German Credit Risk

Participantes:

Gandini, Tiago

Godoy, Tomás

Indice

Nuestro Caso:

Nuestro proyecto consiste en saber si los candidatos que tenemos en nuestro dataset son buenos prospectos para recibir un credito bancario, en nuestro caso más especificamente hablamos de bancos alemanes.

Diccionario de variables:

1: Age (Numerica)

2: Sex (Texto: male o female)

3: Job (Numerica: 0 – sin habilidad y no residente, 1 - sin habilidad y residente, 2 – con habilidad, 3 – alta habilidad)

4: Housing (Texto: own, rent o free)

5: Saving accounts (Texto - little, moderate, quite rich o rich)

6: Checking account (Numerica: en marcos alemanes)

7: Credit amount (Numerica: en marcos alemanes)

8: Duration (Numerica: en meses)

9: Purpose (Texto: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)

10: Risk (Texto: good o bad)

Tabla de versionado:

1: Dataset original

2: Dataset sin la columna „unname: 0“ (generada automaticamente, es una copia del indice) ni la columna „Checking account“, ya que por la cantidad de nulos que tenia, al rededor de un 40%, no nos parecio conveniente dejarla, además luego de sacar esa columna se aplicó el metodo dropna para sacar los pocos nulos que quedaban y por ultimo se reinició el indice.

Objetivo del modelo:

Nuestro objetivo es poder determinar en base a nuestro dataset, con diversos algoritmos de machine learning, si una

persona/candidato representa un riesgo alto o bajo para el banco a la hora de otorgarle un prestamo.

Datos importantes encontrados por el EDA:

1:La mayoria de los solicitantes tienen una edad de entre 25 y 40 años.

2:El riesgo y la cantidad del monto estan directamente correlacionados.

3:Parece ser que a mayor edad tenga el solicitante, más seguro es que vaya a pagar el credito (Probablemente porque ya tendrán altas capacidades laborales y propiedades).

4:Hay una pequeña tendencia femenina a pedir mayores montos en los creditos.

5:Los que rentan tienden a no sacar prestamos para vacaciones/ otros.

6:Parece haber una relación inversa entre pagar alquiler y sacar prestamos(suponiendo que los propietarios lo fueran al 100% y n o que tuvieran hipotecas) para cosas relacionadas al ocio.

7)Es visible como los hombres de todas las edades sacan credito s preferentemente para autos, muy por encima de la cantidad de mujeres.

8)Se puede observar que a partir de los 50 no se sacan tantos creditos para radio/TV.

9)Se puede observar que los hombres jovenes (20-35) sacan más creditos con fines de negocios.

Algoritmos Elegidos:

Los algoritmos que elegimos fueron arbol de decisión, bosque aleatoreo, knn, regresión logistica y svm.

Metricas de desempeño:

	Test	Training	Dif
Arbol	0.695122	0.735552	-0.040430
RF	0.719512	1.000000	-0.280488
kNN	0.621951	0.786340	-0.164389
LogReg	0.727642	0.712785	0.014858
SVM	0.711382	0.674256	0.037126

Iteraciones de optimización:

Gridsearch: 60 iteraciones en RF y 180 en KNN

Randomsearch: 100 iteraciones en KNN

Métricas finales del Modelo Optimizado:

	Test	Training	Dif
Arbol	0.695122	0.735552	-0.040430
RF	0.719512	1.000000	-0.280488
kNN	0.621951	0.786340	-0.164389
LogReg	0.727642	0.712785	0.014858
SVM	0.711382	0.674256	0.037126
RF_Tuneado	0.752033	0.761821	-0.009789
KNN_Tuneado	0.686992	0.709282	-0.022290

Futuras lineas:

Algo ideal, a nivel de escalibilidad, para nuestro proyecto sería poder utilizarlo con datos más actuales e incluso quizá con datos en tiempo real, ya que por lo que hemos visto tiene un buen desempeño a la hora de determinar si al banco le conviene

otorgar los prestamos. Algo aún más ambicioso sería incluso poder llegar a convertir nuestro modelo en una app para los bancos, que ellos puedan correr en sus propios servidores, ya que, a priori, no es tan complejo como podrían llegar a serlo otros modelos, por lo cual el determinar si conviene o no otorgar el prestamo podría hacerse de manera relativamente rapida en lugar de tardar horas/dias como llega a suceder con otros tipos de datasets.

Conclusiones:

Como hemos podido observar, existen muchas tendencias que varian mayormente según género y edad. Por un lado hay una ligera tendencia femenina a pedir prestamos más altos y además, entre los 20 y los 25 años de edad, estos prestamos suelen ser para amoblar, mientras que los hombres entre 20 y 35 años tienden a pedir prestamos para temas relacionados a los negocios. Por último hay una aparente relación inversa entre el riesgo y la edad, que presumimos se puede llegar a deber a la experiencia laboral y a las propiedades del solicitante.

Referido al tema de los algoritmos, parece ser que las dos mejores opciones son random forest y la regresión logistica respectivamente. Luego del hypertunning el mejor modelo de ambos es random forest, ya que es el que más precisión consigue. El resto de los modelos aplicados consiguen resultados similares, pero con una precisión un tanto menor a estos dos últimos.