FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1 - Conclusiones

Grupo 21 - 'df' is not defined

[7506/9558] Organizacion de Datos
Primer cuatrimestre de 2023

Alumno	Padron	Email
Martin Pata Fraile de Manterola	106226	mpata@fi.uba.ar
Tobias Emilio Serpa	108266	tserpa@fi.uba.ar
Juan Francisco Cuevas	107963	jcuevas@fi.uba.ar

1. Introducción

En este proyecto, nos enfrentamos como grupo a un desafío real de ciencia de datos, con el objetivo de realizar todas las etapas del proceso y resolver el problema planteasdo aplicando los conceptos aprendidos a lo largo del curso. La esencia del proyecto consiste en aplicar las diferentes técnicas de análisis exploratorio y preprocesamiento de datos, así como entrenar modelos de clasificación para generar predicciones respecto del objetivo fijado por la cátedra.

2. Checkpoint 1

En esta sección comenzamos a familiarizarnos con el dataset, conociendo los tipos y el significado de cada variable, como así también la relación de las mismas con el target, y entre ellas. Estas relaciones fueron visualizadas gracias a los distintos gráficos y coeficientes calculados, sugeridos por la cátedra. Gracias a toda esta información, pudimos observar que exisían diferentes variables que no mostraban una relación o modelo útil para generar predicciones respecto del target. También generamos nuevas variables a partir de relacionar variables simples, en busqueda de buenos patrones para la predicción. Además, anallizamos todos los valores atipicos y los datos faltantes de las variables, analizando en profundidad cada caso,buscando tomar la mejor decisión para solucionar estos inconvenientes. Consideramos que nos hubiese gustado continuar un tiempo mas el analisis, para así generar mas variables multirelacionadas, las cuales suelen brindar mu buenos parámetros para la predicción.

3. Checkpoint 2

Para este punto, comenzamos a generar nuestros primeros modelos de clasificación, centrándonos en los árboles de decisión. En primer lugar, analizamos cual era el mejor método de encoding para las variables categóricas, tomando como método final el One Hot Encoding. Podemos ver que OHE nos da un mejor puntaje y aunque duplica las columnas reduciendo un poco la importacia de algunas otras variables. A continuación, probamos los distintos tipos de arbloes, e hicimios una busqueda exhaustiva de los mejores hiperparametros (criterio, maxima profundidad, poda, etc) , consiguiendo un modelo final con el cual estamos muy satisfechos.

4. Checkpoint 3

En esta etapa, generamos distintos modelos como: KNN (con sus diferentes variaciones), Random forest, SVM (con todos sus kernels) y los distintos tipos de ensambles hibiridos (generados utilizando modelos genericos, o nuestros modelos creados con anterioridad). Para cada uno de estos, realizamos una busqueda de sus mejores hiperparametros, pero nos hubiese gustado experimentar mas para obtener mejores resultados, no logramos esto ya que nos vimos fuertemente limitados por los recursos computacionales, ya que cada uno de estos algoritmos consumia demasiados recursos. A cada uno de estos modelos, le analizamos sus métricas en busqueda del mejor modelo, el cual fue el generado por XGBoost.

5. Checkpoint 4

Ya en la última etapa, nos dedicamos a experimentar con redes neuronales. Comenzamos normalizando los datos, requerimiento excluyente para la creación de redes neuronales, debido a su sensibilidad a los cambios de escala. Creamos nuestras primeras redes, probando las distintas funciones de activaciones, o que no posean funciones, como así tambien jugamos con la cantidad de capas. Debido al gran consumo de recursos de estos algoritmos, no logramos una buena experimentación con lo que son los distintos tipos de regularización y optimización.

6. Conclusiones Generales

Como era de esperarse, a medida que avanzaban los checkpoints y utlizabamos métodos mas potenes/eficientes, fuimos obteniendo mejores modelos de clasificacion respecto de nuestro target. Empezamos con clasificadores mas sencillos como pueden ser los arboles de decision donde obtuvimos nuestros primeros modelos, los cuales no daban las mejores predicciones pero sabiamos que a medida que los modelos sean mas complejos, nuestras predicciones iban a mejorar. Esta mejora se pudo observar cuando empezamos con los ensambles, los primeros ensambles no daban una mejora altisima pero a medida que ibamos probando ensambles mas potentes y que claramente tenian un gasto computacional mayor (es decir, que tenian mas tiempo de ejecucion), pudimos notar una mejoria. Este cambio se noto demsiado a la hora de probar XGBoost, el cual dio una prediccion muy buena, siendo este modelo el que mejor predijo. Por ultimo probamos las redes neuronales, las cuales parecian mas complejas que cualquier otro clasificador y por ende supusimos que podrian ser las que nos den las mejores predicciones. Pero no fue asi, XGBoost siguio siendo nuestro modelo fuerte, y las predicciones hechas por redes variando las capas y los hiperparametros dieron practicamente las mismas predicciones, sin tener tanta variedad como se pudo notar en otros clasificadores.

7. Limitaciones

Consideramos que habría sido interesante modelar problemas de regresión. Y además nos hubiese gustado tener ejercicios mas práctivos/visuales, un ejemplo sería reducir la dimensionalidad de una imagen, simulando una compresión. Fuera de esto, consideramos que el trabajo nos llevó a abarcar la gran mayoría de temas enseñados por la cátedra y estamos conformes.