Aprendizaje maquina para determinar crédito otorgados a personas físicas y su rentabilidad

Lic. Aranza Alejandra Esteban Avalos

20 de noviembre de 2023

1. Abstract

La capacidad de predecir con precisión los montos de préstamos representa una ventaja tanto para instituciones financieras como para clientes. utilizando herramientas como aprendizaje máquina, se desarrollan análisis con grandes conjuntos de datos para revelar patrones y tendencias en el comportamiento financiero. Este estudio se sumerge aplicando técnicas avanzadas de aprendizaje máquina, como Árboles de Decisión y Regresión Lineal, para predecir los montos de los préstamos en base a ciertas características del conjunto de datos.

Nuestro enfoque se centra en obtener mejor visibilidad a los préstamos personales para en futuro identificar créditos empresariales. Se utiliza un conjunto de datos que contiene información desde los ingresos anuales de los solicitantes hasta sus historiales de crédito. Mediante la implementación de de dichos modelos, buscamos mejorar la eficiencia en la toma de decisiones de préstamos.

A través de este estudio, se busca dar mayor visibilidad al análisis de datos, cuando se combina con herramientas de aprendizaje maquina, puede convertirse en un punto crítico para la toma decisiones más informadas y estratégicas en el ámbito financiero.

2. Introduccion

El análisis se centra en los créditos otorgados a personas físicas, con un enfoque específico en la predicción del monto de préstamos y adicional a esto identificar patrones o las diferentes relaciones entre las variables. Se cuenta con una data detallada que incluye variables como el monto del préstamo, la tasa de interés, el ingreso anual del acreditado, el historial de pagos, entre otras, se buscan los patrones y factores clave que influyen en la decisión de otorgar un préstamo y adicional su potencial rentabilidad.

El objetivo es primero, identificar las características de los préstamos y de los acreditados que están más fuertemente asociadas con la aprobación de montos de préstamo adicionales; y en un segundo plano, evaluar la rentabilidad de estos préstamos adicionales a través de un análisis que integra diversas variables

financieras y personales. Se emplearán técnicas avanzadas de análisis de datos y modelado predicativo, incluyendo regresión y clasificación, para generar insights que puedan orientar las estrategias de préstamos

3. Descripción de los datos

El conjunto de datos utilizado en este estudio representa un compendio detallado de préstamos personales, abarcando múltiples aspectos financieros y personales de los solicitantes. Cada entrada en este conjunto de datos incluye variables como el monto del préstamo, la tasa de interés, el plazo del préstamo, los ingresos anuales del solicitante, el historial de crédito, entre otros. Estas variables ofrecen una visión integral de las circunstancias financieras y el comportamiento crediticio de los individuos, lo que las hace idóneas para el análisis mediante aprendizaje máquina.

	count	mean	std	min	25%	50%	75%	max
loan_amnt	10101.0	12914.283239	8497.637542	1000.00	6000.00	11400.00	17600.00	35000.00
term	10101.0	44.057024	11.334269	36.00	36.00	36.00	60.00	60.00
int_rate	10101.0	12.446663	4.234000	5.42	8.90	12.42	15.27	24.11
installment	10101.0	365.788152	227.094175	21.74	195.27	323.48	483.38	1288.10
grade	10101.0	1.412236	0.699324	0.00	1.00	2.00	2.00	2.00
sub_grade	10101.0	9.571627	7.000548	0.00	4.00	8.00	14.00	34.00
emp_length	10101.0	5.398970	3.537944	0.00	2.00	5.00	10.00	10.00
annual_inc	10101.0	71080.476980	48185.734264	6000.00	43200.00	60000.00	85000.00	1782000.00
purpose	10101.0	8.812593	5.254307	0.00	2.00	13.00	13.00	13.00
addr_state	10101.0	14.629047	14.312617	3.00	4.00	4.00	27.00	45.00
earliest_cr_line	10101.0	9922.499554	2428.795560	-2555.00	8770.00	10238.00	11688.00	13889.00
open_acc	10101.0	9.509752	4.253482	2.00	6.00	9.00	12.00	34.00
total_pymnt_inv	10101.0	13831.297683	10102.759198	96.96	6133.00	11360.34	18616.80	56475.05
total_rec_int	10101.0	2843.840275	3249.521484	11.64	745.01	1607.72	3672.00	22930.84
last_pymnt_d	10101.0	15942.656668	412.344892	14983.00	15709.00	16076.00	16082.00	16801.00
last_pymnt_amnt	10101.0	3380.403727	5379.236981	0.01	281.46	665.57	4382.62	36115.20
last_credit_pull_d	10101.0	16331.686170	472.132996	14981.00	16076.00	16441.00	16801.00	16801.00
default_ind	10101.0	6353.534700	2480.379726	1338.00	4657.00	5934.00	7639.00	18750.00
sub_grade_encoded	10101.0	9.571627	7.000548	0.00	4.00	8.00	14.00	34.00
credit_history	10101.0	6353.534700	2480.379726	1338.00	4657.00	5934.00	7639.00	18750.00

Figura 1: Descripción de variables

La limpieza de datos es un paso crucial en cualquier análisis estadístico y de aprendizaje máquina. En nuestro estudio, este proceso implicó varias fases:

Se identificaron y manejaron los valores faltantes en el conjunto de datos para asegurar la integridad del análisis.

Se revisaron y corrigieron posibles errores e inconsistencias, como valores atípicos o incoherentes.

Los datos se normalizaron para homogeneizar las escalas de las diferentes variables y se realizaron transformaciones necesarias para mejorar la calidad y la relevancia del análisis.

4. Metodología

Pruebas de Hipótesis y Selección de Modelos

Realizamos pruebas de hipótesis utilizando técnicas como RidgeCV, VarianceThreshold y fregression para determinar la relevancia de las diferentes variables. Estos métodos nos ayudaron a identificar las variables más significativas y a refinar nuestro conjunto de datos.

Uso de K-Means para Análisis Exploratorio Como parte de nuestro análisis exploratorio, aplicamos el algoritmo K-Means para identificar agrupaciones naturales dentro del conjunto de datos. Este enfoque nos permitió descubrir patrones subyacentes y relaciones entre las variables, informando nuestra selección y aplicación de modelos predictivos.

'Installment' identificado como un predictor importante, indicando que la estructura de los pagos del préstamo es muy importante para la predicción. 'Term' y 'loanVSfunded' también destacados como relevantes. 'Annual inc' mostró una relación no lineal o compleja, posiblemente debido a la influencia de la regularización.

Con el análisis se puede interpretar que ciertas características financieras son más influyentes en la rentabilidad de los préstamos, con una atención particular en la estructura de pagos y términos del préstamo.

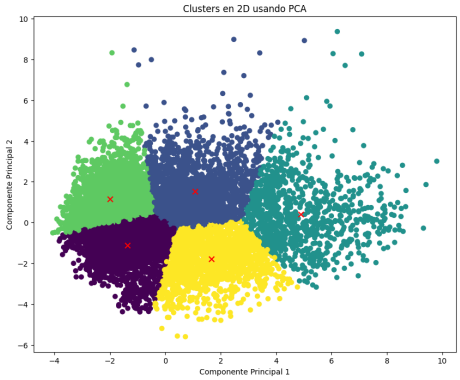

Figura 2: Clustering

Modelado y Evaluación

Seleccionamos Árboles de Decisión y Regresión Lineal como nuestros modelos principales:

Árbol de Decisión

Un arbol de decisión se representa como una serie de divisiones basadas en las caracteristicas. Cada nodo del árbol representa una característica, cada rama representa una decisión y cada hoja representa un resultado. La fórmula general de un árbol de decisión no es tan directa pero la idea básica se puede expresar como una serie de decisiones condicionales

Si $(x_i \leq umbral)$ entonces ir a la rama izquierda, sino ir a la rama derecha

Como resultado obtuve que se capturó relaciones complejas, aunque con un riesgo de sobreajuste.

Regresión Lineal

La regresión es un enfoque estadistico para modelar la relación entre una variable dependiente en este caso, la rentabilidad del préstamo y una o más variables independientes

Proporcionó un enfoque más simplificado y generalizable. Ambos modelos fueron evaluados usando métricas como MAE, RMSE y R^2 para comparar su eficacia y precisión.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

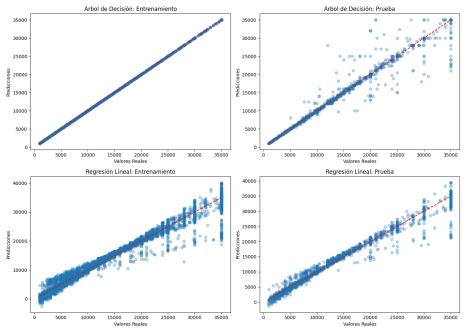

Figura 3: Arboles de decisión vs Regresion

Se comparó la Regresión Lineal con Árboles de Decisión donde los arboles de decisión muestran evidencia de sobreajuste, con un alto rendimiento en el conjunto de entrenamiento y una disminución en los datos de prueba. El sobreajuste es una preocupación para los modelos de árbol. En cambio las regresión lineal mostró un comportamiento más equilibrado y consistente, con una buena generalización entre los conjuntos de entrenamiento y prueba.

5. Exploracion de resultados

El Árbol de Decisión y la Regresión Lineal, se revelaron modelos distintos. La predicción de los montos de préstamos. Cada uno nos contó una historia diferente sobre los patrones y relaciones en nuestros datos.

El Árbol de Decisión, con un MAE y RMSE de cero y un R^2 de 1.0, inicialmente pareció el modelo perfecto, describiendo los datos con una precisión impecable. Sin embargo, este nivel de perfección sugiere un posible sobreajuste.

Por otro lado, la Regresión Lineal muestra una visión más equilibrada. Con un MAE de 848.55, un RMSE de 1571.36 y un R^2 de 0.9656. Aunque no cubrio cada detalle con la misma precisión del Árbol de Decisión.

El modelo de Árbol de Decisión mostró un rendimiento perfecto en el conjunto de entrenamiento, lo cual puede indicar un sobreajuste, donde el modelo se adapta demasiado bien a los datos de entrenamiento, pero podría no generalizar bien a datos nuevos. Por otro lado, el modelo de Regresión Lineal, aunque menos preciso, ofrece resultados más realistas para una situación práctica, con un R^2 muy alto, lo que indica que el modelo puede explicar una gran proporción de la variabilidad en el monto del préstamo.

Esta exploración de los préstamos personales a través del aprendizaje máquina es como estar llena de incertidumbres y posibilidades. Aprender sobre la importancia de equilibrar la precisión y la generalización, la teoría y la práctica. En última instancia, aunque las herramientas y técnicas pueden ser poderosas, su verdadero valor reside en cómo las aplicamos y en las historias que elegimos contar con ellas.

6. Conclusiones

Este estudio ilustra cómo diferentes modelos de aprendizaje máquina pueden ser utilizados para predecir montos de préstamos. Mientras que el Árbol de Decisión muestra una precisión teóricamente perfecta en el conjunto de entrenamiento, es probable que sufra de sobreajuste. En cambio, la Regresión Lineal, aunque no tan precisa, proporciona un balance más realista entre precisión y capacidad de generalización.

[6] [7] [5] [3] [1] [2] [4]

Referencias

[1] Alberto Benavides. *Aprendizaje No Supervisado*. Accedido el: [fecha de acceso]. 2023. URL: https://github.com/albertobenavides/aprendizaje_autom/blob/master/caps/4_no_supervisado.ipynb.

[2] Alberto Benavides. *Aprendizaje Supervisado*. Accedido el: [fecha de acceso]. 2023. URL: https://github.com/albertobenavides/aprendizaje_autom/blob/master/caps/5_supervisado.ipynb.

[3] Alberto Benavides. *Métricas de Desempeño - Aprendizaje Automático*. Accedido el: [fecha de acceso]. 2023. URL: https://github.com/albertobenavides/aprendizaje_autom/blob/master/caps/7_metricas_desempeno.ipynb.

[4] Alberto Benavides. *Selección de Características*. Accedido el: [fecha de acceso]. 2023. URL: https://github.com/albertobenavides/aprendizaje_autom/blob/master/caps/3_sel_carac.ipynb.

[5] scikit-learn. *Clustering*. Accedido el: [fecha de acceso]. 2023. URL: https://scikit-learn.org/stable/modules/clustering.html.

[6] scikit-learn. *Comparing Successive Halving and Random Search*. Accedido el: [fecha de acceso]. 2023. URL: https://scikit-learn.org/stable/auto_examples/model_selection/plot_successive_halving_heatmap.html.

[7] scikit-learn. *Supervised Learning*. Accedido el: [fecha de acceso]. 2023. URL: https://scikit-learn.org/stable/supervised_learning.html.