Proyecto 1: Regresión

Tiziano Abraham Lopez Vargas, Jose Leandro Machaca Soloaga, Luis Renato Carbajal Cortez

Universidad de Ingeniería y Tecnología - UTEC

Cristian López del Alamo

Índice

1. Introducción

Este informe aborda el análisis y predicción de datos relacionados con el Presupuesto Institucional de Apertura (PIA) en una entidad pública. El objetivo de este proyecto es utilizar técnicas de machine learning para desarrollar un modelo de regresión adecuado que nos permita predecir el monto asignado de PIA a partir de una o diversas variables dentro del dataset estudiado. El resultado esperado es encontrar aquel modelo que nos proporcione las estimaciones más precisas que puedan llegar a ser útiles al momento de gestionar y planificar presupuestos.

En la sección 2 empezaremos haciendo una exploración y análisis del dataset que nos ayudará a tomar decisiones al momento de crear el modelo de regresión. En la sección 3 examinamos el diseño del modelo que incluye el tipo de regresión definido, las variablesque se utilizarán para hacer la predicción, las funciones de pérdida y las técnicas de regularización. En la sección 4 discutimos aspectos técnicos de la implementación del modelo. En la sección 5 presentamos los experimentos diseñados para probar la precisión del modelo implementado junto con los resultados obtenidos. En la sección 6 analizamos e interpretamos los resultados obtenidos en la experimentación. Finalmente, en la sección 7 exponemos las conclusiones finales del proyecto.

2. Dataset

2.1. Contextualización

El conjunto de datos proporcionado incluye una gran cantidad de variables que caracterizan diversas facetas del presupuesto y la ejecución financiera de una entidad pública, más específicamente, los datos usados para el entrenamiento y las pruebas hacen referencia a el Programa Nacional Plataformas de Acción para la Inclusión Social (PAIS). Varios factores considerados dentro del dataset incluyen información sobre la entidad, el programa presupuestal, las metas de la entidad, entre otros elementos que al examinar permitiría obtener un panorama completo de la distribución y ejecución del presupuesto.

La variable objetivo del análisis es MTO_PIA, que hace referencia al monto asignado de Presupuesto Institucional de Apertura. Esta variable nos da información sobre la asignación de recursos financieros, cosa que influye en la planificación de proyectos para el periodo de tiempo sobre el cual hacemos el análisis. Determinar cómo se relaciona el MTO_PIA con las variables restantes del dataset nos va a permitir identificar tendencias para predecir la asignación de presupuesto.

2.2. Análisis descriptivo

El dataset cuenta con 88 variables de las cuales 58 representan valores numéricos y 30 representan categorías, ya sea a través de texto o números. No obstante, de las 58 variables numéricas, se nos pide excluir del análisis a todas aquellas que representen montos además de MTO_PIA que es nuestra variable objetivo, dejándonos solo 6 variables numéricas. En la Tabla 1 observamos la lista completa de todas las variables que forman parte del dataset y pueden ser relevantes para el desarrollo del proyecto.

3. Metodología

3.1. Modelo

El modelo de regresión escogido es un modelo de regresión lineal multivariable. Este va a recibir como entrada una matriz de $n \times m$, donde n es el número de datos y m es el número de variables predictoras o características que va a tener cada dato, y nos va a dar como resultado un vector de tamaño n que contiene las predicciones de la variable objetivo para cada dato en la matriz. En la Figura 1 podemos observar un ejemplo de gráfica de una regresión lineal multivariable donde se busca trazar un hiperplano que mejor ajuste a los puntos en el espacio.

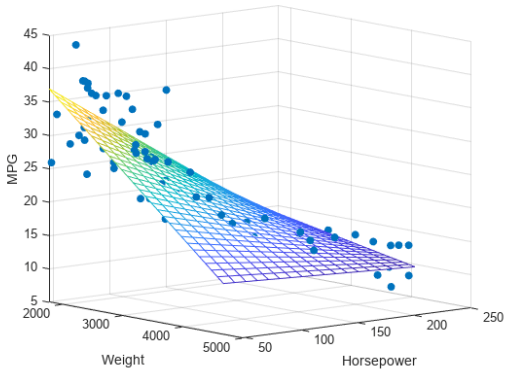

Figura 1. Ejemplo de una regresión lineal multivariable.

Un modelo de regresión lineal multivariable es una expansión del modelo lineal univariable que nos permite trabajar con más de una sola variable predictora. Otra posible alternativa es un modelo de regresión no lineal univariable, sin embargo, consideramos que al disponer de un dataset con una gran cantidad de variables es más conveniente aprovecharlas para hacer un análisis más preciso en lugar de la alternativa que sería definir una única variable que pueda ajustar al modelo de forma polinómica.

3.2. Variables predictoras

En base a la exploración del dataset realizada, seleccionamos 9 variables predictoras: TIPO_PROD_PROY, TIPO_ACT_OBRA_ACCINV, META, CANT_META_ANUAL,

Variable	Descripción	Tipo
FECHA_CORTE	Fecha en el que se genero el dataset	Categórica
ANO_EJE	Año de ejecución del presupuesto	Categórica
SECTOR	Código y nombre de Sector al que pertenece el Pliego Presupuestal	Categórica
PLIEGO	Código y nombre del pliego presupuestal al que pertenece la Entidad.	Categórica
UNIDAD_EJECUTORA	Código y nombre de la cadena institucional que identifica a una Entidad.	Categórica
SEC_EJEC	Código único que identifica a la unidad ejecutora	Categórica
PROGRAMA_PPTAL	Código y nombre de la categoría presupuestal, el cual es un criterio de clasificación del gasto presupuestal	Categórica
TIPO_PROD_PROY	Código y nombre que identifica si es proyecto o producto.	Categórica
PRODUCTO_PROYECTO	Código y nombre del producto o proyecto	Categórica
TIPO_ACT_OBRA_ACCINV	Código y nombre del tipo de actividad, acción de inversión u obra	Categórica
ACT_OBRA_ACCINV	Código y nombre de la actividad, acción de inversión u obra	Categórica
FUNCION	Código y nombre de la Función.	Categórica
DIVISION_FN	Código y nombre de División Funcional.	Categórica
GRUPO_FN	Código y nombre del Grupo Funcional.	Categórica
META	Código de la meta	Categórica
FINALIDAD	Código y nombre de la finalidad.	Categórica
UNIDAD_MEDIDA	Código y nombre de la unidad de medida	Categórica
CANT_META_ANUAL	Cantidad de meta anual establecida	Numérico
CANT_META_SEM	Cantidad de meta establecida al semestre	Numérico
AVAN_FISICO_ANUAL	Información sobre la ejecución financiera anual y la ejecución de las metas físicas programadas de los productos de los Programas	Numérico
AVAN_FISICO_SEM	Información sobre la ejecución financiera semestral y la ejecución de las metas físicas programadas de los productos de los Programas	Numérico
SEC_FUNC	Número secuencial de las metas que tiene la Unidad Ejecutora	Numérico
DEPARTAMENTO	Código y nombre del departamento del ubigeo de la meta.	Categórica
PROVINCIA	Código y nombre de la provincia del departamento del ubigeo de la meta.	Categórica
DISTRITO	Código y nombre del distrito de la provincia del departamento del ubigeo de la meta	Categórica
UBIGEO	Codigo de ubicación geografica de la meta	Categórica
FUENTE_FINANC	Código y nombre de la fuente de Financimiento que agrupa a uno o más Rubros.	Categórica
RUBRO	Código y nombre del rubro que puede utilizar la Entidad.	Categórica
CATEGORIA_GASTO	Código y nombre de la categoría de Gasto.	Categórica
TIPO_TRANSACCION	Código y nombre de la transacción, para este reporte, siempre se presentará.	Categórica
GENERICA	Código y nombre de mayor nivel de agregación de los clasificadores de gasto.	Categórica
SUBGENERICA	Código y nombre del nivel intermedio de agregación (subgenérica nivel 1) de los clasificadores de gasto.	Categórica
SUBGENERICA_DET	Código y nombre del nivel intermedio de agregación (subgenérica nivel 2) de los clasificadores de gasto.	Categórica
ESPECIFICA	Código y nombre de la específica nivel 1, identifica el detalle del gasto.	Categórica
ESPECIFICA_DET	Código y nombre de la específica nivel 2, Identifica el detalle del gasto.	Categórica
MTO_PIA	Monto asignado de Presupuesto Institucional de Apertura.	Numérico

Cuadro 1. Lista de variables a considerar para el proyecto.

CANT_META_SEM, AVAN_FISICO_ANUAL, AVAN_FISICO_SEM, SEC_FUNC y CATEGORIA_GASTO. Para la selección de estas variables se tomó en cuenta las siguientes consideraciones:

1. Ignoramos aquellas variables que solo tienen un valor único parar todos los registros del dataset.
2. La variables tiene que aportar valor al análisis, es decir, no puede ser un identificador como un nombre o código ya que limita la capacidad de generalizar.
3. En los casos donde tenemos variables que son subcategorías de otras, consideramos solo aquellas que son más específicas ya que nos aportan mayor información.

Para considerar variables categóricas dentro de la regresión las convertimos en valores numéricos que representan a cada categoría.

3.3. Funciones de pérdida

Para medir qué tan bueno es nuestro modelo haciendo las predicciones hacemos uso de una función de pérdida. Antes de definir la función de pérdida a usar, realizamos un análisis de los valores atípicos presentes en nuestras variables predictoras. Los resultados de este análisis se pueden observar en la Figura 2. Observamos que 5 de las 9 variables predictoras presentan valores atípicos. La máxima cantidad de valores atípicos es 183 para las variables CANT_META_SEM y AVAN_FISICO_SEM, representando un 12.5 % de los datos totales. Teniendo esto en cuenta, consideramos la función de pérdida minimum absolute error (MAE) ya que, a diferencia del minimum squared error (MSE), es menos sensible al ruido en los datos.

	Variable	Outliers
0	TIPO_PROD_PROY_NUM	6
1	TIPO_ACT_OBRA_ACCINV_NUM	6
2	META	0
3	CANT_META_ANUAL	0
4	CANT_META_SEM	183
5	AVAN_FISICO_ANUAL	0
6	AVAN_FISICO_SEM	183
7	SEC_FUNC	0
8	CATEGORIA_GASTO_NUM	36

Figura 2. Conteo de valores atípicos por variable predictora.

Asimismo, decidimos implementar tres funciones de pérdida MAE: sin regularización, con regularización L1 y con regularización L2. La regularización tiene como objetivo reducir la complejidad del modelo para evitar casos de overfitting, de modo que el modelo tenga mejor capacidad de generalización. Considerar tres funciones diferentes de pérdida nos va a permitir comparar el comportamiento de cada una con nuestro modelo para encontrar aquel método que nos da los mejores resultados.

4. Implementación

La implementación completa, junto con todos los archivos necesarios para su ejecución, se encuentran en el siguiente repositorio de GitHub: https://github.com/JLeandroJM/Project1_MachineL.

Los archivos train_final.csv y test_final.csv contienen los datasets filtrados para el entrenamiento y las pruebas respectivamente, solo contienen las variables que definimos previamente para el modelo. El archivo metaData.csv contiene las descripciones de todas las columnas originales del dataset. Los notebooks contienen la implementación de los experimentos y las funciones usadas en estos.

5. Experimentación

5.1. Diseño de experimentos

El modelo implementado fue entrenado usando el 70 % de los datos en el dataset de entrenamiento, mientras que el 30 % restante fue usado para las pruebas. Esto se hizo porque, a pesar de que tenemos un dataset de pruebas, este no contiene los valores reales de la variable objetivo y es usado solo para la competencia de Kaggle asociada al proyecto.

Si bien se llegó a implementar los tres tipos de funciones de pérdida definidos, debido a errores durante la ejecución no pudimos realizar los experimentos con todas estas. El experimento principal y del cual se obtuvieron resultados es el realizado sin regularización.

Para todos los experimentos se consideró $\alpha = 0.01$, $\lambda = 0.01$ y $e = 10000$, donde α es el learning rate, λ es el parámetro regularizador y e es el número de epochs

5.2. Resultados

En la Figura 4 observamos una gráfica del error en función al número de epochs durante el entrenamiento en el experimento sin regularización. En la Figura **??** observamos una comparación visual de los valores que predijo el modelo sobre la data de prueba y los valores reales.

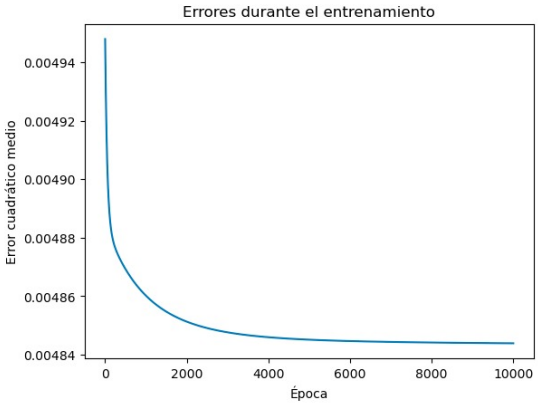

Figura 3. Gráfica de error contra número de epochs para entrenamiento sin regularización.

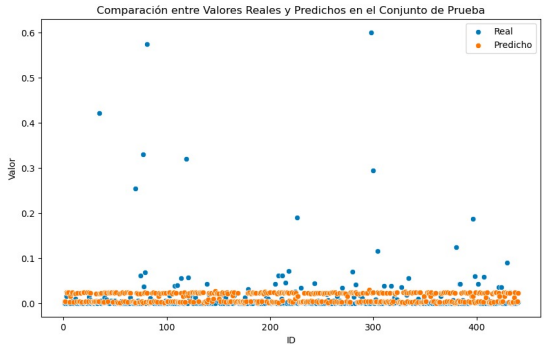

Figura 4. Comparación de los valores generados por el modelo con la data real de testeo.

6. Discusión

Para el experimento sin regularización, la Figura 4 demuestra que conforme aumenta el número de epochs el valor de error disminuye, lo que indica que el modelo mejora su capacidad de predicción mientras más iteraciones pasa sobre el dataset de entrenamiento. Asimismo, esta reducción en el error se da con mayor intensidad en los primeros epochs hasta que empieza a converger. El mínimo valor de error registrado es 0.004843875100194083.

7. Conclusiones

La realización de este proyecto nos llevó a los siguientes resultados:

- Logramos implementar un modelo de regresión lineal multivariable para la predicción del monto de Presupuesto Institucional de Apertura.
- En los experimentos registramos un error mínimo de 0.004843875100194083 durante el entrenamiento con la función de pérdida sin regularización.

Asimismo, contamos con las siguientes limitaciones durante el desarrollo:

- Las opciones de modelos de regresión a implementar se limitaron a aquellos vistos en la teoría del curso: lineal univariable, lineal multivariable y no lineal univariable.
- No contamos con los valores reales de la variables objetivo para el dataset de pruebas, por lo que para visualizar los resultados de forma local se tuvo que particionar el dataset de entrenamiento. Esto hizo que se redujera la cantidad de datos que alimentaron al modelo lo cual pudo afectar su desempeño en la sección 5.2.

Para finalizar, recopilamos algunas recomendaciones para futuros trabajos:

- Finalizar las implementaciones de las funciones de pérdida con regularización y comparar el desempeño con la función de pérdida sin regularización.
- Implementar múltiples modelos de regresión y realizar los mismos experimentos sobre estos para elegir el modelo adecuado en base a los mejores resultados y corroborar si la justificación del modelo seleccionado inicialmente tiene fundamentos prácticos.
- Repetir los experimentos múltiples veces usando diferentes particiones aleatorias del dataset para analizar la varianza entre resultados.