Memoria

1.Tema y datos obtenidos

El objeto del presente proyecto de Machine Learning es obtener un modelo de predicción de precios de billetes de tren en España.

Para ello vamos a utilizar un dataset obtenido de datamarket.es, que contiene billetes de tren con todo su árbol de precios y plazas disponibles para las principales rutas en España. Las características de este dataset son las siguientes:

- Frecuencia de actualización: actualizado cada 4h
- Volumen estimado: 15.000 registros diarios
- Histórico: desde 2018 en adelante

2. Preprocesado

En primer lugar y para una mejor interpretabilidad y manejo de la información para hacer un modelo más compacto y manejable:

- Se han eliminado ciudades duplicadas, ya que las que tienen tilde aparecían escrita con ella y sin ella.
- Se han agrupado los tipos de trenes, porque se nombraban de diferente forma, siendo de la misma categoría.
- Se han reducido los tipos de vagones.
- Se ha aplicado una técnica de encoding para transformar las variables categóricas y así poder trabajar sobre ellas. Concretamente LabelEncoder.
- Se han creado nuevas columnas a partir de la columna que contenía las fechas y hora de salida. Lo cual hace más robusto el modelo con información muy útil y relevante a la hora de viajar.
- Se han eliminado columnas.
    - La que mostraba la compañía carecía de sentido ya que había un valor único que era: Renfe.
    - Además, una vez transformadas el resto de columnas de variables categóricas de tipo “object” a variables numéricas, se ha prescindido de las originales.

Estos primeros pasos han sido muy acertados y fundamentales para la obtención de los resultados. Todo esto se recoge encapsulado en una única función de preprocesado guardada en un archivo “.py” por dos motivos, porque se deben repetir para cada modelo y porque deja más limpio el código.

3. Modelos

Para la elaboración de los modelos he decidido realizarlos en jupiter.notebook diferentes ya que no quería que la generación, transformación y entrenamiento de un modelo pudiera influir de algún modo en los sucesivos. Además todo queda más limpio y ordenado y es más sencillo a la hora de revisar los procesos, ya que un jupiter.notebook muy largo y con información muy variada puede resultar muy farragoso.

El modelo de regresión lineal ha sido la primera opción, ya que es el más sencillo de entender y fácil de realizar. No ha dado buenos resultados ya que la distribución de las variables son algo más complejas. R2 score igual a 0.356.

Con la regresión polinómica se ha incrementado considerablemente el grado de acierto de la predicción. He elaborado tres modelos de este tipo de grado 2, grado 3 y grado 4. Siendo el de grado 4 el que mejor se ha comportado con los datos empleados, no obstante, su entrenamiento fue ya bastante pesado llegando a tardar más de 40 minutos para ello. R2 score igual a 0.874. En este punto para guardar el modelo se hace en dos partes: la primera se guarda la transformación polinómica y otro modelo guardado con la regresión.

El modelo de árbol de decisión ha sido ya un salto cuantitavo en lo que se refiere a la métrica de predicción, ha logrado un R2 score igual a 0.978. Un dato muy bueno.

Como teníamos la gran mejora con el árbol de decisión hemos probado un RandomForest de regresión y tratar de mejorar el porcentaje de acierto con 30 Y 50 árboles diferentes. Y ha vuelto a subir el R2 score a 0.984 en los dos casos. Posteriormente he realizado una prueba para identificar las variables más influyentes, tras ello y al valorar que la menos influyente podría eliminarla para quitar ruido al modelo, me percaté que así lo empeoraba.

También he realizado un modelo de SVM de regresión que no planteó mejora alguna. Después he tratado de ajustarlo mediante un Gridsearch, el cual ha sido una mala decisión ya que mi dataset es muy pesado para realizar una comparación de todas las combinaciones posibles indicadas como parámetros. Tras 5296 minutos de ejecución decidí pararlo y descartar este modelo.

Por último, probé un auto-ML que en ningún caso mejoraba mis métricas.

4. Presentación

Para la exposición al público objetivo he elegido un formato de diapositivas mediante Powerpoint, en el que se abordan las siguientes cuestiones:

- ¿Qué pretendes resolver con el modelo de ML? Aportar información para reestructurar y modernizar la movilidad de los empleados de una empresa.

- ¿Qué solución aportas con el modelo de ML? El precio de los billetes de tren.

- ¿Qué resultados has obtenido? Una previsión de precios con un porcentaje de 0.98 de precisión.

- Modelos de regresión lineal, modelos regresión polinómica, modelos de árboles de decisión y modelos de SVM.

- ¿Cuáles han sido las variables de mayor impacto? La duración de trayecto, la tarifa aplicada y el tipo de tren.

- ¿Qué decisiones o acciones te permiten llevar a cabo tu modelo? Todas las de preprocesado expuestas más arriba.

Enlaces de interés

https://www.kaggle.com/datamarket/viajes-en-tren