Complete a Data Science Capstone Project: Predicting Credit Card Default

Diego Araya Mèndez

2020

Tabla de contenido

Introducción:

El uso de tarjetas de crédito está muy extendido en los Estados Unidos. Alrededor del 70% de todos los estadounidenses usan al menos una tarjeta de crédito y había 364 millones de cuentas de tarjetas de crédito abiertas en los Estados Unidos a fines de 2017, según la Asociación Americana de Bancos (González-García, 2018). Con el uso de tarjetas de crédito viene el riesgo de incumplimiento, que se define como no hacer un pago por 180 días (Konsko, 2014). Este proyecto está basado en una base de datos de una entidad financiera X en Taiwán, si bien los datos no son de Costa Rica es buen parámetro para que las compañías financieras ya se públicas o privadas se beneficien. La realidad país no está ajena a este tipo de problemática.

Objetivo:

El objetivo de este proyecto es predecir efectivamente el incumplimiento del pago con tarjeta de crédito utilizando factores demográficos, datos crediticios, estados de pago, estados de cuenta e historial de pagos.

Sub Objetivos:

- Determinar cómo la probabilidad de pago predeterminado varía según las categorías de diferentes variables demográficas.
- Determinar los predictores más fuertes de pago predeterminado entre las variables

Interesados:

El mercado interesado en este proyecto, naturalmente, es el sector financiero el cual emite tarjetas de crédito. Dichas compañías tienen un interés significativo en predecir qué clientes incumplirán con sus pagos porque tales incumplimientos les cuestan dinero y, por lo tanto, prefieren no extender el dinero a personas con una alta probabilidad de incumplimiento. Un buen modelo de predicción les permitirá prestar a buenos clientes.

Un buen modelo de predicción también permitirá que las compañías de tarjetas de crédito realicen intervenciones tempranas y efectivas con los clientes existentes que tienen probabilidad de incumplimiento. Este estudio puede hacer que la deuda sea más llevadera ya que la entidad podría dar mejor asesoría financiera a sus clientes, permitiendo que el cliente mantenga un buen crédito y que la compañía reciba el pago. En circunstancias

menos afortunadas, una intervención más tardía puede provocar que las compañías de tarjetas de crédito no reciban pagos en comparación con respecto a otros acreedores.

En resumen, las compañías de tarjetas de crédito son las principales partes interesadas en este proyecto, puede mejorar los filtros de aceptación y rechazo de las solicitudes de crédito, así como la decisión de a quién extender las intervenciones y la mejora de la decisión da como resultado mayores ganancias para la compañía.

Vale la pena mencionar que este problema no es muy diferente de otros problemas como la predicción de bancarrota, y que una solución a un problema podría generalizarse fácilmente al otro. Por lo tanto, otros interesados en este problema podrían incluir financieras, bancos y otros grandes acreedores.

Set de Datos

La fuente de datos para este problema es un conjunto de datos de Kaggle.com que contiene aproximadamente 30,000 registros de información sobre clientes de tarjetas de crédito en Taiwán desde abril de 2005 hasta septiembre de 2005. Puede encontrarlo aquí:

https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/home

Estas son las 25 variables del set de datos:

ID: ID of each client

LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit

SEX: Gender (1=male, 2=female)

EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

MARRIAGE: Marital status (1=married, 2=single, 3=others)

AGE: Age in years

PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)

PAY_2: Repayment status in August, 2005 (scale same as above)

PAY_3: Repayment status in July, 2005 (scale same as above)

PAY_4: Repayment status in June, 2005 (scale same as above)

PAY_5: Repayment status in May, 2005 (scale same as above)

PAY_6: Repayment status in April, 2005 (scale same as above)

BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)

BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)

BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)

BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)

BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)

BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)

PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)

PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)

PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)

PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)

PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)

Análisis de los datos

Como el conjunto de datos se obtuvo desde Kaggle, estaba bastante limpio. Sin embargo, se encontraron inconsistencias en los datos que debían corregirse.

Primero, los datos se examinaron en busca de valores faltantes o nulos utilizando el método Pandas .info (). Ninguna columna tenía valores nulos, y los valores de 0 formaban parte del dominio en las columnas en las que se encontraban, por lo que no era necesario hacer nada.

A continuación, se examinaron los valores atípicos en las columnas. Se determinó que no se deben excluir valores del conjunto de datos porque, aunque había muchos valores que cumplían con la definición estadística de ser un valor atípico (mayor que el valor del tercer cuartil + 1.5 veces el rango intercuartílico o menor que el valor del primer cuartil menos 1.5 veces el rango intercuartil) todos estos valores tenían sentido en contexto y no había evidencia de que los valores se hubieran ingresado incorrectamente o de que los datos estuvieran corruptos de alguna manera.

Y, sin embargo, algunos valores sí tuvieron que cambiarse porque violaron la descripción del conjunto de datos. Hubo tres instancias donde esto ocurrió.

En primera instancia, se descubrió que las columnas PAY_X (las columnas que determinan cuántos meses estuvo atrasado un cliente a crédito) contenían valores de -2 y 0 cuando solo deberían haber contenido valores de -1 (lo que representa ningún mes de retraso en pagos) y enteros positivos (que representan el número de meses de retraso para el que se aplicó). Dada la definición de la columna, no tiene sentido tener valores negativos (porque

no puede adelantarse pagos, solo atrasado o lo que está a tiempo). Por lo tanto, se tomó la decisión de reemplazar todos los valores negativos en estas columnas con un 0.

En la segunda instancia, se descubrió que la columna Matrimonio contenía valores de 0, cuando debería haber contenido solo 1 para "Casado", 2 para "Soltero" o 3 para "Otro". Se tomó la decisión de recodificar este valor con un 3 para “otro”, ya que los valores de 0 y 3 se tenían asignados para "Otro".

Finalmente, había valores de 0 para Educación que la descripción del conjunto de datos no daba información al respecto. Además, había una columna para "Otro" y 2 valores para "Desconocido". Dado que todos estos valores representan esencialmente lo mismo, se decidió agrupar todos estos valores bajo una sola codificación. Por lo tanto, los valores de 0 (no contabilizados en la descripción del conjunto de datos), y 5 y 6 (ambos valores para "desconocido") se recodificaron a valores de 4 que representan "Otro".

Además, también detectamos de que la columna PAY_0 tenía un nombre extraño, ya que las otras columnas PAY_X tenían valores en el rango (2, 6). Por lo tanto, se cambió el nombre de la columna a PAY_1.

Hubo 2 componentes principales en la etapa de Análisis Exploratorio de Datos de este proyecto. Primero, se examinaron las correlaciones entre nuestra variable de interés o variable dependiente, "default". Y segundo, se analizaron las correlaciones y las relaciones entre varias variables independientes.

Para el cálculo de la correlación entre el incumplimiento y otras variables, se calculó una correlación simple para cada una de las variables y el incumplimiento, y sus resultados se ordenaron de mayor a menor. Se descubrió que ninguna de las variables tenía una correlación significativa o incluso moderada (definida como tener valores de correlación absolutos superiores a 0,5), pero que la variable LIMIT_BAL (que mide el crédito extendido a un cliente) y las variables PAY_X (que representan el estado de reembolso sobre 6 meses) se correlacionaron más con el incumplimiento. En el caso de LIMIT_BAL, se correlacionó más negativamente con el valor predeterminado que cualquier otra variable, con un valor de correlación de aproximadamente -0.15, y en el caso de las variables PAY_X se descubrió que tenían correlaciones que variaban de aproximadamente 0.24 para PAY_6 todo el camino a aproximadamente 0.4 para PAY_1 con la fuerza de la correlación aumentando con el tiempo posterior (es decir, menor X en PAY_X).

En el caso de las variables PAY_X, tenía sentido que estuvieran más fuertemente correlacionadas con el incumplimiento que cualquier otra variable o grupo de variables, ya que miden el estado de reembolso que intuitivamente puede entenderse como muy

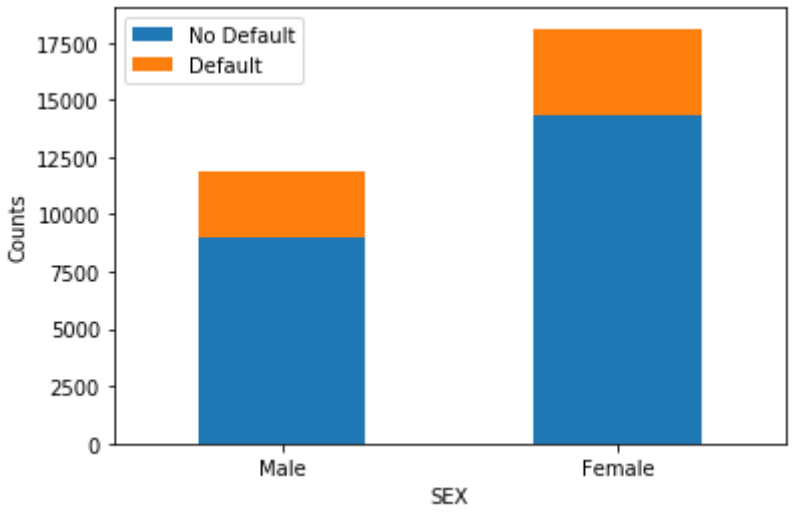
relacionado con el incumplimiento. Además, tenía sentido que la correlación aumentara con el tiempo, ya que estar atrasado en los pagos en septiembre debería estar más correlacionado con el incumplimiento en octubre que en los pagos en abril.

En el caso de la variable LIMIT_BAL y la razón de la correlación negativa con el incumplimiento, una hipótesis de trabajo es que la agencia de crédito solo extendió más crédito a las personas que tenían más confianza para pagar. Por lo tanto, las personas con más crédito deberían haber tenido mayores capacidades de pago y, por lo tanto, tasas de incumplimiento más bajas.

Del Análisis Exploratorio de Datos, el encontrar relaciones entre variables, se descubrieron varias relaciones, pero solo se mencionarán las 4 principales por brevedad:

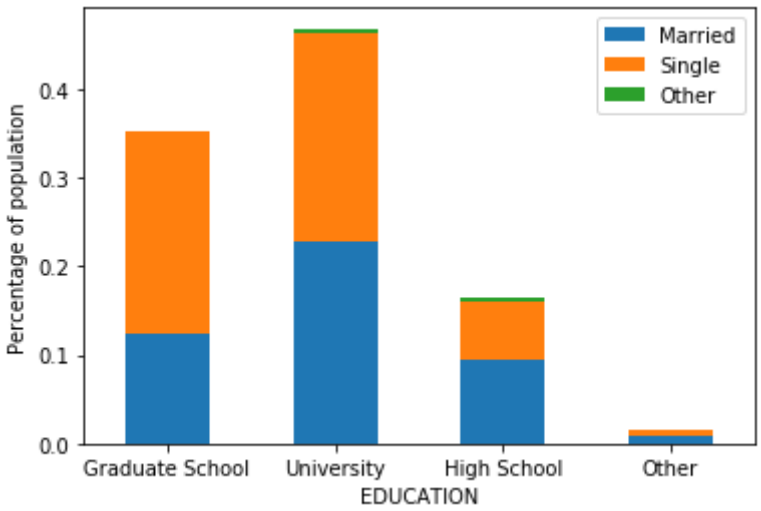
1- El incumplimiento de los hombres es mayor que el de las mujeres.

Esto se puede ver en el siguiente cuadro:



default	No Default	Default	percentage default
SEX			
Male	9015	2873	24.2%
Female	14349	3763	20.8%

2- Las personas que solo tienen educación secundaria tienen tasas de matrimonio más altas que las personas con educación de graduate school education.

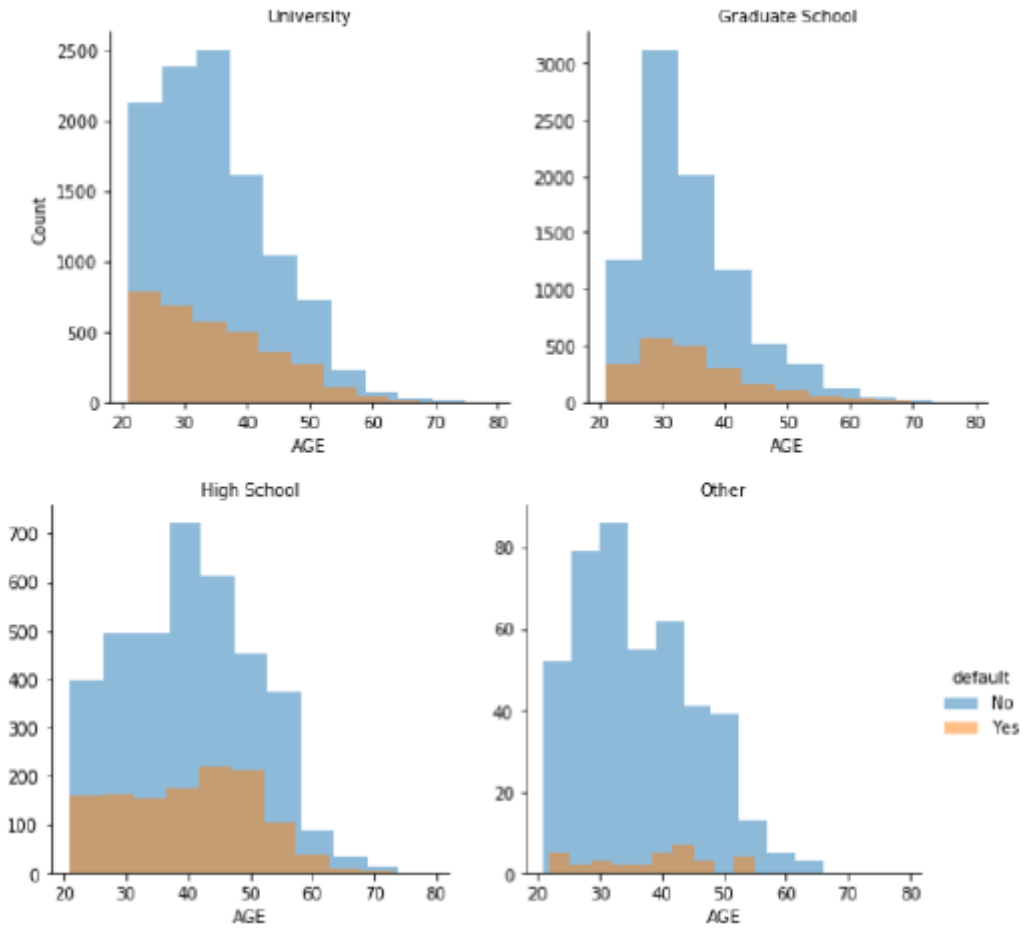


MARRIAGE EDUCATION	Married	Single	Other	Total
Graduate School	35.2%	64.3%	0.5%	100.0%
University	48.8%	50.0%	1.2%	100.0%
High School	58.2%	38.8%	3.0%	100.0%
Other	50.0%	48.3%	1.7%	100.0%

3- La edad media de las personas que solo tienen educación secundaria (aproximadamente 40.3) es considerablemente más alta que la edad promedio de las personas con educación de posgrado (aproximadamente 34.2).

default EDUCATION	No Default	Default
Graduate School	34.1	34.6
University	34.7	34.7
High School	40.3	40.2
Other	36.0	38.2

Los histogramas a continuación también muestran que una mayor educación se asocia con una edad más joven. Además, tanto en el cuadro como en los histogramas a continuación, vemos que las edades de los morosos y no morosos no son significativamente diferentes.



4- Las personas casadas tienen más probabilidades de incumplimiento que los solteros (23.5% vs. 20.9%)

default	No	Yes	Percentage Default
MARRIAGE			
Married	10453	3206	23.5%
Single	12623	3341	20.9%
Other	288	89	23.6%

Estos hallazgos son interesantes para nosotros, y podemos especular sobre por qué son ciertos, pero debemos recordar que son subproductos de nuestro proceso para lograr nuestro objetivo, que es predecir el incumplimiento utilizando 6 técnicas diferentes de aprendizaje automático y determinar si existe Una diferencia entre ellos.

Enfoque para resolver el problema:

Este proyecto implementará 6 técnicas de minería de datos, como se describe en el documento Son:

1- Logistic Regression (LR)
2- K-Nearest Neighbor (KNN)
3- Support Vector Machine (SVM)
4- Decision Trees (DT)
5- Random Forest (RF)
6- Naive Bayes

Los entregables para este proyecto incluyen el código del proyecto, un documento o informe del proyecto y una presentación de los hallazgos del proyecto en una plataforma de diapositivas.

Resultados:

Anteriormente dijimos que nuestra meta era predecir el incumplimiento de pago de las tarjetas de crédito utilizando las características que tenemos disponibles que contienen información sobre la demografía del cliente y la información financiera del cliente.

Para hacer esto, probaremos 6 modelos diferentes de clasificación para ver cuál produce los mejores resultados:

- Logistic Regression
- K Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Decision Tree, Random Forest
- Naïve Bayes.

Comenzamos pre procesando los datos. Hacemos un buen uso del objeto Pipeline () de scikit-learn y de las características de pre procesamiento. Luego dividimos nuestras características en 2 listas, una que contiene características categóricas y la otra que contiene características numéricas / continuas. Luego aplicamos las transformaciones apropiadas: a las características continuas, aplicamos la transformación MinMaxScaler () para que todos los valores se re escalen entre 0 y 1; a las características categóricas aplicamos la transformación OneHotEncoder (). Estas transformaciones son importantes porque muchos clasificadores como KNN usan métricas de distancia para hacer

clasificaciones. Envolvemos cada una de estas transformaciones de Pipeline en un objeto ColumnTransformer () para que puedan aplicarse a sus características apropiadas.

Dividimos los datos mediante la función train_test_split () de scikit-learn. Hecho esto estamos listos para adaptar nuestros modelos a los datos.

Antes de hacerlo, es importante aclarar cómo calificaremos cada modelo. La métrica más natural e intuitiva es la Accuracy. Sin embargo, hay 2 problemas al usar solo la Accuracy:

1) Nuestro conjunto de datos presenta un desequilibrio de aproximadamente 78% a 22%, donde el 78% de los registros están etiquetados como "no predeterminados" y el 22% fueron etiquetados como "predeterminados". En problemas de clasificación, cuanto mayor es el desequilibrio en el conjunto de datos, menor Accuracy informativa es como una métrica, ya que se pueden lograr altas precisiones prediciendo la etiqueta más común.

2) En el contexto de nuestro problema, no todos los errores cometidos por un modelo son iguales. Hay 2 tipos de errores que podemos cometer:

a. Falsos positivos: predecir que una persona tendrá un incumplimiento cuando no lo hará

b. Falsos negativos: predecir que una persona no incumplirá cuando si lo hará

En nuestro problema, nuestra parte interesada clave es cualquier institución que otorgue crédito. Para tales instituciones, el costo de un falso negativo es mucho mayor que el costo de un falso positivo. Por lo tanto, debemos de ser cautos, debemos asignar mayor valor a los modelos que producen menos falsos negativos al costo de más falsos positivos. Esta compensación es conocida en problemas de clasificación como recuperación de precisión compensación. En nuestro caso, queremos maximizar el recuerdo (la proporción de valores predeterminados pronosticados correctamente el número de incumplimientos reales) a costa de la precisión (la proporción de predicciones correctas valores predeterminados para el número total de valores predeterminados previstos).

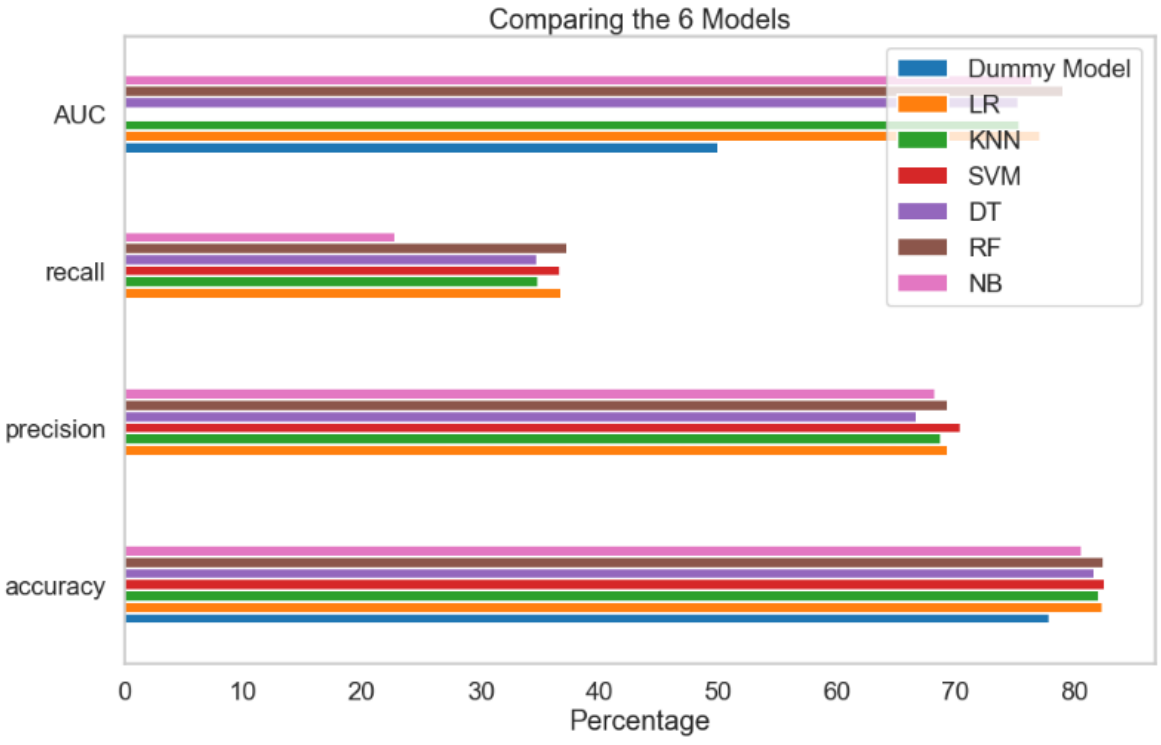
Así que ahora sabemos que el Recall debería ser nuestra métrica de puntuación principal. Además de Accuracy, Precisión y Recall, también realizaremos un seguimiento de la puntuación de AUC para cada modelo, que rastrea el equilibrio entre verdadero- positivo. Finalmente, antes de ejecutar nuestros modelos, creamos un modelo ficticio que predecirá "no default".

Ajustamos los modelos a los datos de entrenamiento, siguiendo los pasos anteriores, predecimos en el conjunto de pruebas y calificamos cada uno de los modelos, agregando cada puntaje a una tabla y así visualizar sus puntajes.

Después de ajustar los modelos y calificarlos, obtenemos una tabla de nuestros modelos y sus puntajes:

	Dummy Model	LR	KNN	SVM	DT	RF	NB
accuracy	0.779	0.824	0.821	0.826	0.817	0.825	0.806
precision	0.000	0.694	0.688	0.704	0.667	0.694	0.683
recall	0.000	0.368	0.349	0.367	0.348	0.373	0.228
AUC	0.500	0.771	0.754	NaN	0.753	0.791	0.765
Time to Train	0.116	122.608	174.053	968.909	14.403	189.350	0.119

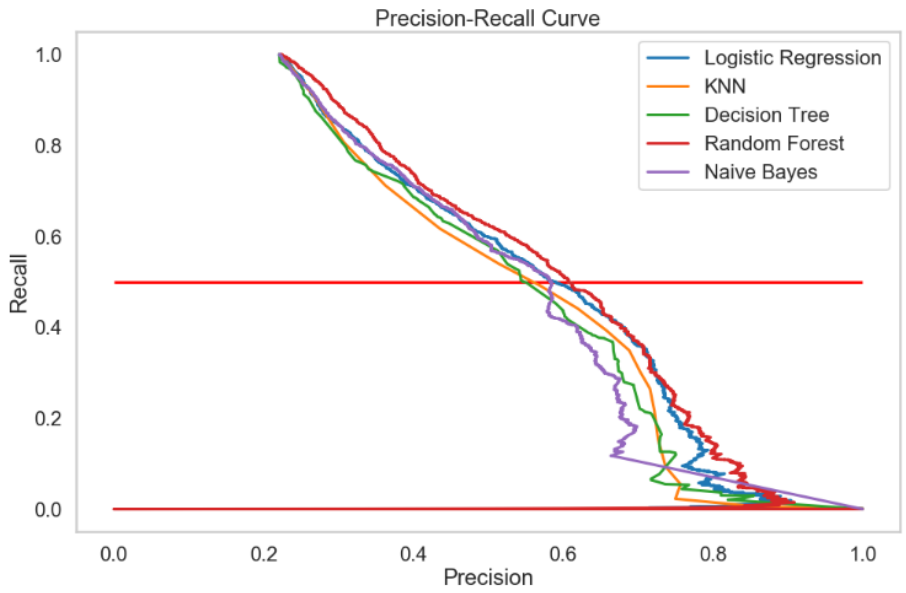
Mirando los resultados, podemos ver dos puntos clave. Primero, que todos nuestros modelos se desempeñaron mejor que el modelo dummy. El modelo dummy tenía puntajes de recall y AUC de 0 ya que nunca predijo predeterminado y tenía una precisión de aproximadamente el 78% ya que ese era el porcentaje de registros no predeterminados en nuestro conjunto de datos En contraste, nuestros modelos lograron precisiones que van desde 80.6% para Naïve Bayes hasta 82.5% para Random Forest. Esto significa que nuestros modelos están funcionando, ya que están encontrando algunos medios para distinguir los morosos de los no morosos. El segundo punto es que, el modelo de regresión logística obtuvo el mejor rendimiento (36.8%) de recall por un poco sobre Random Forest (36.7%). El siguiente cuadro también proporciona una forma útil de ver el rendimiento de los modelos.



Observando la tabla y el gráfico, parece que la Regresión logística y Random Forest tienen la mejor puntuación. Están muy parecidos con respecto a las métricas. Sin embargo, la tabla y el grafico no cuentan toda la historia. Lo que también debemos hacer es mirar un gráfico de precision-recall que muestra cómo el recall cambia con la precisión.

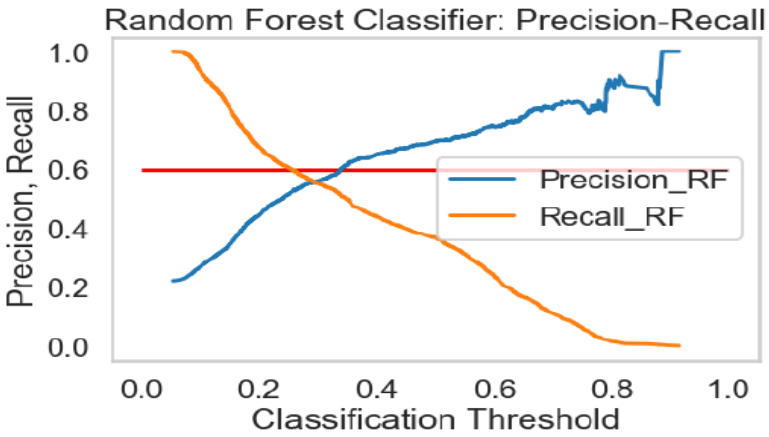
Cuando trazamos curvas de recuperación de precisión para cada uno de nuestros modelos (excepto SVM, esto no se puede hacer desde

SVC () en scikit-learn no tiene el método predict_proba ()), obtenemos la siguiente figura:



Vemos que el modelo Random Forest produce la mayor recall para una precisión dada sobre todos los modelos a excepción de valores de precisión entre aproximadamente 0,63 y 0,78. En ese rango, regresión logística en realidad produce valores más altos de reacall. Porque los valores de precisión dentro de nuestra tabla cayeron dentro de en ese rango, la Regresión logística tuvo una mayor recall que el Random Forest en nuestra tabla. Esto puede ser cambiado alterando el valor de probabilidad de umbral para la predicción para que terminemos en diferentes puntos de nuestras curvas de precisión de recall. Esto nos permitirá alcanzar valores más altos de recall, que deseamos.

La figura a continuación nos muestra cómo cambian la precisión y recall en función del umbral para Random Forest.



Vemos valores de clasificación inferiores a aproximadamente 0.3, el valor recall es mayor que la precisión. por lo tanto, nosotros deberíamos cambiar el umbral de probabilidad utilizado por nuestro clasificador a un nivel predeterminado de 0.5 para lograr mayor recall. Cuando hacemos esto y usamos un nivel de 0.25, logramos un recall del 60.6%, precisión del 51.5%, y accuracy del 78,6%. Claramente, entonces, nuestra accuracy (y precisión) ha dado como resultado del uso de un nuevo valor más bajo para el umbral, lo cual tiene sentido. Por otro lado, nuestro reacall está muy arriba, lo cual es lo que realmente nos importa También podemos generar la matriz de confusión que acompaña el uso de un umbral valor de 0.25. Notamos que contiene más falsos positivos que falsos negativos, como debería:

PREDICTION TRUE	pay	default	Total
pay	5730	1279	7009
default	739	1252	1991
Total	6469	2531	9000

Concluimos que Random Forest se desempeñó mejor. Al alterar el umbral de probabilidad para la clasificación, somos capaces de lograr tasas de recuperación más altas que las que podemos lograr con todos los demás clasificadores. Ya que este es nuestra métrica clave de puntuación, esto convierte a Random Forest en nuestro mejor modelo para el desafío particular de clasificar clientes como morosos o no.

Referencias

Default of Credit Card Clients Dataset. (n.d.). Retrieved from www.kaggle.com:

https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/home
Gonzalez-Garcia, J. (2018, April 26). *Credit card ownership statistics*. Retrieved from

Gonzalez-Garcia, J. (2018, April 26). *Credit card ownership statistics*. Retrieved from www.creditcards.com: https://www.creditcards.com/credit-card-news/ownership-statistics.php

Konsko, L. (2014, September 2). *I Defaulted on My Credit Card – Now What?* Retrieved from www.nerdwallet.com: https://www.nerdwallet.com/blog/credit-cards/credit-card-default-what-to-do/

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 2473-2480.