Equipo

"Los Internacionales" está integrado por Verónica Bianchini, Federico Baiocco, Federico Aballay, Micaela Marzoa, Cristian Vergara y Andrea Campetella.

Desafío elegido: Desafío #1

Comentarios: Elegimos analizar los resultados de los exámenes aprender con el objetivo de entender mejor las razones detrás de los diversos desempeños escolares en Argentina.
La selección de este desafío particularmente nos resultaba desafiante al permitirnos explotar la funcionalidades de la ciencia de datos para comprender mejor las variables que afectan el nivel educativo nacional.

Un aprendizaje es un fenómeno complejo atravesado por múltiples factores, muchos de los cuales son difíciles de aprehender en métricas estandarizadas. Sin embargo, nos resulta interesante poder intentar explicar esta realidad utilizando entre otras cosas la riqueza de los formularios complementarios que acompañan las evaluaciones respondidos por los principales actores de la institución educativa.

Una vez entregados los datasets tuvimos que redireccionar nuestro enfoque y tomamos las primeras decisiones en relación a los datos. Optamos por enfocar el análisis a nivel alumno de primaria ya que la cantidad de datos disponibles era significativamente mayor en el nivel primario. También decidimos hacer foco en el desempeño en lengua, para acotar el universo de análisis. También decidimos excluir los resultados de 2013 por la discontinuidad anual, la metodología diversa con la que se tomaron las pruebas y las reformas en políticas educativas que se efectuaron tras ese periodo: siendo todas estas variables que no podemos plasmar en el modelo, excluir 2013 nos permite evitar "ruido" en el mismo. . Conservamos la idea de reconstruir el contexto de los resultados pero desde el aspecto de gestión de recursos económicos y del personal, con este criterio hicimos la selección de variables.

Qué datos usaron. Descripción de variables y fuente usada

Con el criterio mencionado anteriormente las variables que consideramos fueron índice socioeconómico, nivel de icse, acceso a internet, tasa con nbi, cargos específicos, gasto por alumno. También seleccionamos datos que especificara particularidades de la escuela como ámbito, gestión, dependencia, provincia, nivel, año, sexo, si era técnica o no. Toda la información que utilizamos fue del repositorio de Argx Edu excepto la variable de gasto por alumno. Para tener un valor más preciso convertimos este valor a dólares utilizando el tipo de cambio por promedio anual del Banco Central[1]. También utilizamos los índices de canasta basica sacados del Indec[2] para ponderar los costos de vida relativos en cada región.

Nos hubiera gustado incluir datos 2018, pero el tiempo no nos permitió ir a la búsqueda de todos los valores pertinentes a este año para todas nuestras variables.

Nuestra primera aproximación al problema también incluyó las variables respecto a la posesión de tecnologías (pc, celular, etc), pero debido a la alta cantidad de valores faltantes en estas variables, terminamos excluyendolas.

Cómo procesan la información. Incluyendo el criterio para el análisis que hayan hecho y referencias a material que han consultado.

Importamos las librerías a usar y los datos necesarios para armar nuestra base de datos. Los datos que necesitamos se encuentran alojados en diversos archivos, por lo cual el segundo paso será combinarlos en un dataset para el análisis.

[1] Banco central
[2] indec

Luego, analizamos la forma de nuestro dataset: que información tenemos, cuáles son las variables, su tipo, número de casos. Además, identificamos la cantidad de casos nulos o no responde.

Para el tratamiento de datos faltantes, en casos en que el porcentaje de datos faltantes no es muy grande (hasta un 20% aproximadamente) decidimos completarlos con valores como su media, moda o mediana. En otros casos, como por ejemplo en la variable sexo, directamente decidimos descartar la columna ya que la mayor parte de sus valores son faltantes. En el caso de datos faltantes en nuestras variables de destino, decidimos excluirlos del análisis para no hacer asumpciones que sesguen el modelo. Como mejora, pensamos en completar estos datos faltantes haciendo un análisis multivariable, es decir, teniendo en cuenta los valores de otras variables para completar faltantes.

Una vez que procesamos los datos completando los nulos, correremos análisis descriptivos de nuestras variable para entender media, mediana, moda, desvío, máx y min. También generamos histogramas de nuestras variables y buscamos correlaciones entre variables del dataset.

Finalmente, excluimos variables del data set final, como se puede ver en los notebooks, convertimos las variables necesarias en dummy y aplicaremos ponderadores.

Una vez que nuestros datos estaban limpios y codificados a valores numéricos (ya sea con dummies o valores numéricos), comenzamos a modelar nuestras variables para explicar el desempeño de lengua de los alumnos nivel primario. Para este primer análisis decidimos enfocarnos en el nivel de desempeño satisfactorio. Hacer el preprocesamiento de está forma no nos resultó. Por lo tanto pre procesamos los datos de nuevo de otra manera.

Con los datos preprocesados en el segundo análisis, elegimos arrancar con un random forest simple, sin mucha optimización de hiperparámetros, para poder encontrar sobre todo qué importancia tiene cada una de las features que seleccionamos. El modelo utilizando estas variables tiene un poder explicativo bajo (0.386), sin embargo es una buena primera

aproximación y nos permitirá entender la importancia de nuestras variables en la predicción del resultado.

Luego cambiamos el enfoque y comenzamos a seleccionar distintas variables y procesarlas de distintas maneras para ver si encontrabamos alguna o algunas variables que fueran clave al momento de predecir. Por lo general, las variables relacionadas a nivel socioeconómico afectaban más que las otras.

Qué herramientas utilizaron. Breve descripción de las herramientas. (software) utilizado.

Para procesar los datos utilizamos Python con sus librerías más conocidas para análisis de datos:

- Pandas: Para trabajar los datasets, hacer joins, transformar y en algunos casos visualizar tablas.
- Numpy: Para trabajar con matrices.
- Sklearn: Para los modelos predictivos (Random Forest, Logistic regression y KMeans) y para obtener las métricas de evaluación de los modelos.
- Seaborn y matplotlib para graficar.

Resumen de desafíos que encontraron y cómo los resolvieron

A medida que intentábamos enriquecer la cruza de variables se nos presentaros algunos desafíos. El uso de la ponderación en variables categóricas no nos resultaba claro, comprendemos la función de estos valores pero no su aplicación. La asistencia por parte de los mentores fue muy útil para poder continuar. La cantidad de datos proporcionados también nos dificulto seleccionar un criterio para el analisi por lo que los informes estadísticos ya realizados y publicados en la página de Argx Edu fue un guia que nos sirvió para la toma de ese tipo de decisiones. Otras variables que consideramos pertinentes para nuestro análisis no las pudimos incluir por falta de información. Los salarios y gasto objeto solo tenemos información hasta el 2016 y cantidad de alumnos por clase tampoco pudimos conseguir esa información ya que sólo proporcionaba la

cantidad por curso. Algunas de las respuestas más interesantes del cuestionario, no contaban con los ids necesarios para ser agregadas al modelo.

Además, nos encontramos en algunas etapas con dificultades para entrenar modelos predictivos debido a el desbalanceo de clases. Para esto, probamos primero balancear las clases haciendo undersampling de la clase mayoritaria. Esto mejoró las predicciones de nuestro modelo, pero sobre una muestra muy pequeña que probablemente no represente a toda la población. En otra etapa, probamos agrupar los niveles de desempeño en 2 clases: una para el nivel de desempeño por debajo del básico o básico y otra para el nivel satisfactorio o avanzado. Esto también resultó en un modelo más preciso que los anteriores, pero no nos fue suficiente para encontrar los patrones que buscamos.

Resumen de conclusiones y ultimos analisis

No pudimos predecir el desempeño del total de los alumnos con las variables que tenemos y la manera en la que las procesamos hasta el momento usando random forests. Además, sin importar si aplicamos o no los ponderadores, obtenemos resultados muy similares. Proseguimos a analizar los datos y tratamos de buscar clusters. Hicimos un análisis de clusters para ver qué similitudes encontrabamos entre los distintos grupos que se forman. Quisimos ver por ejemplo, si alumnos que entran en un mismo cluster, tienen niveles de desempeño similares. Buscamos la cantidad de clusters que mejor separe estos datos. Graficamos la distancia media (inercia) en función al número de clusters. Según nuestro criterio, no encontramos un número de clusters "optimo" con estas features. No vemos un "codo" en donde podamos decidir que ese es el número óptimo de K. "Tomamos otro subconjunto de features y tratamos de separar en clusters teniendo en cuenta estas variables. Analizamos la inercia con este conjunto de features para ver si podemos distinguir mejor los clusters. Creemos que se podría dividir en 4/5 clusters. Tomamos 5 en un principio. Todos los alumnos de este cluster repitieron primaria. La mayoría son de un nivel socioeconómico medio. En el cluster 1 pudimos identificar como tendencia que los alumnos repitieron primaria. No encontramos tendencias en el nivel de desempeño. En el cluster 3, tenemos a los alumnos que no repitieron primaria y que son de un nivel socioeconómico bajo. Estos alumnos del cluster 4 tienen sobre todo nivel de desempeño bajo. Los alumnos del cluster

5 en general no repitieron primaria y tienen un nivel de desempeño por debajo del básico. Todos los cluster tienen alumnos de todos los niveles de desempeño. En general, están agrupados teniendo en cuenta si repitieron primaria y su nivel socioeconómico. Esto nos parece poco explicativo, ya que no mejora el resultado obtenido con nuestro modelo benchmark (random forest).

Nuestra conclusión general de estos primero modelos generales es que es difícil predecir la performance de la totalidad de los alumnos encuestados en base a las variables existentes en nuestro modelo. Además, las clases están desbalanceadas y esto afectó mucho a la performance de los modelos, en algunos casos probamos balancearlos y esto mejoró los resultados, pero sobre muestras muy chicas que probablemente no sean representativas de toda la población.

Antes de abrir teorías sobre cuáles son los datos faltantes para poder explicar dicha performace, decidimos intentar un modelo con un universo de análisis más acotado. Retomando nuestros primeros resultados del análisis exploratorio, podría haber una relación entre el nivel socioeconómico y las competencias en lengua.

Sin embargo, esta relación dista de ser deterministica: hay alumnos de nivel socieconómico bajo que logran niveles satisfactorios y avanzados. Nuestro interés es entonces comprender qué factores influyen en esto.Comenzamos a analizar a los alumnos que pertenecen a un índice socioeconómico bajo, con el objetivo de encontrar cuales son los factores que influyen en este nivel socioeconómico, para que un alumno alcance el nivel. Encontramos diferencias en el porcentaje de alumnos con indice socioeconómico bajo que obtienen un nivel de desempeño alto y el porcentaje de alumnos de un indice socioeconómico superior que obtienen el mismo resultado. En nivel socioeconómico bajo solo un 20.19% de alumnos alcanzar el nivel de desempeño avanzado, mientras que en niveles socioeconómicos superiores, el nivel de desempeño avanzado es el más común con un 38.28%". Los datos están desbalanceados por lo que probamos usar undersampling para balancear el dataset. Al hacer esto las predicciones mejoraron pero de todas formas estamos trabajando sobre una muestra muy reducida y los resultados no nos convencen. Finalmente, para tener un dataset más balanceado también probamos dividir el nivel de desempeño en 2 categorías (basico con por debajo del básico,

avanzado con satisfactorio). Obtuvimos mejores resultados pero todavía no es suficiente para definir las características que distinguen a alumnos de un nivel socioeconómico bajo.

Como proximos pasos queremos entender mejor las implicaciones de ciertas decisiones tomadas como el manejo de casos nulos.

También nos gustaría poder agregar datos como las variables pedagógicas de la encuesta (ejemplo las relacionadas al bulling o atención en clase).

Creemos que agregando estas variable lograremos una mejor comprensión de los factores impactando el desempeño.