Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Trabajo Practico N°1

Materia: Aprendizaje Automático

Integrantes:

* Martin Llada
* Rodrigo Mauriño
* Juan Jose Iguaran

Nro. de grupo: 6?

Fecha de entrega: 27/05/2019

1. Resumen

En este trabajo se implementa un modelo de Árbol de decisión con el objeto de predecir el otorgamiento (o no) de préstamos hipotecarios en el Estado de Nueva York para el año 2015. A partir de entrenar diferentes modelos -combinando diferentes niveles de altura y valores para los hiperparámetros- se arriba a que el árbol con una profundidad de 6 y que utiliza, como criterio de selección de atributo el método Gini gain reporta los mejores resultados relativos en término de la performance. Esto se evidencia al computar las medidas de accuracy (en promedio, 82%) y ROC AUC (en promedio, 78%). Por otro lado, se encuentra que este árbol reduce levemente su capacidad de predicción cuando el atributo más relevante en el conjunto de datos, posee datos faltantes o ruido. Por último, con el objeto de comparar este método de clasificación con otra técnica, se comparó la performance del árbol de decisión mencionado con el método Naive Bayes con distribución de probabilidad Bernoulli. Si bien, ambos modelos presentan buenas medidas de desempeño (utilizando ROC AUC), se evidencia que para el conjunto de datos de validación el algoritmo de Naive Bayes posee una media más elevada y un menor desvío estándar.

1. Introducción

El objetivo del trabajo es predecir y analizar otorgamientos de préstamos hipotecarios en el estado de Nueva York en 2015 utilizando herramientas en Python.

Se utilizaron algoritmos predictivos de Arboles de Decisión y Naive Bayes para clasificar. Para cada modelo se analizó el desempeño frente a la ocurrencia de datos faltantes generados y tolerancia al ruido.

El Dataset utilizado en el proyecto proviene del portal Kaggle y es de dominio público. Es una fracción del conjunto de datos original de créditos hipotecarios otorgados desde el 2007 hasta la actualidad para todos los estados de Estados Unidos. Estos reportes son confeccionados por la entidad Consumer Finance Protection Board (CFPB).

Se tomaron en cuenta distintas variables relacionadas con el solicitante como la situación económica, género, etnicidad, estado habitacional del inmueble, objetivo de la hipoteca y ubicación además de datos respecto a la agencia inmobiliaria, entre otros atributos. Si bien inicialmente se contaron con 6 resultados distintos que clasificaban la decisión tomada para cada caso, el análisis se hizo sobre el agrupamiento de estos 6 resultados en 2 posibles clasificaciones: Otorgado y No Otorgado.

3. Datos

El Dataset original para el estudio posee 439.655 casos, 77 atributos y 1 atributo clasificatorio que puede tomar 6 valores distintos que representan la acción a tomar respecto al crédito hipotecario. Con el objetivo de simplificar el análisis se substrajo al azar mediante una distribución uniforme un subset menor al 1% del total de casos, dejando solamente 3478 casos para estudiar.

Inicialmente se aplicó un pre filtro para mejorar la calidad de datos y preservar únicamente los atributos más relevantes para el estudio. Con el pre filtro se eliminaron atributos redundantes, atributos constantes, número identificador de cada caso y columnas que contenían principalmente datos faltantes. Se presentan en la **Tabla 1** los 28 atributos utilizados en el análisis.

|  |  |  |  |
| --- | --- | --- | --- |
| **Columna** | **Descripción** | **Tipo de Dato** | **Clasificación de datos** |
| agency\_code | Agencia inmobiliaria | Numérico | Categórico |
| applicant\_ethnicity | Etnicidad de solicitante | Numérico | Categórico |
| applicant\_income\_000s | Ingresos de solicitante en miles | Numérico | Continuo |
| applicant\_race\_1 | Raza de solicitante | Numérico | Categórico |
| applicant\_sex | Genero de solicitante | Numérico | Categórico |
| application\_date\_indicator | Indicador fecha de aplicación | Numérico | Categórico |
| census\_tract\_number | Número de Censo | Numérico | Categórico |
| co\_applicant\_ethnicity | Etnicidad de co-solicitante | Numérico | Categórico |
| co\_applicant\_race\_1 | Raza de co-solicitante | Numérico | Categórico |
| co\_applicant\_sex | Genero de co-solicitante | Numérico | Categórico |
| county\_name | Nombre de condado | String | Categórico |
| hoepa\_status | Estado hoepa | Numérico | Categórico |
| lien\_status | Estado de derecho de retención | Numérico | Categórico |
| loan\_purpose | Propósito de préstamo | Numérico | Categórico |
| loan\_type | Tipo de préstamo | Numérico | Categórico |
| msamd | msamd | Numérico | Categórico |
| owner\_occupancy | Ocupación de solicitante | Numérico | Categórico |
| preapproval | Pre aprobación | Numérico | Categórico |
| property\_type | Tipo de Propiedad | Numérico | Categórico |
| purchaser\_type | Tipo de comprador | Numérico | Categórico |
| sequence\_number | Número secuencial | Numérico | Categórico |
| hud\_median\_family\_income | Ingreso medio de familia | Numérico | Continuo |
| loan\_amount\_000s | Valor de préstamo en miles | Numérico | Continuo |
| number\_of\_1\_to\_4\_family\_units | Unidades familiares | Numérico | Continuo |
| number\_of\_owner\_occupied\_units | Unidades ocupadas | Numérico | Continuo |
| minority\_population | Población minoritaria | Numérico | Continuo |
| population | Población | Numérico | Continuo |
| tract\_to\_msamd\_income | Ingreso msamd | Numérico | Continuo |

(Tabla 1: Atributos estudiados)

Como se puede observar el Dataset cuenta con una amplia variación datos de distinta naturaleza: 8 datos continuos y 20 datos categóricos, uno de los cuales se presenta como Tipo de dato String, county\_name. Debido a los procesos clasificatorios que vamos a someter a este Dataset es conveniente transformar la columna county\_name en un dato Numérico. Por lo cual se procedió a realizar un Label Encoder en la columna county\_name y se transformó cada valor distinto en String a un valor numérico utilizando números enteros consecutivos.

La columna del atributo clasificatorio se llama action\_taken y presenta valores categóricos numéricos, a continuación se presentan en la Tabla 2:

|  |  |  |  |
| --- | --- | --- | --- |
| **action\_taken** | **Descripción** | **Cantidad** | **Proporción** |
| 1 | Loan originated | 1951 | 56.10% |
| 2 | Application approved but not accepted | 129 | 3.71% |
| 3 | Application denied by financial institution | 630 | 18.11% |
| 4 | Application withdrawn by applicant | 353 | 10.15% |
| 5 | File closed for incompleteness | 142 | 4.08% |
| 6 | Loan purchased by the institution | 273 | 7.85% |

(Tabla 2: Clases originales)

Hay una claro desbalance entre las 6 posibles clases, donde la primera clase tiene casi la misma cantidad de casos que la suma del resto. Nuevamente por la naturaleza de los modelos predictivos que vamos a usar en el análisis de los datos debemos transformar la clase action\_taken en valores binarios. Se tomó el criterio de clasificar los códigos 1,5 y 6 como prestamos otorgado y el resto de los casos representan el caso contrario.

|  |  |  |  |
| --- | --- | --- | --- |
| **action\_taken** | **Descripción** | **Cantidad** | **Proporción** |
| 0 | Aprobado | 2366 | 68.03% |
| 1 | No Aprobado | 1112 | 31.97% |

(Tabla 3: Clases usadas para el estudio)

Una vez aplicada la transformación vemos ahora un leve desbalance hacia la clase “Aprobado”.

Bibliografia

<https://www.consumerfinance.gov/data-research/hmda/explore>

<https://www.kaggle.com/jboysen/ny-home-mortgage>