**Trabajo 01: Modelos de riesgo de crédito**

**Integrantes**

Camilo Andres Granada Mejia

Alejandro Zapata Quintero

**Introduccion**

En un sector economico con tantas extensiones como lo es el de las finanzas, es natural esperar que con el intensivo flujo de informacion hayan errores y contratiempos en el sistema, que pueden porvocar fallas que vayan desde simples errores de diccion hasta el colapso total de las diferentes funciones que componen la maquinaria empresarial. No se habla solamente de errores a nivel logistico o errores de diseño sino tambien de errores de planeación que pueden verse desde la estructuracion de los diferentes proyectos de una empresa hasta la selección de un público objetivo.

En este caso se hará acercamiento a un modelo crediticio en el que aun con una probabilidad de error significativa bajo ciertas regulaciones las ganancias podrían estar garantizadas. Para eso se hará uso de datos historicos recopilados de sistemas que han probado ser funcionales.

**Desarrollo**

**1) Limpieza de la base de datos**

Antes de hacer cualquier tipo de análisis en los datos es necesario revisar su distribucion en la base de datos para evitar que algun tipo de anomalía afecte la consistencia del modelo.

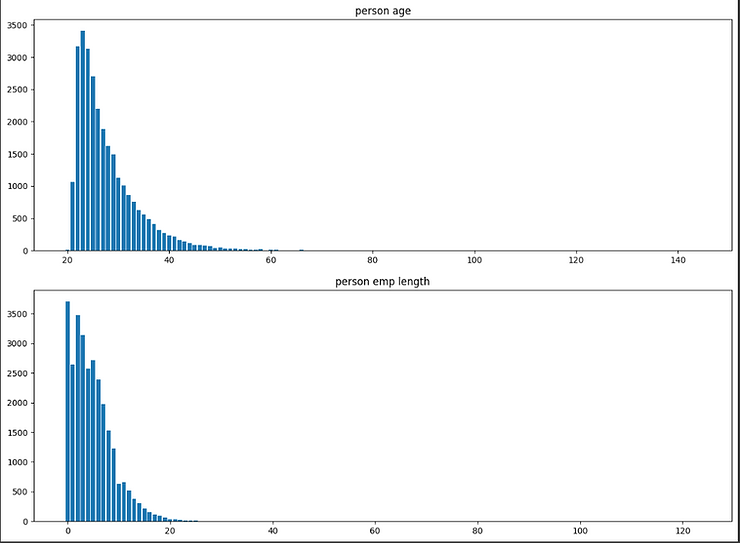
A continuación se presenta una vision general del set de datos.



**Figura 1:** Comportamiento de la informacion

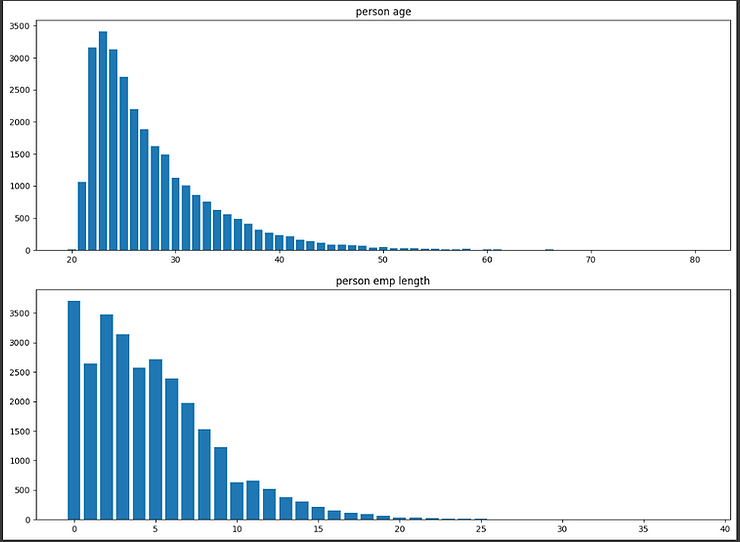
Si bien el comportamiento de los datos está entre lo esperado, se puede apreciar que hay inconsistencias en la cantidad de datos en las diferentes variables, esto se debe a la presencia de datos nulos (o lo que es lo mismo, la ausencia de información) y para mantener una estructura limpia y consistente se elige eliminar todas las filas afectadas por este evento.

Adicionalmente se aprecian valores atípicos en las variables que representan la edad y la longitud del prestamo sobrepasando los montos razonables y aunque es claro que deben eliminarse lo mejor es revisar primero la distribucion de los datos en ambas variables.



**Figura 2:** Distribucion de datos original

Es claro que los datos mencionados son en efecto atípicos respecto al comportamiento natural del resto de la informacion, viendolo en detalle se encuentra que la edad presenta un tope en los ochenta años y la longitud del prestamos al rededor de treinta y ocho, por lo tanto se decide limitar la edad a ochenta y la longitud del prestamos a cuarenta, como consecuencia se obtiene lo siguiente:



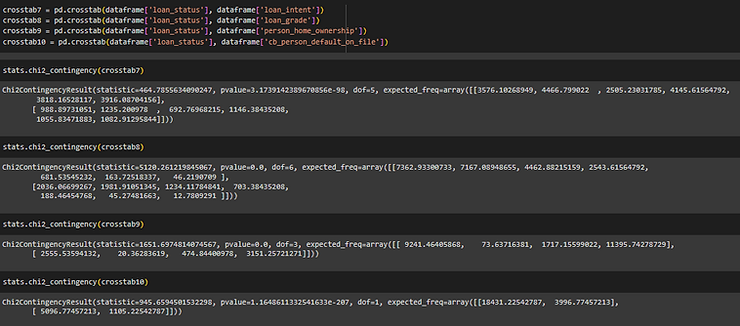
**Figura 3:** Distribucion de datos limitada

Ahora el comportamiento de los datos es razonable.

**2) Análisis de las variables**

Con la base de datos limpia se puede por fin hacer consideraciones sobre la forma en la que se distribuyen los datos. Aunque no parece apropiado se dará primero una revision a las variables categóricas del set de datos. Para ello se usará la prueba de chi2 estadística que comparará la variable objetivo "loan\_status" con las demás.

Cabe mencionar que "loan\_status" es una variable categorica con categorías numéricas, por lo que es facil equivocarse al hacer un acercamiento a los datos pero a su vez facilita la revisión de los mismos por medios estadísticos gracias a su naturaleza.

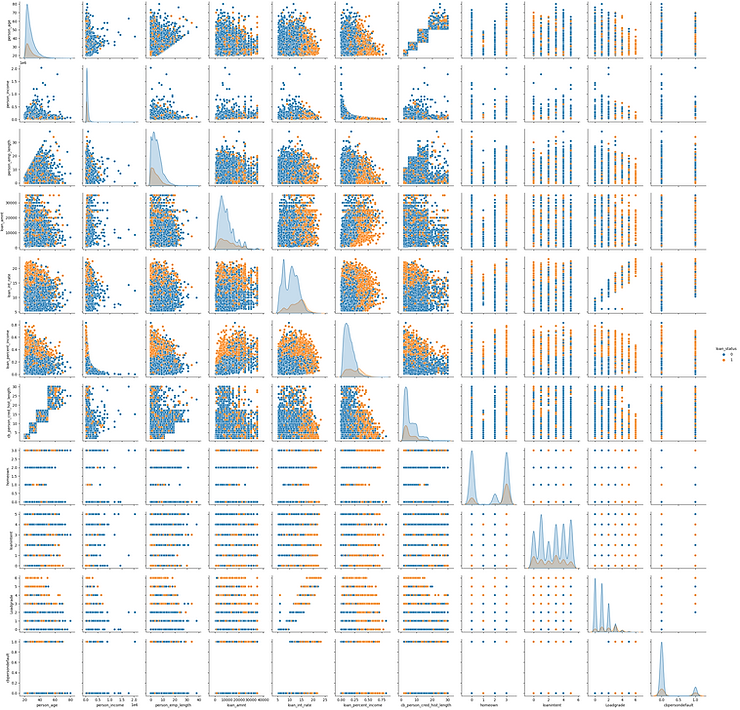


**Figura 4:** Prueba Chi2

De inmediato puede verse que los valores p obtenidos son extremadamente pequeños o directamente cero por lo tanto no es necesario descartar variables categóricas (solo desde este criterio).

Seguido a ello se decide realizar una prueba Anova a las variables numéricas, los resultados obtenidos son en efecto muy similares a los de la prueba chi2 independientemente del orden o la cantidad de variables analizadas (dos como mínimo), esto sugiere que las variables numéricas pueden usarse libremente pero no parece apropiado por lo que se elige ignorar esta prueba.

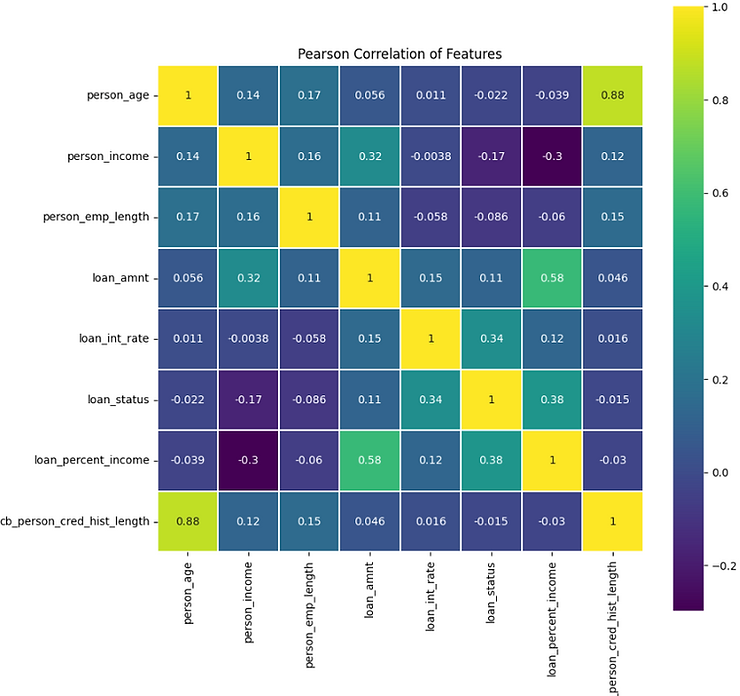
Antes de tomar la decision sobre cuales variables se escogen y el motivo, es interesante observar la distribucion de las personas que cumplen con los pagos de su credito y las que no respecto a las demás variables. Para ello se hace uso de diagramas de pares.



**Figura 5:** Diagramas de Pares

Inmediatamente puede verse cierta dominancia de los datos azules (estado cero/no paga) sobre los datos naranja (estado uno/paga), lo cual implica que para que el modelo funcione aun siendo una minoría, las personas que cumplen con sus obligaciones crediticias deben compensar en gran medida las deficiencias de la otra parte de la población.

Finalmente se realiza una matriz de correlación sobre las variables numéricas para tomar decisiones.



**Figura 6:** Matriz de correlacion

De inmediato puede verse la relevancia y prevalencia de las variables "loan\_int\_rate" y "loan\_percent\_income". En consecuencia debido a lo visto anteriormente se elige tomar estas variables y las categóricas en su totalidad.

**3) Desarrollo del modelo**

Luego de analizar la distribucion de los datos se procede a programar el modelo.

En primer lugar las variables que representan la edad, los ingresos y el monto prestado se convierten a categorías usando intervalos de valores.

Para la edad se decide usar los siguientes intervalos como categorías: '20-25', '26-35', '36-45', '46-55', '56-65'.

Para los ingresos se usan 'Bajo', 'Medio-Bajo', 'Medio', 'Medio-Alto', 'Alto' como categorías con 0, 25000, 50000, 75000, 100000 representando los límites de los intervalos respectivamente. De manera similar se usan las categorías 'Pequeño', 'Mediano', 'Grande', 'Muy Grande' para el monto prestado con los limites respectivos 0, 5000, 10000, 15000.

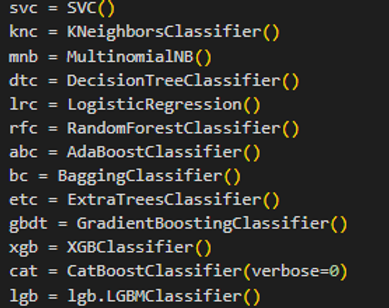
Seguido a ello se calculan las tasas prestamo-ingresos, prestamo-tiempo de empleo y tasa de interes-prestamo, pues es necesario tener una relacion matemática de estas variables para ser usadas posteriormente en el modelo.

Se hace una nueva reorganizacion de los datos solo para asegurar que la consistencia de la base de datos se mantiene pues es relativamente sencillo violentar la distribucion de la informacion, esto se hace solo como un paso preventivo.

Se crean nuevos sets de datos para entrenamiento y testeo del modelo nombrados x\_train y x\_test respectivamente y junto a ellos los modelos y\_train y y\_test que contienen la informacion objetivo relacionada a sus contrapartes x\_train y x\_test, esto implica que la columna de la variable objetivo no esta presente en ninguno de los sets de datos nombrados por x.

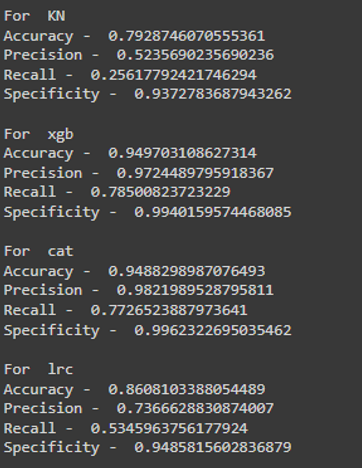
Usando escaladores de preprocesamiento datos se reorganiza la informacion en los nuevos sets solo para darle un formato entendible por el modelo de machine learning que se usará más adelante.

Se importan múltiples regresiones para evaluar los datos, pero primero se revisa el comportamiento de dichas regresiones con respecto a los datos. Las regresiones seleccionadas son las siguientes:



**Figura 7:** Regresiones a evaluar

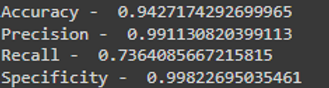
Se hace una evaluación en grupos de cuatro y de acuerdo a los resultados obtenidos las mejores son las siguientes:



**Figura 8:** Mejores regresiones

Se toman las tres mejores y se analizan los datos con ellas, luego se aplica la funcion de arboles de decisión aleatorios para encontrar los coeficientes de las categorias presentadas en los datos, incluyendo las categorías creadas para variables como la edad y los ingresos.

Este último proceso presenta los siguientes resultados.

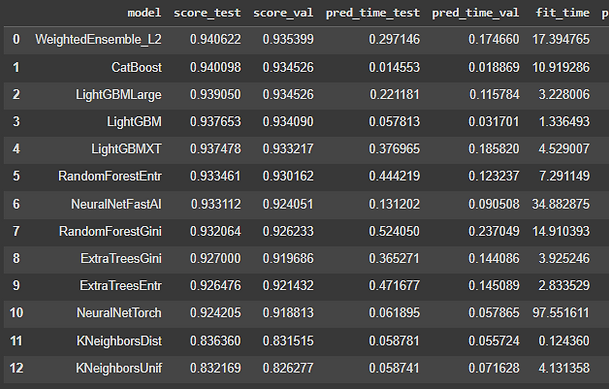


**Figura 9:** Scorecard del árbol de decisiones



**Figura 10:** Coeficientes de las categorias

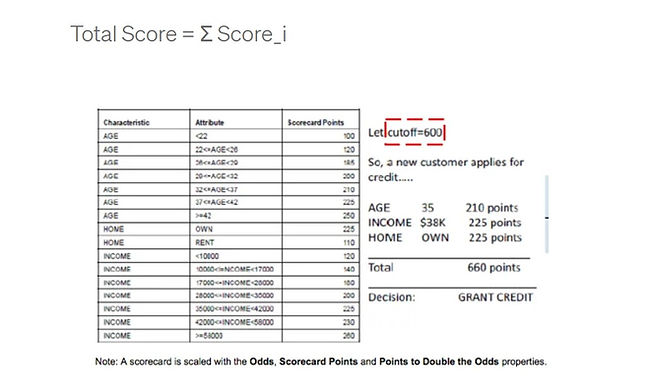
Finalmente se importan las funciones de machine learning del modelo Autogluon ML que permite hacer predicciones en los datos para comportamientos a futuro, no se hará uso inmediato de los resultados pero los resultados de los tests son prometedores, lo que confirma la consistencia del modelo.



**Figura 11:** Scorecard de Autogluon

**4) Asignacion del puntaje**

El puntaje final asignado a las diferentes personas respecto a los datos presentados en el set, incluyendo las categorías, se calcula usando una funcion simple dominada por la siguiente tabla.



**Figura 12:** Scorecard de referencia

**5) Conclusiones**

* Fue posible desarrollar el modelo usando multiples regresiones como referente, para después pulír el resultado usando regresiones de la teoría de análisis de decisiones.
* La consistencia de los datos se mantiene durante todo el desarrollo gracias a la insistencia con la organizacion y el mantenimiento del set.
* Aunque no fue considerado durante el desarrollo, en los tests preliminares se pudo apreciar una diferencia significativa en la correlacion de algunos datos bajo la presencia de nulos, por ello su eliminación fue necesaria.

**Referencias**

**[1]**Kaggle, 2020. Credit Risk Dataset: Dominio Público. Disponible en: https://www.kaggle.com/datasets/laotse/credit-risk-dataset

**[2]**Kaggle, 2023, Credit Risk Prediction. Disponible en: <https://www.kaggle.com/code/anshtanwar/credit-risk-prediction-training-and-eda#11.-Auto-ML-using-AutoGluon>

**[3]**Tina, 2019. Intro to Credit Scorecard. Disponible en: <https://towardsdatascience.com/intro-to-credit-scorecard-9afeaaa3725f>