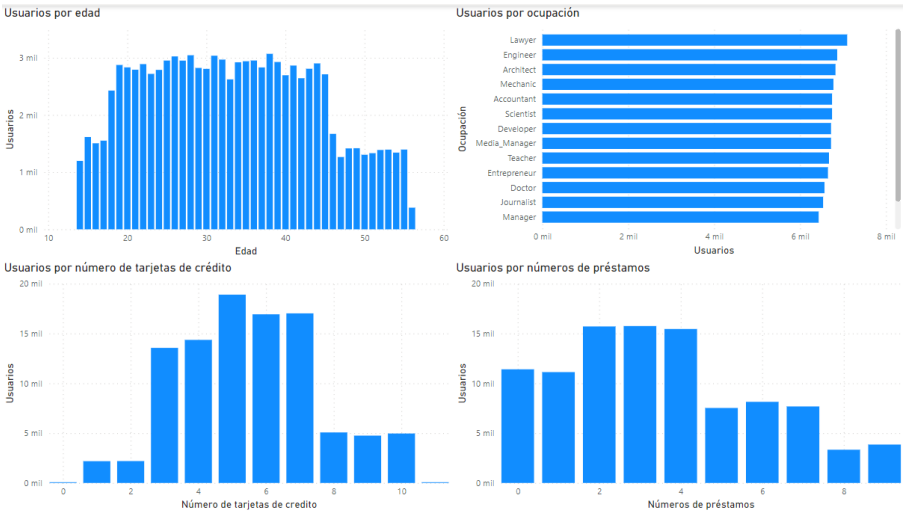
Por último, se identificaron las variables categóricas y numéricas de todo el conjunto de datos.

3.2.4. Exploración de datos

En la exploración de datos se realizaron tablas de frecuencia y gráficos de distribución para entender el comportamiento de las variables, identificar el balanceo de la etiqueta de salida, encontrar datos extraños y comprender valores, a continuación, se presenta la exploración de datos:

- **Distribución de la etiqueta de salida:**
  La categoría más frecuente es “Standard” que contiene el 53,2% de los registros, seguida por “Poor” (29,0%) y finalmente “Good” (17,9%). Con base en esto, los datos presentan desbalanceo que, aunque no es severo, se deberá probar si esto incide en la calidad de los modelos implementados.

- **Distribución de los usuarios por edad, ocupación, número de tarjetas de crédito y cantidad de préstamos:**



- **Comportamiento de los usuarios al momento del pago:**

```
Value_Payments_Behaviour
Large      25412
Medium     33639
Small      40949
Name: Value_Payments_Behaviour, dtype: int64
```

- **Comportamiento de los usuarios al momento de gastar:**

```
Spent_Behaviour
High     46056
Low      53944
Name: Spent_Behaviour, dtype: int64
```

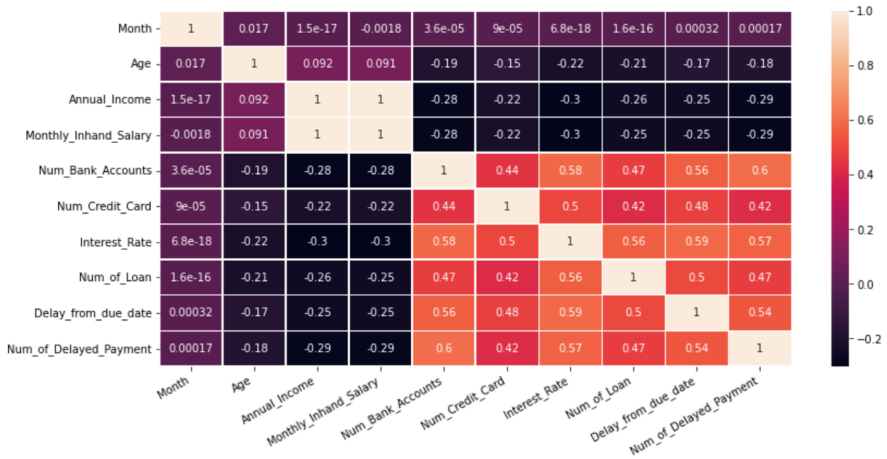
- **Tipos de préstamos:**

```
PaydayLoan                40568
Credit-BuilderLoan        40440
NotSpecified              39616
HomeEquityLoan            39104
StudentLoan               38968
MortgageLoan              38936
PersonalLoan              38888
DebtConsolidationLoan     38776
AutoLoan                  37992
NoData                    11408
dtype: int64
```

- **Análisis univariante:** se realizó el describe de todas las variables del dataframe, en el notebook anexo se encuentra el desarrollo completo:

	Age	Monthly_Inhand_Salary	Num_Bank_Accounts	Num_Credit_Card	Interest_Rate	Num_of_Loan	Delay_from_due_date
count	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05
mean	5.827583e-16	-2.302947e-16	-5.204726e-16	-1.224321e-15	-5.384804e-17	-6.434353e-15	-2.635253e-17
std	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00
min	-1.794405e+00	-1.221945e+00	-2.455343e+00	-2.676988e+00	-1.548065e+00	-1.444147e+00	-1.490243e+00
25%	-8.654482e-01	-8.067610e-01	-9.131897e-01	-7.418987e-01	-8.616672e-01	-6.266004e-01	-7.476566e-01
50%	-2.938663e-02	-3.456440e-01	-1.421132e-01	-2.581264e-01	-1.752694e-01	-2.178271e-01	-2.075940e-01
75%	8.066749e-01	5.524840e-01	6.289632e-01	7.094183e-01	6.255281e-01	5.997194e-01	4.674843e-01
max	2.107215e+00	3.454464e+00	2.171116e+00	2.644508e+00	2.227123e+00	2.234812e+00	2.762750e+00

- **Análisis multivariante:** se construyó un mapa de calor para ver la correlación lineal entre cada par de variables:

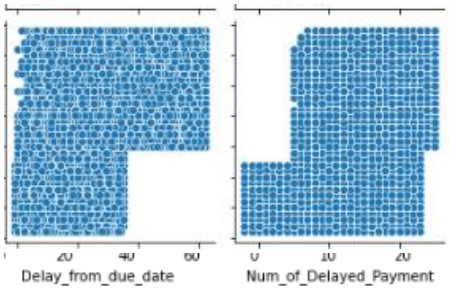


Encontrando las siguientes interpretaciones de correlación:

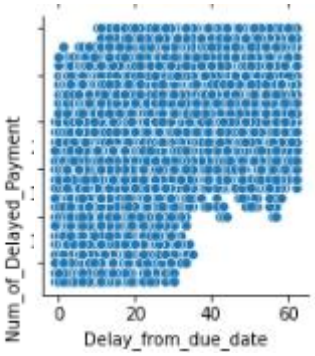
- Entre más cantidad de cuentas bancarias tiene una persona mayor será el número medio de pagos retrasados.
- A mayor cantidad de días promedio de retraso desde la fecha de pago mayor será la tasa de interés.
- Existe una correlación perfecta entre los ingresos mensuales y anuales; esto indica que una de estas variables no se debe tener en cuenta en el modelo, dado que sería redundante.

También se realizó la matriz de correlación para ver gráficamente como se comportaban los datos entre pares de variables, los más representativos fueron:

- Interest rate: acá se ve una correlación entre la tasa de interés con la variable que hace referencia a los días de retraso desde la fecha de pago y con el promedio de pagos retrasados por una persona.



- Num of delayed payment: existe una alta correlación entre el número medio de días de retraso desde la fecha de pago y número medio de pagos retrasados por una persona.



3.3. Preparación de datos

En esta etapa el objetivo es preparar los datos para adaptarlos a las técnicas de aprendizaje automático que se utilizaran en el modelado.

3.3.2. Selección de variables

El objetivo del problema es predecir la clasificación del score crediticio, por ello se transformó esa variable de tipo categórica a numérica:

```
y = df.Credit_Score.map(lambda x: 0 if x=="Poor" else x) \
                   .map(lambda x: 1 if x=="Standard" else x) \
                   .map(lambda x: 2 if x=="Good" else x)
y.unique()
```

```
array([2, 1, 0], dtype=int64)
```

Para el desarrollo de este proyecto se trabajará con todas las variables que entrega el dataset; en caso que se quiera pasar a producción, se podría desarrollar un modelo que elimine las variables que no tendría un cliente

nuevo, dado que asumir el riesgo de dejarlas todas puede generar ruido al modelo.
Por tanto, este modelo funcionaría perfecto para los clientes actuales que quieran solicitar un préstamo nuevo.

3.3.3. Limpieza de datos

Durante la exploración de datos se encontró que existían muchos datos atípicos, nulos, entre otros caracteres que requerían especial atención y tratamiento, en el notebook de limpieza anexo se encuentran los pasos de la limpieza de datos.

3.3.4. Ingeniería de características

Una vez realizados todos los procesos anteriores se identificó que había campos a los cuales se les debía realizar ingeniería de características, en el notebook anexo se encuentra detallado todo el proceso de ingeniería de características.

3.3.5. Tratamiento de alta dimensionalidad

Al momento de hacer toda la limpieza e ingeniería de características al dataset se encontró que se tiene una gran cantidad de variables con proceso de dumificación que proviene de variables numéricas y categóricas, por tanto, se requirió hacer el proceso de PCA, pero acotado al método Factor analysis of mixed data (FAMD) para que pudiese cumplir con esta condición. (Blaufuks, 2021)

El proceso de FAMD se aplicó a las variables categóricas, en el caso nuestro la ocupación, tipo de préstamo, comportamiento del gasto y pago, entre otras. El proceso consistió aplicar la raíz cuadrada a la probabilidad de cada columna y centrar los datos con la media.

También se realizó el StandardScaler de la librería de sklearn, con el fin de estandarizar las variables numéricas eliminando la media y escalando a la varianza unitaria. Luego se procedió a aplicar el PCA, colocando como parámetro que los componentes genere el 80% de la varianza explicada.

3.4. Modelado

3.4.2. Definición de datos de train y test

En este caso los datos son de tipo panel, dado que un mismo cliente se repite a lo largo del tiempo; es por ello que esta condición implica que el split para

los datos de entrenamiento y testeo se realice con un proceso llamado estratificación.

La división estratificada consiste en dividir los datos con una característica específica, en este caso el ID Cliente, garantizando así que un mismo individuo este con todos los registros completos en el train o test.
La librería usada para esta técnica es: GroupShuffleSplit de sklearn.model_selection:

```
gss = GroupShuffleSplit(n_splits=2, test_size=0.3, random_state=42)
for train_index, test_index in gss.split(X, groups=index):
    print("Train:", train_index, "Test:", test_index)
    X_train = X.iloc[train_index, :]
    y_train = y[train_index]
    X_test = X.iloc[test_index, :]
    y_test = y[test_index]
```

```
Train: [   16    17    18 ... 99965 99966 99967] Test: [    0     1     2 ... 99997 99998 99999]
Train: [    0     1     2 ... 99989 99990 99991] Test: [    8     9    10 ... 99997 99998 99999]
```

Con un conjunto de entrenamiento del 70% y testeo del 30%, adicionalmente se puso el random_state en 42 para fijar el punto de partida para esta división. Este proceso de split se realizó sin PCA y con PCA, con el fin de generar modelos con ambos procesos.

3.4.3. Generación de modelos y parámetros

En esta fase se determinó que los modelos a desarrollar en el proyecto serían Regresión Logística, Random Forest, XBoost y LGBM, siendo el primero el best line model como punto de referencia.

En los modelos la primera decisión que se tomo fue generarlos con el hiperparámetro de GridSearch, con el objetivo de encontrar la mejor combinación de hiperparámetros eficiente en cada modelo. Para la regresión logística toma la generalización L1 y L2 para prevenir el overfitting y la regularización. En métodos de ensamble se toma la profundidad de los árboles.

3.5. Evaluación

3.5.2. Métricas de evaluación

En este caso se van a evaluar los modelos con la métrica de precisión para la clase 0 - Poor, dado que el negocio tendrá mayor riesgo si presta dinero a alguien que salió etiqueta estándard o good, sabiendo que era poor. En este caso se castigan los verdaderos negativos.
Los resultados de la precisión en test para cada modelo son:

SIN PCA	
Modelo	**Precisión**
Regresión Logística	0,66
Random Forest	0,70
LGBM	0,76
XGBoost	0,63

CON PCA	
Modelo	**Precisión**
Regresión Logística	**0.70**
Random Forest	**0,68**
LGBM	**0,70**
XGBoost	**0,60**

3.5.3. Cross-validation

La técnica del cross validation se realizó para estimar el error de generalización o test, para ello se toman los datos de entrenamiento dividiéndolos en 5 particiones, en este caso la métrica usada fue F1-score, aplicando el F1-Macro de sklearn, dado que el problema enfrentado es de clasificación multiclase.
El cross validation con mayor % en los modelos sin pca fue el arrojado con XGBoost y con respecto a los modelos generados con pca, los resultados de CV fueron más bajos.

4. Análisis

El objetivo del proyecto es un problema de clasificación del score crediticio en 3 clases, Poor, Standard and Good, para lograrlo se realizó todo el proceso de exploración de datos e ingeniería de características para aplicar los modelos de Regresión Logística, Random Forest, XGBoost y LGBM.
Estos modelos se crearon desde dos puntos de vista, el primero tomando todas las variables y el segundo aplicando el proceso de PCA con el método de FAMD. Es importante mencionar que 46 variables resultantes después de la ingeniería de

características no se considera un problema de dimensionalidad, por ello se tomó la decisión de correr los modelos de las dos formas.
En este caso el mejor desempeño medido desde la métrica de precisión enfocado a la clase 0, fue para el modelo LGBM con todas las variables.

Al momento de correr todos los modelos el XGBoost fue el más demorado en tiempo, por lo que este modelo aprende por el descenso del gradiente y sus resultados se van acumulando, generando así el ensamble de modelos. Por otro lado, el LGBM fue más rápido y preciso porque cuando se genera un árbol se expande por las hojas que menor pérdida generan, por tanto, no generan más nodos.

5. Conclusiones

- Las técnicas usadas para el tratamiento de datos disminuyen la probabilidad de pérdida de información, por tanto, los métodos abordados son una forma de reconstrucción de información.
- Aplicar técnicas de reducción de dimensionalidad en matrices con pocas características es innecesario, ya que genera perdida de interpretabilidad y no mejoro de manera considerable el rendimiento de los modelos.
- Con la generación de todos los modelos se comprueba que el Grid Search si mejora la precisión ya que permite entre otros controlar el sobreajuste de los modelos, especialmente Random Forest al evaluar distintas profundidades de los árboles.
- Las técnicas de Gradient Boosting fueron más precisas que las técnicas lineales y el Random Forest tradicional.
- LGBM es mejor que XGBoost por ser más rápido y con una mejora precisión, además de tener menos costo computacional.
- Los modelos generados son para clientes que ya tienen historial crediticio en la entidad financiera.
- Aplicar redes neuronales con el volumen de datos actual es costoso computacionalmente y la precisión no mejora sustancialmente.

6. Implicaciones éticas

Esta herramienta tiene implicaciones importantes tanto para la empresa como para los clientes. Es importante analizar las repercusiones de esta tecnología en dos escenarios: en el de éxito, en el cual el modelo segrega correctamente a las personas en sus puntajes crediticios correspondientes, y en el que no.

En caso de éxito, esto implicaría reducción considerable en el tiempo y esfuerzo que deba invertir la empresa para analizar las solicitudes de créditos, lo que se traduce tanto en menos pérdidas para la compañía como una atención más ágil para los solicitantes.

Por otro lado, en caso de fallo del modelo en producción, las implicaciones pueden ser más serias. Se pueden observar diferentes fallos para ver sus repercusiones. Si el modelo tiende a dar puntajes altos a personas que en realidad tendrían un score bajo, la empresa va a sufrir pérdidas graves. Similarmente, si el modelo clasifica con un puntaje bajo a una persona que en realidad tendría uno alto, la persona no va a poder a acceder a mejores créditos y su reputación ante la empresa se vería deteriorada.

Por último, el aspecto más importante en relación con la ética es que el modelo no tenga sesgos sobre características como raza, género y nacionalidad, porque, en caso de estar sesgado, estaría dando puntajes injustos sobre características que no tienen relevancia a la hora de evaluar a un individuo, convirtiéndose la herramienta en un detrimento para la igualdad y agravando problemas existentes como el racismo, la xenofobia y el machismo.

7. Aspectos legales y comerciales

El potencial comercial de este proyecto para las entidades financieras reside en la posibilidad de diseñar un algoritmo que comprenda las variables claves que se deben tener para generar una clasificación precisa del score crediticio, generando así, mejoras en procesos manuales de asignación del score, un mayor retorno de los créditos otorgados, entre otros beneficios para las entidades financieras que usen inteligencia artificial en este proceso. Es por ello que este proyecto merece la pena desarrollarse.

En este sentido los aspectos legales que se deben considerar al momento de exponer los resultados son los acuerdos de confiabilidad y tratamiento de información de los clientes.

En esta investigación en particular los datos son públicos, por tanto no hay acuerdos de confidencialidad y los resultados que se obtengan serán de código abierto.

8. Bibliografía

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications*, 42-53.

Blaufuks, W. (25 de 05 de 2021). *Towards Data Science*. Obtenido de https://towardsdatascience.com/famd-how-to-generalize-pca-to-categorical-and-numerical-data-2ddbeb2b9210

Feng, S., Zhiyuan, Y., Xingchao, Z., & Dao, L. (2022). Reject inference in credit scoring using a three-way decision and safe semi-supervised support vector machine. *Information Sciences*, 614-627.

IBM. (17 de 08 de 2021). *IBM*. Obtenido de https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

kaggle. (2022). *kaggle*. Obtenido de https://www.kaggle.com/datasets/parisrohan/credit-score-classification

Koutanaeia, F. N., Sajedib, H., & Khanbabaeic, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, 11-23.

Machado, M. R., & Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert Systems with Applications*, 1-12.

Masui, T. (01 de 2020). *Towards Data Science.* Obtenido de Towards Data Science: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

Weinberger, K. (2018). *Machine Learning for Intelligent Systems*. Obtenido de https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote09.html

9. Anexos

Anexo 1. Descripción de variables

Anexo 2. Limpieza de datos

Anexo 3. Ingeniería de características

Anexo 4. Exploración de datos

Anexo 5. Modelos sin PCA

Anexo 6. Modelos (RF,XGBoost) con PCA

Anexo 7. Modelos (LGBM) con PCA