Ficha Técnica

Hackathon Banistmo 2018

Equipo 31 RRAgile
Modalidad A

Integrantes:

René Delgado
Rafael Ferrero
Raul Ramsay
Manuel Argüelles

Sección 1: Descripción de variables empleadas:

A continuación se presentan las variables que decidimos utilizar en nuestro modelo final. Cabe destacar que no todas se utilizan para entrenar los modelos pero fueron parte de nuestro análisis.

Nombre	**Tipo**	**Descripción**	**Origen**
v_1	int	Corresponde a los días de morosidad	hackaton_training_v1.csv para clientes existentes, APC para el resto
v_2	int	Corresponde al número de días con el trabajo actual	hackaton_training_v1.csv
v_4	int	Corresponde a la edad de la persona	hackaton_training_v1.csv
v_6	double	Corresponde al sueldo de la persona	hackaton_training_v1.csv
v_8	int	Corresponde al score de originación colocada por el banco al momento de la solicitud	hackaton_training_v1.csv
v_10	int	Corresponde al score de cobranza	hackaton_training_v1.csv

Fuentes alternativa de Datos para la evaluación de riesgo

Como parte del ejercicio, se consideraron otras fuentes alternativas de datos para que pudiesen ser incluidas en el Score de Comportamiento de Cobranza. Para nuestro caso, Fuentes Alternativas de Datos serán aquellas que no forman parte de las variables involucradas en el cálculo típico del Score de Originación. Estos datos pueden ser historiales de pago en servicios y utilidades, como lo es agua, luz, teléfono e internet; Perfiles de los clientes armados con base en su comportamiento en redes sociales, tales como reputación, perfil psicológico o número de contactos en redes como facebook, twitter, LinkedIn; historial de transacciones con entidades privadas como mueblerías; entre otros.

Estas nuevas fuentes, no deberían ser consideradas estrictamente necesarias, ya que no todos los clientes del banco, o los que aún no lo son, podrán tener datos en estas plataformas. Pero, a pesar de esto, consideramos que podría existir valor en estos datos que podrían impactar, de forma negativa o positiva el Score de Comportamiento de Cobranza de aquellos que sí tienen información disponible.

1. **Autoridad de Innovación Gubernamental (AIG),** cuenta con consultas en línea. La vigencia del seguro de auto de un conductor podría ser considerado con mejor Score de Originación; Contar con una póliza vencida podría denotar poca responsabilidad en los pagos impactando negativamente el Score. Datos requeridos: Copia de la licencia del conductor/dueño del vehículo (Número de control) y número de cédula. Enlace de la fuente: AIG-SOAT

2. **Registro Público,** mantiene una base de datos de acceso al público en la cual se pueden extraer datos generales de las fincas, como su dueño, miembros de junta directiva de una sociedad anónima, hipotecas, entre otros. Si un cliente, cuenta con una finca, esto se puede interpretar como un bien que puede servir de garantía en caso de un incumplimiento de un crédito. Adicionalmente, podemos obtener el número de finca de la propiedad y cotejar con la Dirección General de Ingresos si la misma está paz y salvo (ver punto 3). Datos necesarios: número de cédula y nombre completo. Enlace a la fuente: Registro Público

3. **Dirección General de Ingresos (DGI),** permite consultas en línea sobre las fincas y el estado de sus respectivos impuestos. Un cliente con sus fincas paz y salvo en impuesto inmueble denota ser más responsable con el pago de sus créditos. Enlace de la fuente: DGI-MEF

4. **Órgano Judicial,** cuenta con la búsqueda de las partes que estén involucradas en un proceso civil. Realizar una búsqueda por el nombre del cliente permite validar si el mismo ha sido demandado civilmente por incumplimiento de responsabilidades financieras: Enlace de la fuente: Órgano Judicial

5. **Panama Compras,** la plataforma en línea por la cual se realizan la mayoría, por no decir todas, las compras del Gobierno de Panamá. Buscar por nombre del oferente

podría permitir determinar si el cliente del banco, a través de sus ventas al estado, tiene buen manejo de sus finanzas, entrega a tiempo, entre otros aspectos financieros que podrían aportar al SCORE. Enlance: PanamaCompras

6. **Autoridad del Canal de Panamá (ACP)**, cuenta con una plataforma para publicar sus licitaciones y permitirle a los proveedores enviar sus ofertas, esta data es de acceso a las personas que se registran al sistema de forma gratuita. En este sistema de podrían ubicar transacciones de personas que participen como proveedores del Canal de Panamá. Enlace: MiCanaldePanama

7. **Instituto Nacional de Acueductos y Alcantarillados (IDAAN),** permite extrar el historial de pago de sus clientes solamente introduciendo el Número de Cliente (NIT).

8. **Autoridad de Tránsito y Transporte Terrestre (ATTT)**, a través de la plataforma de su proveedor de sistemas permite acceder al historial del conductor, incluyendo sus faltas y si las mismas han sido canceladas a la entidad. Enlace: ATTT-Sertracen

9. **LinkedIn**, siendo una red social mayormente para temas profesionales, podemos asumir que la información ubicada en este perfil del cliente podría ser de mayor utilidad que las redes sociales ajenas a la orientación de negocios. Un perfil con una cantidad considerable de contactos podría ser considerado una condición positiva para un cliente, así como también se evidencie en su perfil su nivel educativo, ambas variables han sido consideradas como favorables y que podrían mejorar el Score de Originación de un cliente. De forma contraria, el no contar con un perfil en LinkedIn, o no tener muchas conecciones no debería perjudicar al cliente castigándose su Score de Originación. Enlace: LinkedIn

Sección 2: Descripción de procedimiento de Extracción, Transformación y Carga (ETL):

a. Selección de variables: método de selección de variable

El proceso de selección de variables consistió en los siguientes pasos:

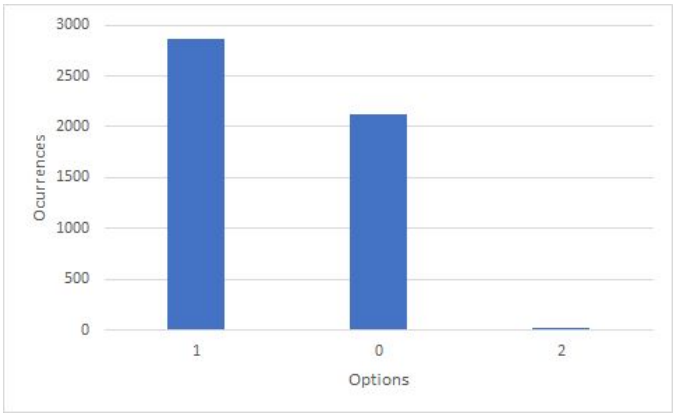
- Se identificó el tipo de datos de cada columna para descartar el campo identificador de secuencia v_0 y el campo representativo a una identificación v_1.
- Se realizó un proceso de análisis de la data restante extrayendo el valor promedio, la desviación estándar, el valor mínimo y el valor máximo de cada columna.

	v_0	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_10	v_11	v_12
AVG		42.68741	1868.599	150.9614	41.75926	0.578747	1012.017	16.51914	410.276	-393.142	218.3546	48.41785	35.27777
MEDIAN		19	1815	146	40	1	520.28	0	538	19	133	56	38
MIN		0	0	0	0	0	0	0	0	-999	0	0	0
MAX		2868	3564	321	96	2	122021.7	12000	975	1423	983	81	72
VAR		7047.146	1129985	10472.83	178.3085	0.250203	8083084	70436.19	123461.2	295697.8	60338.06	271.879	269.8504
STDEV		83.95568	1063.114	102.3471	13.35456	0.500253	2843.361	265.4247	351.4056	543.8354	245.6626	16.4904	16.42877

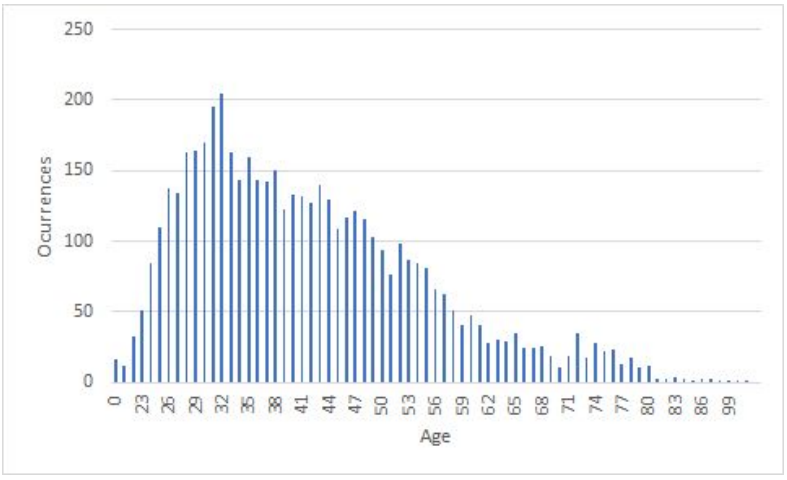
- Se generó la matriz de correlación de las columnas restantes contenida en el dataset para determinar las variables más significativas.

	v_1	v_2	V_3	v_4	v_5	v_6	v_7	v_8	v_9	v_10	v_11	v_12
v_1	1	0.03499	0.01799	-0.06802	0.01657	-0.0152	-0.01422	-0.06854	0.03346	-0.31066	0.02724	-0.01517
v_2	0.03499	1	-0.0091	-0.15813	0.04453	-0.012	0.01054	0.03376	0.0279	-0.08005	0.0678	0.01782
V_3	0.01799	-0.0091	1	-0.06166	0.00341	-0.01116	-0.01298	0.01438	0.01554	-0.02308	0.15237	-0.04163
v_4	-0.06802	-0.15813	-0.06166	1	0.02319	-0.02544	0.00498	-0.26128	-0.3446	0.21658	-0.06711	0.03575
v_5	0.01657	0.04453	0.00341	0.02319	1	0.04556	0.00078	-0.03472	-0.05323	-0.03584	0.0034	0.06158
v_6	-0.0152	-0.012	-0.01116	-0.02544	0.04556	1	0.10771	0.25135	0.29237	0.11077	-0.03707	-0.02393
v_7	-0.01422	0.01054	-0.01298	0.00498	0.00078	0.10771	1	0.05426	0.05354	0.0339	0.01636	0.00475
v_8	-0.06854	0.03376	0.01438	-0.26128	-0.03472	0.25135	0.05426	1	0.63311	0.15519	-0.02155	-0.09252
v_9	0.03346	0.0279	0.01554	-0.3446	-0.05323	0.29237	0.05354	0.63311	1	-0.05404	-0.07754	-0.11756
v_10	-0.31066	-0.08005	-0.02308	0.21658	-0.03584	0.11077	0.0339	0.15519	-0.05404	1	-0.01178	0.01946
v_11	0.02724	0.0678	0.15237	-0.06711	0.0034	-0.03707	0.01636	-0.02155	-0.07754	-0.01178	1	0.33368
v_12	-0.01517	0.01782	-0.04163	0.03575	0.06158	-0.02393	0.00475	-0.09252	-0.11756	0.01946	0.33368	1

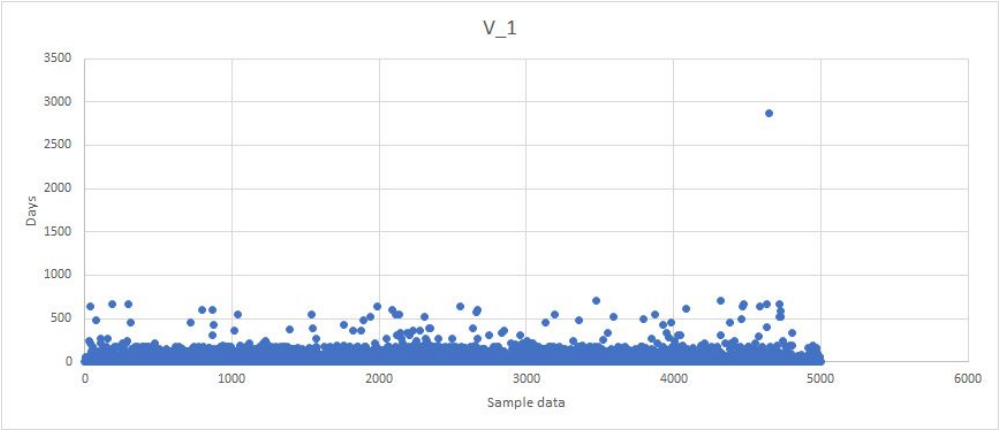
- Se realizó un proceso de descarte para identificar las columnas pertenecientes al sexo por la cantidad de valores [0,1,2]. Se escogió la v_5 como el sexo.



- Se identificó la columna de edad por el set de datos que en un rango entre 21 y 100.



- Se identificaron las columnas de salarios como la v_6 y la de ingresos extra como la v_7 porque eran las únicas columnas que contenían valores decimales.
- Se identificó la columna de días de morosidad como la columna v_1, por la distribución de su data. ⅔ de la misma se ubica entre los 30 y 60 días, con valores menos probables fuera de esos rangos.



- Si asumimos que el score de cobranza real, está fuertemente relacionado a los días de morosidad, buscamos la correlación más alta correspondiente a la columna v_1, la cual resultó ser la columna v_10.
- A partir de la hipótesis del paso anterior y considerando que la columna v_10 contiene la mayor cantidad de valores absolutos de correlación alto con las demás columnas, se estableció como la variable dependiente o score de cobranza real.

b. Ingeniería de variables: transformaciones, agregaciones o variables generadas

- De acuerdo a la matriz de correlación, se procedió a descartar aquellas columnas que guardaban la menor correlación con las demás y sobre todo con la columna de score de cobranza real (v_10): v_0, v_3, v_5, v_7, v_11 y v_12.
- De la columna v_4, correspondiente a la edad, se descartaron todos los valores iguales a 0 por considerarlo fuera de la muestra representativa para una persona mayor de edad. Se eliminaron las filas correspondientes de la matriz de data.
- De la columna v_6, correspondiente al salario, se descartaron los valores mayores a $30K por no tener una muestra representativa para ese rango. Se eliminaron las filas correspondientes de la matriz de data.
- De la columna v_1, correspondiente a los días de morosidad, se descartaron los valores mayores a 1000 por no tener una muestra significativa en ese rango.
-

Sección 3: Descripción de Metodología de Modelamiento:

a. Algoritmo Seleccionado

Para métodos alternativos de valoración de riesgo, se decidió implementar modelos de machine learning que pudieran ser más efectivo y precisos en la estimación de comportamiento de un cliente. Con el dataset entregado y estas técnicas de predicción, se

buscó entender realmente las variables más influyentes que pudieran determinar el comportamiento de pagos y lograr una tasa de error menor a la del banco.

Luego de evaluar varios modelos de machine learning, se seleccionó el *Random Forest Regressor*. Se optó por un modelo de regresión, en lugar de clasificación, para adaptarnos mejor a la naturaleza del scoring utilizado en el dataset.

Este algoritmo de estimación consiste en combinar un número N de árboles de decisión en muestras parciales de la data de entrada y luego promediando sus resultados para mejorar la precisión del modelos. Se usaron siempre las mismas muestras para entrenar los árboles de decisión que conforman el Random Forest Regressor.

Para la implementación del algoritmo en python se utilizó la librería de machine learning de código abierto de scikit-learn.

b. Algoritmos Comparados

Para la elección del algoritmo utilizado, se compararon los siguientes algoritmos:

- Decision Tree Regressor: Este modelo de árbol de decisión consiste en la aplicación de un conjunto de reglas en cascada que se arma través del análisis de data de prueba o de entrenamiento. Se utilizó la librería scikit-learn.

c. Forma de separación del set de datos de validación, prueba y entrenamiento

El dataset entregado fue separado en dos conjuntos. El 40% se utilizó para el entrenamiento y 60% para la validación del modelo. Ambos conjuntos se eligieron de forma aleatoria para tener mayor uniformidad en ambas muestras. Para esto se utilizó el método *train_test_split* de scikit-learn.

d. Metodología de reducción de error

Para lograr una mayor reducción de error, se realizaron varias pruebas modificando las siguientes variables:

- Se varió la cantidad de estimadores en el modelo de Random Forest Regressor. Esta variable indica la cantidad de árboles de decisión a utilizar en el algoritmo.
- Se excluyeron ciertas filas presentes en las columnas seleccionadas, como se explicó en la sección de transformación de datos, donde los valores se encontraban muy alejados del valor promedio. Al entrenar los modelos con data más representativa, la tasa de efectividad aumentó.

e. Métrica de Rendimiento:

Al escoger modelos de regresión, se utilizó el RMSLE como métrica principal de rendimiento de nuestra propuesta.

1. RMSLE de entrenamiento y de validación

El RMSLE o Root Mean Squared Logarithmic Error, se calcula usando el paquete metrics de la librería sklearn. En las pruebas de entrenamiento y validación arrojaron mejores resultados para el modelo de Random Forest Regressor, por lo cual se seleccionó sobre el modelo de Decision Tree Regressor.

Como criterio de comparación adicional, también se evaluó el promedio y la desviación estándar entre los resultados reales y las predicciones arrojadas por cada modelo.

En la evaluación del modelo Random Forest Regressor, los datos fueron:

RMSLE	2.791
Promedio	153.084
Desviación Estándar	136.960

Sección 4: Referencias:

a. Citar todos los artículos científicos o páginas web utilizados en el desarrollo del modelo.

- Age Of Social Credit: China
- Women more prone than men to miss card payments
- FICO's 5 factors: The components of a credit score
- Deep learning with Python
- Scikit Learn
- Kaggle
- More Education = More Income: NY Times