Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

REVISIÓN Y ACTUALIZACIÓN DE MODELO PARA LA CLASIFICACIÓN DE PROYECTOS DE REGALÍAS

Entidades involucradas

Departamento Nacional de Planeación

- Dirección de Vigilancia de las Regalías.
- Dirección del Sistema General de Regalías.
- Dirección de Desarrollo Digital.

Sector	Lenguaje	Fuente de datos
Planeación	Python.	DNP-Sistema General de Regalías.

Presentación

El año pasado, la unidad de científicos de datos (UCD) del DNP desarrolló un modelo para predecir dificultades en la ejecución de proyectos de regalías, a partir del texto de sus documentos de formulación. Este modelo fue desarrollado con el objetivo de apoyar al Sistema General de Regalías (SGR) de diversas formas, incluyendo evidenciar fallas tempranas en la formulación de proyectos, priorizar seguimiento a proyectos y entidades ejecutoras y optimizar recursos. A partir de este proyecto se viene realizando una revisión metodológica, para identificar oportunidades de mejora del modelo. Entre las acciones que se han tomado se incluyen la actualización de la información utilizada para entrenar el modelo, la inclusión de otras variables, la revisión de los modelos/algoritmos utilizados y el acercamiento a las áreas de formulación y seguimiento de proyectos de regalías, para tratar de comprender mejor sus necesidades y hacer del modelo lo más útil posible. Este trabajo se presenta a continuación, junto a los resultados obtenidos.

Last year, DNP's data science team developed a model to predict difficulties in the execution of royalty investment projects, based on the text of the project formulation documents. This model was developed with the objective of supporting the general royalty system (Sistema General de Regalías, SGR) in different ways, including detecting early failures in the formulation of projects, prioritizing follow-up of projects and executing entities, and optimizing resources. A methodological revision is currently being made to this project, to identify opportunities to improve the model. Actions are being taken such as updating and searching for new sources of information, reviewing the models/algorithms used, and approaching the areas responsible for formulation and monitoring of royalty investment projects, to better understand their needs and to make the model as useful as possible.

Objetivo general

Hacer una revisión completa tanto del código y metodología desarrolladas como de la información utilizada para el proyecto de regalías, para de esta forma identificar oportunidades de mejora en el modelo desarrollado.

Objetivos específicos

- Revisar y actualizar las fuentes de información actuales con las cuales se entrena y prueba el modelo.
- Buscar nuevas fuentes de información, o nuevas variables que puedan agregar valor y mejorar el desempeño del modelo.
- Evaluar la calidad del IGPR como variable objetivo del modelo, y evaluar los umbrales establecidos anteriormente.
- Hacer una revisión de los algoritmos y herramientas utilizados para desarrollar el modelo.
- Entrenar de nuevo el modelo, y validar si hay una mejora en el desempeño.

Metodología

La metodología para la revisión metodológica y redesarrollo de este proyecto involucra varias etapas. A continuación, se describen cada una de estas etapas.

Entendimiento del problema y recolección de datos

Lo primero que se realizó fue ganar entendimiento sobre el modelo realizado el año pasado por la UCD. Para ello, se llevaron a cabo 3 acciones:

Revisión del código existente

Lo primero fue hacer una recolección e inventario del código que se desarrolló para la primera etapa de este proyecto. Este código estaba dividido en varios archivos, tanto en R como en Python, y algunos de los archivos auxiliares no estaban. Se hizo el estudio de todo el código y se aplicaron algunas correcciones, así como una depuración para seleccionar solo las partes que pudieran ser utilizables en esta segunda fase.

Entendimiento del Sistema General de Regalías (SGR) y sus necesidades

Se hizo una revisión a cómo funciona el proceso de formulación de proyectos para el SGR, con el fin de entender mejor el problema y de ganar un poco de conocimiento de negocio. Adicionalmente, se tuvieron reuniones con las dos direcciones del DNP asociadas a regalías: la Dirección de Vigilancia de las Regalías (DVR) y la Dirección del Sistema General de Regalías (DSGR). El objetivo de estas reuniones fue reactivar el contacto y conocer mejor sus necesidades desde un punto de vista de analítica de datos. De estas reuniones salieron insumos valiosos para esta fase e incluso para futuras fases del proyecto.

Identificación de fuentes de información y recolección de los datos

Finalmente, en esta etapa se estudiaron las distintas fuentes de información disponibles para el desarrollo de los modelos. Por un lado, se actualizó la información correspondiente al seguimiento de avance de proyectos de regalías realizado por el Órgano Colegiado de Administración y Decisión (OCAD) y al Índice de Gestión de Proyectos de Regalías (IGPR). Esta información fue descargada desde la página web del DNP y utilizada para actualizar los datos de la base de entrenamiento del modelo.

Por otra parte, con apoyo de la Oficina Tecnologías y Sistemas de Información (OTSI) del DNP se exploró la base de datos interna que contiene información sobre los proyectos de regalías y el sistema de Metodología General Ajustada (MGA). En esta base de datos también se encuentra la información sobre las fichas de Estadísticas Básicas de Inversión (EBI) de los proyectos, las cuales tienen información textual sobre la formulación del proyecto.

Con esta información se hizo un nuevo filtro, para evaluar qué variables efectivamente están disponibles en el sistema para un proyecto en su etapa de formulación. En este filtro se perdieron algunas variables, tales como el OCAD asignado a un proyecto o la entidad ejecutora del mismo. Todos estos datos fueron sometidos a análisis exploratorio y procesamiento. También se utilizó la información disponible para construir nuevas variables, con el objetivo de tener un mayor número de características que puedan enriquecer los modelos predictivos.

Formulación de variables objetivo

El objetivo es realizar un modelo de aprendizaje supervisado, lo que requiere utilizar una variable objetivo, que es la que se intenta predecir. Esta variable debe expresar el buen o mal desempeño que ha tenido un proyecto de regalías, de manera que permita identificar si un determinado proyecto tiene problemas en su ejecución. En su primera fase, el proyecto trabajó con el IGPR, y el modelo se encargaba de predecir si para un proyecto en específico el promedio del IGPR dentro de una ventana de tiempo iba a ser mayor o menor a un valor de umbral.

Inicialmente se consideró una ventana de observación comprendida entre el último trimestre del año 2016 y el primer trimestre del 2018 (6 reportes trimestrales del IGPR de los proyectos de regalías). Sin embargo, luego de una reunión con la Dirección de Vigilancia de las Regalías se acordó utilizar todos los reportes disponibles (10 hasta el momento), por lo que la ventana de observación se amplió para cubrir desde el primer trimestre de 2016 hasta el segundo trimestre de 2018.

En este nuevo desarrollo se consideraron varias candidatas, todas relacionadas con el IGPR, para ser la variable objetivo del modelo. Se realizaron las siguientes variaciones:

- *Tipo de cálculo*: a partir de las mediciones de IGPR se hicieron dos cálculos: obtener el promedio y el valor mínimo de cada proyecto en la venta de tiempo.
- *Estado de los proyectos*: también se hicieron estos cálculos aplicando distintos filtros a los proyectos, dependiendo de su estado de contratación. En algunos casos se excluyeron los proyectos cuyo estado en la ventana de tiempo era “Sin contratar”, en otros casos solo se tomaron en cuenta solamente los proyectos cerrados y terminados, y en otros casos se utilizaron todos los proyectos sin importar su estado.

Finalmente, se tuvo en cuenta una variable extra. Dentro del cálculo del IGPR, uno de los factores que da puntos es que el proyecto no haya requerido de medidas por parte del Sistema de monitoreo, seguimiento, control y evaluación (SMSCE). Este factor fue tomado por aparte, y se construyó una variable que da a un proyecto como bueno si nunca requirió medidas del SMSCE en la ventana de observación, y como malo en caso contrario. Las variables objetivo que fueron consideradas y sus respectivas descripciones se presentan en la Tabla 1.

target_prom_all	Promedio de los valores de IGPR que tuvo el proyecto en la ventana de observación.
target_prom_contratados	Promedio de los valores de IGPR que tuvo el proyecto en la ventana de observación. Se excluyen IGPRs si en ese momento el estado del proyecto era "Sin contratar".
target_prom_finalizados	Promedio de los valores de IGPR que tuvo el proyecto en la ventana de observación. Solo se toman en cuenta IGPRs si el estado del proyecto es "Cerrado" o "Terminado".
target_min_all	Mínimo de los valores de IGPR que tuvo el proyecto en la ventana de observación.
target_min_contratados	Mínimo de los valores de IGPR que tuvo el proyecto en la ventana de observación. Se excluyen IGPRs si en ese momento el estado del proyecto era "Sin contratar".
target_min_finalizados	Mínimo de los valores de IGPR que tuvo el proyecto en la ventana de observación. Solo se toman en cuenta IGPRs si el estado del proyecto es "Cerrado" o "Terminado".
target_SMSCE	Medida binaria que indica si el proyecto ha requerido medidas del Sistema de monitoreo, seguimiento, control y evaluación (SMSCE) durante la ventana de observación.
Ventana de observación	Del último trimestre de 2016 hasta el primer trimestre de 2018.

Tabla 2: Variables objetivo que fueron consideradas

Tratamiento de variables

Una vez se tienen los datos y las variables objetivo, es necesario preparar los datos para el desarrollo del modelo. Esto requiere una serie de acciones que varía dependiendo del tipo de variable. A continuación, se muestra el procesamiento que se realizó para cada tipo de variable presente en los datos.

Variables continuas

Para las variables continuas de tipo numérico (como, por ejemplo, el valor total del proyecto) se realizaron tres operaciones. En primer lugar, se consideró el problema de la información faltante. La mayoría de los algoritmos de inteligencia artificial que trabajan con variables numéricas son muy sensibles o no responden bien cuando faltan datos. Para solucionar esto, se decidió reemplazar cualquier dato faltante con el promedio de la variable numérica a la que pertenece. Este enfoque funciona bastante bien cuando el porcentaje de datos faltantes de una variable es bajo, como era el caso en este proyecto. Adicionalmente, se decidió acotar las variables numéricas por lo alto y por lo bajo con sus percentiles del 99% y 1%, respectivamente. Esto se hace para eliminar valores atípicos, que pueden afectar análisis posteriores. Finalmente, los datos numéricos fueron normalizados para facilitar el entrenamiento de los modelos.

Variables categóricas

Para las variables categóricas (como el departamento en el cual se piensa realizar el proyecto) es necesario hacer algún procesamiento que permita llevar las categorías a números que puedan ser utilizados por los modelos de aprendizaje. En este caso se hicieron dos tipos de procesamiento. El primero de estos fue aplicar *One Hot Encoding* (OHE), el cual consiste en crear una nueva variable para cada posible valor que pueda tomar una variable categórica. La principal desventaja de este método es que el tamaño de la base puede crecer muchísimo, sobre todo si hay variables categóricas que tengan muchos posibles valores. Teniendo en cuenta esto, algunas variables categóricas con muchos posibles valores fueron agrupadas (de acuerdo a la tasa de malos de cada valor) antes de aplicar el OHE.

Adicionalmente, también se estudió la tasa de ocurrencia de la variable objetivo para cada valor de las variables categóricas. Esto permitió calcular el *weight of evidence* (WOE), una medida numérica de cada categoría que se calcula basada en la ocurrencia de eventos (proyectos malos) para cada valor de las categorías, y que fue utilizada para representar las variables categóricas como variables continuas. En este caso la presencia de datos faltantes no es un problema tan grave, pues los datos faltantes pueden ser representados como una nueva categoría.

Variables de texto

Para hacer cualquier tipo de análisis con las variables de texto, es recomendable procesarlas primero. Teniendo esto en cuenta, se les dio el siguiente tratamiento a las variables de texto:

1. Unir todos los campos de texto de cada proyecto en un solo texto.
2. Hacer una limpieza al texto unificado, lo que incluye:
    - Pasar todo el texto a minúsculas.
    - Eliminar información no textual (caracteres especiales, puntuación, símbolos).
    - Quitar del texto palabras de menos de 3 letras.
    - Quitar del texto palabras que no aporten significado, tales como artículos, pronombres y preposiciones, entre otras.
3. Una vez se tienen los textos depurados, se aplicaron técnicas de lematización a los mismos.

En el proceso de entrenamiento de los modelos se pudo observar que lematizar los textos en este caso no implicaba una mejora significativa en el desempeño de los modelos. Teniendo en cuenta esto, y los altos costos en tiempo de lematizar los textos, finalmente se decidió aplicar solo los primeros dos pasos del procesamiento de texto.

Vectorización y agrupamiento de textos

Al igual que con las variables categóricas, las variables de texto deben ser convertidas a alguna representación numérica, para poder ser aprovechadas por los modelos de aprendizaje automático. En este caso se utilizó la técnica *Doc2Vec*, que permite representar un documento entero, a nivel tanto

semántico, como un vector numérico. Por medio de esta técnica, cada texto pre procesado fue convertido a un vector de dimensión 150. Luego, estos vectores fueron agrupados por medio de métodos no supervisados. En particular, se usaron los métodos de *Birch* y *K-means*, para encontrar conjuntos de textos similares entre sí. Adicionalmente, el algoritmo de K-means fue ejecutado teniendo en cuenta dos medidas de distancia diferentes: la distancia euclidiana, que es la distancia más popular al realizar este algoritmo, y la distancia coseno, que fue tenida en cuenta porque es la que mide similitud entre vectores producidos por el algoritmo *Doc2Vec*.

Para determinar el número de grupos a definir en el *K-means* con distancia euclidiana, se hizo un barrido utilizando distintas cantidades de grupos, y se estudió la inercia (varianza interna) de los grupos generados (Figura 1). Al realizar este estudio se determinó que 16 es un número de grupos en el cual la inercia es baja, y no varía mucho más después. Por este motivo, se decidió agrupar los vectores de los textos en 16 grupos. Para los otros dos algoritmos de agrupamiento el proceso para determinar el número de grupos fue más empírico, buscando un número que asegurara una buena distribución de los proyectos entre los grupos. Las agrupaciones logradas por los 3 algoritmos se convierten en nuevas variables categóricas del modelo, y son tratadas como tal para ser utilizadas en los modelos.

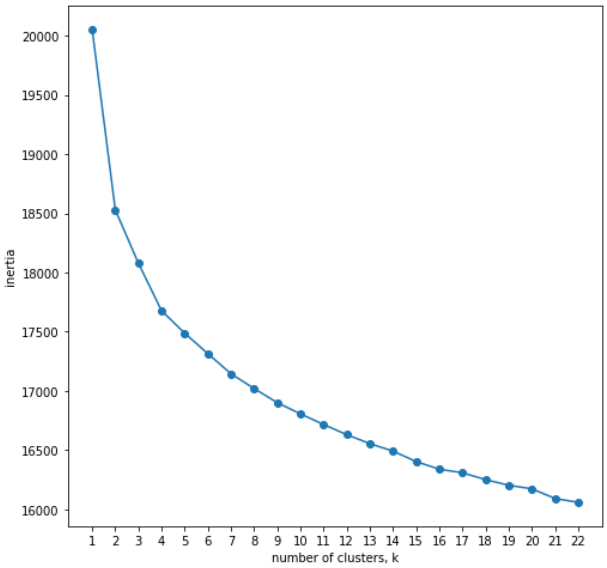

Figura 1: Barrido paramétrico de K-means para encontrar el número de grupos óptimo

Reducción de dimensionalidad y selección de variables

Luego de realizar todas las etapas anteriores, se llega a una base de modelamiento que tiene 70 variables. Es posible que no todas las variables aporten información valiosa al modelo, mientras que si pueden tener efectos negativos como introducir ruido e incrementar los costos computacionales de entrenamiento. Para reducir el número de variables y dejar solo información que describa los datos y aporte al modelo, se tuvieron en cuenta dos alternativas.

La primera alternativa fue utilizar el método de análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos. Aplicando este método, se obtuvo una base de modelamiento de 25 variables, que conservaban más del 94% de información sobre la variabilidad de los datos. La segunda alternativa fue la realizar selección de características utilizando la variable objetivo y algún método estadístico. En este caso, se realizó un análisis de la varianza (ANOVA), aplicando un *F-test* para determinar qué variables aportan más información sobre las variables objetivo. Utilizando este método se llegó a una base de modelamiento con 30 variables. Esta base fue utilizada para obtener una tercera base reduciendo el número de características a 14, al quitar variables muy similares entre sí y/o que no tenían peso en el modelo final.

Entrenamiento de modelos

Al llegar a esta etapa se tienen 3 bases de modelamiento y 8 alternativas de variable objetivo. En todos los casos, se planteó un problema de clasificación, en el cual el modelo debe distinguir entre proyectos "buenos" (sin ninguna dificultad) y proyectos "malos" (presentaron alguna dificultad en la ventana de observación). Por tal motivo, es necesario adaptar las alternativas existentes de variable objetivo para que separen proyectos entre buenos y malos. En el caso de **target_SMSCE** (referirse a la Tabla 1 para conocer la descripción de esta variable) no hay que hacer nada, pues la variable ya separa los proyectos que requirieron alguna medida del SMSCE (malos) de los que no (buenos).

Al utilizar las variables objetivos basadas en el puntaje IGPR, fue necesario definir umbrales; valores de corte para los cuales el proyecto es calificado como bueno si tiene un puntaje igual o superior, y como malo si el puntaje es más bajo. Los valores de estos umbrales fueron determinados utilizando medidas estadísticas de la distribución IGPR de los proyectos, tales como el promedio, la mediana y la desviación estándar. Según la variable objetivo y el umbral utilizados, la distribución de proyectos buenos y malos en la base puede cambiar sustancialmente. En cuanto a los modelos, se tuvieron en cuenta tres algoritmos de aprendizaje de máquina: máquinas de soporte vectorial (SVM por sus siglas en inglés), también utilizado en el proyecto de identificación de vías terciarias, y dos modelos basados en árboles de decisión.

Al utilizar las variables objetivo basadas en el puntaje IGPR, fue necesario definir umbrales; valores de corte para los cuales el proyecto es calificado como bueno si tiene un puntaje igual o superior, y como malo si el puntaje es más bajo. Los valores de estos umbrales fueron determinados utilizando medidas estadísticas de la distribución IGPR de los proyectos, tales como el promedio, la mediana y la desviación estándar. Dependiendo de la variable objetivo y el umbral utilizados, la distribución de proyectos buenos y malos en la base puede cambiar sustancialmente.

El primero de los modelos basados en árboles es el modelo de *Random Forest*, que consiste en entrenar un alto número de árboles de decisión, cada uno utilizando una muestra aleatoria de los registros y las variables de la base de modelamiento. Estos árboles individuales son ruidosos y tienen mal desempeño por sí mismos, pero una vez se unen las predicciones de todos para obtener una sola predicción, el resultado es un modelo robusto que suele mostrar buen desempeño. El segundo modelo utilizado es el *Extreme Gradient Boosting* (también conocido como *XGBoost*), y consiste en combinar una serie de clasificadores débiles (en este caso, los árboles de decisión) de manera escalonada para obtener un modelo fuerte. Teniendo en cuenta lo anterior, se entrenaron un total de **72 modelos**, producto de la combinación de 8 variables objetivo, 3 bases de modelamiento y 3 algoritmos de aprendizaje de máquina.

Validación de los modelos

Para validar los modelos, la base de modelamiento se dividió en dos grupos: uno de entrenamiento y uno de prueba. A partir de la base de entrenamiento se aplicó la técnica de *K-fold cross validation* (validación cruzada de K iteraciones), que consiste en partir la base en *K* segmentos y realizar *K* modelos, siempre utilizando *K-1* segmentos para entrenar y *1* segmento para validar. En este caso se eligió 5 como el número de segmentos a utilizar. Esta técnica de validación se combinó con una búsqueda sistemática, conocida en inglés como *grid search*, para encontrar los hiper parámetros de los modelos que logran obtener un mejor desempeño en la validación cruzada. Una vez se obtienen estos hiper parámetros, se prueba cada modelo en los datos de prueba, para validar el desempeño del modelo en datos que nunca ha visto antes.

Con respecto a las métricas utilizadas para evaluar el desempeño de los modelos, inicialmente se consideró el porcentaje de acierto de clasificación. Sin embargo, la mayoría de las variables objetivo utilizadas tenían clases bastante desbalanceadas, lo cual influía en el porcentaje de acierto final de una manera no ideal. Por ese motivo, se utilizó como medida el área bajo la curva ROC, que tiene en cuenta de mejor manera el desbalanceo entre clases. Al utilizar el área bajo la curva se pudo validar de mejor manera el desempeño de los modelos, y se observó que no estaba siendo muy bueno, por lo que se decidió cambiar el objetivo del problema: en vez de utilizar los modelos para crear un clasificador duro que etiquete los proyectos como

buenos o malos, se decidió asignar un *score*, o puntaje, a cada proyecto. Este puntaje básicamente es la probabilidad de que un proyecto sea bueno, multiplicada por mil. De esta forma, el puntaje es un valor de 0 a 1000, y un proyecto es mejor a medida que tenga un puntaje mayor. Para calificar la calidad de este puntaje, se decidió utilizar como medida de desempeño el índice de Kolmogorov–Smirnov (KS), que cuantifica la distancia entre 2 distribuciones de datos, y permite medir la capacidad de un *score* para discriminar proyectos entre buenos y malos.

Puesta en producción del modelo

Actualmente se está terminando el código para llevar los modelos desarrollados a producción. El objetivo es que, para cualquier proyecto formulado, sea posible introducir su código de identificación BPIN y a partir de esto el código sea capaz de:

- Consultar las diversas fuentes de información, y obtener las características del proyecto.
- Cargar los modelos desarrollados previamente.
- Hacer todo el preprocesamiento de las variables del proyecto.
- Aplicar el modelo de clasificación.
- Entregar un puntaje o score para el proyecto.

Todo el desarrollo hecho en esta segunda fase del proyecto quedará implementado en servidores del DNP, para que las personas involucradas en el proceso de evaluación de proyectos de regalías puedan hacer uso de estas herramientas e involucrarlas al proceso, para así hacer una evaluación más completa de los proyectos.

Resultados

Al realizar la validación de todas las posibles combinaciones de los modelos, se encontró que el modelo que mejores resultados arrojó fue hecho con:

- Variable objetivo: target_SMSCE
- Base de modelamiento: Base depurada a partir de análisis ANOVA
- Algoritmo de aprendizaje: XGBoost

Este modelo obtuvo un índice KS de 0.345 en los datos de entrenamiento y de 0.296 en los datos de prueba, al separar los puntajes del modelo en 8 grupos distintos (en la Tabla 2 se puede ver esta distribución para los conjuntos de datos de entrenamiento y prueba). En ambos casos el valor del KS es bueno e indica una efectiva separación entre proyectos buenos y malos a través de los rangos. Se puede ver en los datos de entrenamiento el grupo de proyectos con mejores puntajes (superiores a 906) tiene una tasa de aparición de proyectos con dificultades de tan solo el **5.0%,** mientras que en el grupo de peores puntajes (menores a 605) más de la mitad de los proyectos (**57.5%**) presentan dificultades. En el grupo de datos de prueba estas diferencias son un poco menores, aunque aún permiten separar de manera efectiva los proyectos por rango, presentando una tasa de proyectos malos de **8.9% y 50.8%** en el mejor y peor grupo, respectivamente.

Esta tabla de puntajes, también conocida como *scorecard*, permite establecer un criterio para definir uno o varios puntajes de corte en los puntajes, que definan el nivel de confianza de un proyecto. Este criterio debe ser definido en colaboración con las personas encargadas de evaluar los proyectos de regalías en su etapa de formulación.

Rango de score	Número de proyectos	Malos	Buenos	Tasa de malos	Tasa de buenos	KS
> 906	1211	61	1150	5.0%	95.0%	12.1%
(880.0, 906.0]	1262	109	1153	8.6%	91.4%	22.1%
(857.0, 880.0]	1206	162	1044	13.4%	86.6%	28.3%
(825.0, 857.0]	1235	194	1041	15.7%	84.3%	32.9%
(785.0, 825.0]	1265	254	1011	20.1%	79.9%	**34.5%**
(721.0, 785.0]	1239	307	932	24.8%	75.2%	32.5%
(605.0, 721.0]	1238	391	847	31.6%	68.4%	25.7%
<= 605	1241	713	528	57.5%	42.5%	0.0%

Tabla 3a: Distribución de puntajes en los datos de entrenamiento, en 8 rangos

Rango de score	Número de proyectos	Malos	Buenos	Tasa de malos	Tasa de buenos	KS
> 906	280	25	255	8.9%	91.1%	8.6%
(880.0, 906.0]	296	32	264	10.8%	89.2%	16.4%
(857.0, 880.0]	305	41	264	13.4%	86.6%	22.5%
(825.0, 857.0]	306	48	258	15.7%	84.3%	26.9%
(785.0, 825.0]	331	61	270	18.4%	81.6%	**29.6%**
(721.0, 785.0]	330	74	256	22.4%	77.6%	29.1%
(605.0, 721.0]	293	96	197	32.8%	67.2%	21.3%
<= 605	305	155	150	50.8%	49.2%	0.0%

Tabla 4b: Distribución de puntajes en los datos de prueba, en 8 rangos

El índice KS mide la separación de dos distribuciones de datos; en este caso, de las distribuciones acumuladas de proyectos buenos y malos a lo largo de los rangos de puntajes. A medida que esta separación sea mayor, más exitoso será el modelo discriminando proyectos buenos y malos. La Figura 2 muestra gráficamente las distribuciones de proyectos, tanto para los datos de entrenamiento (izquierda) como de prueba (derecha).

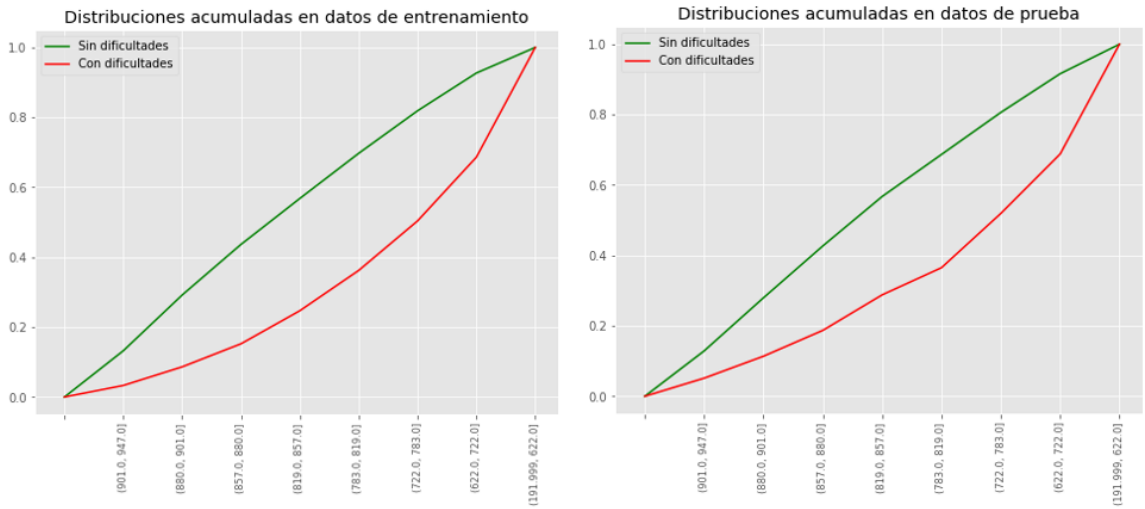

Figura 2: Distribuciones acumuladas de los proyectos buenos y malos en datos de entrenamiento y prueba

Adicionalmente, se puede observar que el peor de los 10 grupos cubre un rango de puntajes bastante amplio, llegando hasta 605. Esto hace pensar en que la distribución de los puntajes está sesgada hacia los valores más altos. Esta noción puede confirmarse al ver gráficamente la distribución de los puntajes (Figura 3). Sin embargo, esto es fácilmente solucionable si se alinea el puntaje de los proyectos para centrar su distribución, o si los criterios de selección de proyectos se definen teniendo en cuenta este sesgo. Lo importante es que los proyectos estén distribuidos de una forma más o menos uniforme a lo largo de todos los rangos; condición que se cumple en este caso.

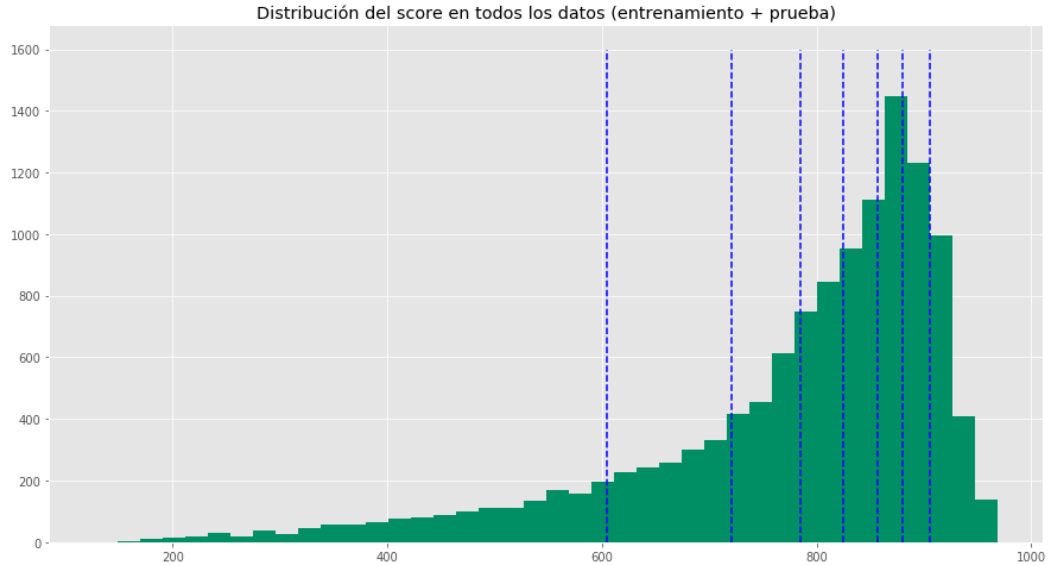

Figura 3: Distribución de los puntajes de los datos de entrenamiento + prueba

Finalmente, el algoritmo de clasificación *XGBoost* está basado en árboles de decisión, y esto permite obtener una medida de la "importancia" de las variables predictoras utilizadas. Esta importancia se calcula basada en el número de veces que cada variable fue utilizada en un nodo para tomar una decisión en alguno de los árboles del modelo, y se asume que una variable predictora es más importante a medida que es más utilizada en los nodos. En la Tabla 3 se pueden observar las importancias de las variables utilizadas en los modelos. Se puede observar que el monto y la duración del proyecto son las variables que más peso tienen en este modelo, pero que todas las variables utilizadas aportan al modelo final.

Variable	Importancia (%)
VALOR_TOTAL_scaled	22.820513
VALOR_SGR_scaled	17.179487
Longitud_proyecto_scaled	16.923077
DEPARTAMENTO_cat_woe	9.74359
kmeans_euc_cluster_woe	7.179488
SECTOR_cat_woe	4.102564
financiamiento_externo_1	4.102564
REGION_cat_woe	3.846154
Causas_indirectas_scaled	3.076923
kmeans_cos_cluster_woe	2.820513
Efectos_indirectos_scaled	2.307692
Causas_directas_scaled	2.051282
Numero_metas_scaled	2.051282
Efectos_directos_scaled	1.794872

Tabla 3: Importancia relativa de las variables

Conclusiones

1. El trabajo hecho en esta segunda fase del proyecto de regalías permitió tener en cuenta datos mucho más variados en comparación a la primera fase. Fue posible utilizar los aspectos positivos realizados anteriormente y potenciarlos con nuevos elementos para obtener mejores resultados.
2. El hecho de utilizar el modelo para devolver un puntaje en vez de una etiqueta permite también cuantificar qué tan "bueno" o "malo" va a ser un proyecto en su etapa de formulación, y esto permitirá a las áreas encargadas de la evaluación de proyectos saber qué proyectos requieren una revisión más a fondo, o qué proyectos, de ser aprobados, requerirán una labor de vigilancia más ardua. También es importante notar que esta metodología puede ser reutilizada si cambian luego las variables predictoras o la variable objetivo que se desea predecir.
3. Lo más importante es que estas herramientas puedan ser realmente útiles en el proceso de evaluar y aprobar proyectos de regalías. En este momento es importante acercarse a las direcciones del DNP encargadas de estos temas para asegurarse de que así sea (que entiendan los alcances de los modelos desarrollados y puedan utilizarlos). Es posible que surjan nuevos requerimientos para mejorar, cambiar o añadir, pero estos ya serían contemplados dentro de una fase 3 del proyecto.

Socialización

Actualmente se está trabajando en la puesta en producción de los modelos, y en la capacitación a las direcciones de regalías del DNP para su uso. Para este proceso se han adelantado reuniones con el área de tecnología del DNP para definir los requisitos técnicos y la mejor forma de implementar el proyecto en la infraestructura de la entidad. Adicionalmente, los resultados del proyecto ya han sido compartidos a la DVR, a la DDD y fueron presentados el 4 de diciembre en la Universidad del Rosario, en el evento "Corrupción y Big Data: retos y oportunidades".