**Datatón 2018 – Bancolombia**

**Personal Financial Managers**

**Romel Casadiegos Barrios**

**Luis Alberto Gaviria García**

**Angy Melissa Medina Berrio**

**Medellín 2018**

**Resumen descriptivo del procedimiento**

En la siguiente descripción se plasma el proceso llevado a cabo con el fin de obtener un modelo que genere valor para el usuario de una aplicación PFM. Se ofrecen diferentes aproximaciones a la solución del reto desde varias miradas de la utilidad potencial de los datos.

1. Análisis de texto:

En una primera fase se realiza un análisis de texto basado en un modelo LDA - Topic Modeling and Latent Dirichlet Allocation – el cual es un tipo de modelado estadístico que permite descubrir los temas abstractos que aparecen en una colección de documentos. Como resultado del modelo se obtiene un listado de tópicos con las palabras que pertenecen a cada tema y su peso relativo que explica el nivel de pertenencia de la palabra a cada tema.

El modelo se construye mediante un análisis bag of words y en determinar las frecuencias de cada palabra en un documento y en todos los documentos.

Con base a este análisis se construye un modelo llamado Term frequency – Inverse document frequency (tfidf), donde:

TF = (Número de veces que la palabra clave aparece en el documento) / (Número total de palabras en el documento)

IDF = (Número total de documentos) / (Número total de veces que aparece la palabra en todos los documentos)

Finalmente se entrena el modelo LDA, el cual permite predecir tópicos relacionados a un nuevo documento que no ha visto el modelo.

Este algoritmo tiene las siguientes aplicaciones:

* Sistema de recomendación: Los tópicos hallados pueden ser usados para sugerirle al usuario una clasificación de sus transacciones PSE, ya que mediante la predicción de nuevos documentos, obtenemos un score de los tópicos posibles que se adaptan según un análisis de la relación de las palabras nuevas con el diccionario de palabras entrenado para un histórico de transacciones.
* Etiquetado del dataset suministrado: Mediante el análisis de la relación entre las palabras, podemos hallar un número de tópicos que generalicen los temas tratados dentro de las transacciones vistas por el modelo. Con lo cual podemos asignar una clasificación descriptiva de cada tópico basados en las palabras relacionadas y su nivel de relación, luego estas clasificaciones serán asignadas a cada transacción, para obtener finalmente un dataset etiquetado. Este dataset será el insumo para un modelo supervisado de clasificación multiclase.
* Columna categórica: Los tópicos hallados podrán formar una columna categórica que harán parte de las variables predictoras de un modelo supervisado multiclase, etiquetado por medio de análisis de los sectores y subsectores.

Hallazgos:

* Se hace necesario mayor cantidad de registros y palabras descriptoras de los registros para aumentar el nivel de relación entre las palabras y mejorar así la calidad del modelo LDA. Ya que actualmente no es posible encontrar tópicos fuertemente diferenciados debido a la débil relación entre las palabras y su contexto.

1. Etiquetado del dataset de transacciones:

Debido a que el análisis de texto no obtuvo una diferenciación clara de los tópicos, se opta por realizar un etiquetado según un análisis de agrupación de subsectores, basándose en 3 clases: Pago de obligaciones, Personal y hogar y otros. Los cuales pueden dar un acercamiento al tipo de transacción del usuario y poder generar una clasificación automática de los registros dentro de la aplicación.

Para este etiquetado se tuvo en cuenta aproximadamente 3 mm de registros que contienen el etiquetado del sector.

En el proceso de etiquetado se realiza un análisis manual de referencias, sectores, subsectores y descripción para obtener un indicio del tipo de transacción y poder clasificar en alguna de las 3 clases propuestas.

Hallazgos:

* Tener una dataset con una mayor cantidad de transacciones balanceadas en cuanto a la clase a la que pertenece cada registro, permitiría aumentar el rendimiento del modelo.

1. Limpieza y consolidación de los datos:

Con el fin de realizar una limpieza de los datos y obtener un Dataset útil para el modelado, se realiza el siguiente preprocesado:

* Filtrado de datos: Se eliminan los registros de clientes sin información en la edad, y se remplazan valores categóricos que no están proporcionados en el diccionario suministrado por BC, como los hallados en tipo de vivienda y rango de ingresos.
* Análisis de outliers de edades: Se eliminan las edades atípicas o datos erróneos como: negativos o extremos, según un análisis de outliers determinado por Isolation Forest. Finalmente perduran las edades de 10 a 82 años.
* One hot encoding dataset pagadores: Se realiza un one hot encoding de las clases categóricas del dataset pagadores como: "seg\_str",'ocupacion',"genero","edad","ingreso\_rango","tipo\_vivienda","estado\_civil","nivel\_academico"
* Análisis outliers valor transacciones: Inicialmente se eliminan las transacciones que son menores o iguales a 0 pesos. Según el score de anomalías se decide filtrar las transacciones mayores a 10mm, se eliminaran 1738 transacciones, estas se definen como transacciones atípicas y corresponden a menos del 0.01%

Hallazgos:

* Se hace necesario un mayor nuevo de transacciones disponibles para el entrenamiento, que contengan una mayor proporción de registros correspondientes a la clase 3 – otros. Ya que esta desproporción puede afectar el entrenamiento de los modelos de clasificación, influenciando la clasificación a aquellas clases de mayor cantidad de registros.

1. Modelado Semi\_supervisado:

Se plantea el uso de un modelo semi-supervisado capaz de predecir la etiqueta de registros sin clasificación, debido a que se cuenta con una cantidad de registros significativa que no pudo ser etiquetada a partir del sector o subsector.

Según el dataset preprocesado obtenido en el anterior numeral, se implementa un modelo semi supervisado.

El dataset es dividido en entrenamiento y pruebas, se le incorpora una reducción de dimensionalidad con el fin de reducir la complejidad del modelo. Luego se implementa el modelo label\_propagation este se basa en definir fronteras de datos etiquetados y tras cada iteración redefine las fronteras basándose en la distancia de los nuevos datos, sin importar que estos datos no tengan una etiqueta definida. Finalmente se aplica GridSearchCV, el cual basado en validación cruzada y en un score balanced\_accuracy encuentra el mejor modelo dado por los diferentes hiperparámetros.

Se elige balanced\_accuracy como medición del score para modelos multiclase desbalanceados.

Se ajusta el modelo y se selecciona el mejor modelo según los hiperparámetros encontrados.

Hallazgos:

* El modelo utilizado no fue capaz de generalizar la clasificación, obteniendo un bajo score. Se esperaría obtener un mejor rendimiento del modelo con la incorporación de nuevas variables explicativas que ayuden a encontrar relaciones entre los datos.

1. Modelo supervisado SVC:

Se propone el uso de un modelo supervisado basado en el dataset etiquetado por medio del análisis de sector y subsector.

Se toma una proporción de datos que balanceen la muestra.

Al dataset se le incorpora una reducción de dimensionalidad, luego se implementa el modelo SVC para clasificación multiclase. Finalmente se aplica GridSearchCV, el cual basado en validación cruzada y el score balanced\_accuracy encuentra el mejor modelo dado por los diferentes hiperparámetros.

Se elige balanced\_accuracy como medición del score para modelos multiclase desbalanceados y se ajusta el modelo.

Hallazgo:

* El modelo no fue capaz de generalizar la clasificación y aprender patrones de los datos que le permitan dar una buena clasificación. Obteniendo una clasificación en una sola clase y por tanto un score bajo, dado por la clase predicha.

1. Modelo supervisado RF y Adaboost Dataset:

Se toma una proporción de datos que balanceen la muestra, se divide el dataset en entrenamiento y prueba. Se incorpora una reducción de dimensionalidad.

Luego, se implementa el modelo RandomForestClassifier para clasificación multiclase y se implementa el modelo AdaBoostClassifier el cual es un tipo de modelo de ensamble en el cual se basa en etimadores DecisionTreeClassifier. El cual por medio de pesos ajusta las instancias clasificadas incorrectamente de modo que las clasificaciones posteriores se centren más en los errores cometidos.

Finalmente se aplica GridSearchCV para cada tipo de modelo, el cual basado en validación cruzada y el score balanced\_accuracy encuentra el mejor modelo dado por los diferentes hiperparámetros.

Se elige balanced\_accuracy como medición del score para modelos multiclase desbalanceados y se ajusta el modelo.

Se presenta la matriz de confusión en la cual se analiza el desempeño del modelo según el número de predicciones de cada clase dadas por: VP, VN, FP, FN.

Se presenta un reporte del score del modelo basado en un comparativo de métricas

Hallazgos:

* Se obtuvo un mejor rendimiento del modelo RandomForest respecto al modelo Adaboost.
* Se observa que el modelo RandomForest generaliza bien la frontera entre las clases, debido a que tanto su rendimiento en entrenamiento como en pruebas es similar, logrando un score del 86% en test.
* El modelo puede tener oportunidades de mejora, explorando nuevos hiperparámetros que puedan mejorar el score, sin embargo se requiere mejor capacidad de cómputo.
* El modelo puede ser mejorado aumentando la cantidad de datos, debido a que su curva de aprendizaje mejora a medida que se aumentan los datos.