**Trabajo Final**

Una importante red de gestión inmobiliaria online precisa desarrollar un modelo predictivo que permita predecir el precio al que venderán las propiedades que ofrecen , para así optimizar sus acción comercial e incrementar sus ventas. El objetivo de este trabajo final es la presentación completa de ese modelo, incluyendo para ello una presentación final, incluyendo un anexo con un informe de ingeniería , en el que se consignarán todos los recursos y material adicional utilizados para su desarrollo tecnológico.

El archivo se denomina **ventas.csv,** contiene 611393 ventas históricas concretadas en Argentina y sus campos se indican a continuación.

**Variables**

1. **Fecha -** Fecha de operación
2. **lat -** Latitud.
3. **lon -** Longitud.
4. **l1 -** Nivel administrativo 1: país.
5. **l2 -** Nivel administrativo 2: usualmente provincia.
6. **l3 -** Nivel administrativo 3: usualmente ciudad.
7. **l4 -** Nivel administrativo 4: usualmente barrio.
8. **Operación -** Tipo de operación (Venta, Alquiler).
9. **Tipo -** Tipo de propiedad (Casa, Departamento, PH).
10. **Ambientes -** Cantidad de ambientes.
11. **Habitaciones -** Cantidad de dormitorios.
12. **Baños -** Cantidad de baños.
13. **Superficie\_total -** Superficie total en m².
14. **Superficie\_cubierta-** Superficie cubierta en m².
15. **Moneda -** Moneda del precio publicado.
16. **Título -** Título del anuncio.
17. **Descripción -** Descripción del anuncio.
18. ***Variable de salida (objetivo):***   **Precio -** Precio publicado en el anuncio.

***Nota:*** hay varios datos faltantes en algunas variables. Estos valores faltantes pueden tratarse como un nivel particular de una variable categórica o procesarse mediante técnicas de eliminación o imputación.

**Modelo de Predicción**

Se deberá presentar los procesos y las estrategias implementadas, tanto de entrenamiento como de validación/selección de hiperparametros, a partir de las cuáles se arribe al modelo de mejor performance predictiva. Sólo para el modelo final seleccionado se deberá presentar el siguiente informe:

1. **Algoritmo utilizado:**
   1. Diseño del entrenamiento
      1. Ingeniería de atributos: conversión de variables, procesamiento,normalización/estandarización, imputaciones, selección de variables, componentes ppales, etc.)
      2. Determinación de Set de entrenamiento y Set de validación. Validación cruzada.
   2. Modelo predictivo utilizado y sus parámetros e hiper parámetros.
   3. Justificación de la selección de hiper parámetros utilizados.
2. **Performance del modelo final:**
   1. R2 Ajustado
   2. CME (Cuadrado Medio Esperado)
   3. Justificación e interpretación de los resultados

**Forma de entrega**

Se enviará a [f](mailto:federico.cardoso.e@gmail.com)cardoso@fi.uba.ar un directorio comprimido al que llamarán TP\_FIUBA\_Regresion\_Grupo\_Nro”, siendo “Nro” el número de grupo.

El contenido del directorio será el siguiente:

1. Presentación del trabajo. Archivo Word. Se deberá presentar un informe profesional en el que se expliquen los procesos llevados a cabo, su fundamentación y los resultados a los que se arribe.
2. Archivo completo de desarrollo del modelo en formato “.py” . Se considerará fundamental la presentación de las funciones y métodos, la definición de variables locales y globales utilizadas y la claridad en la documentación del código (comentarios utilizando el símbolo # antecediendo al texto explicativo)
3. Script o función de python que ejecute las predicciones sobre nuevos datos (a utilizarse para la prueba a realizarse en el siguiente punto, Calificación y “Competencia”)

**Calificación y “Competencia”**

El trabajo será calificado en función de la presentación, justificación y resultados de los puntos I y II del apartado “Modelo de Predicción”. A su vez se ha reservado un set de prueba externo de 68146

registros (extraído del dataset original) que no se incluyó en la entrega del material del presente trabajo, sobre el cual se correrá cada uno de los modelos presentados por los grupos, para luego determinar cuál de ellos es el de mejor performance. Este formato del tipo “competencia” es utilizado en la industria para el desarrollo comunitario de modelos, siendo su paradigma las competencias de la plataforma online “Kaggle”, entorno en el que se han desarrollado siguiendo esta metodología variados reconocidos modelos de machine learning, destacándose el sistema de recomendación de películas de Netflix.

La medida de performance utilizada para calificar y posicionar a los trabajos a los trabajos será: **CME de validación.**

**IMPORTANTE:**

Tener en cuenta que el script final a entregar, deberá ser capaz de:

* Incorporar los datos de testeo final que están separados y el modelo nunca vió.
* Transformarlos con las mismas operaciones que fueron realizadas sobre los datos del entrenamiento.
* Predecir el valor del precio de los inmuebles para cada observación.

Los ganadores recibirán importantes premios y reconocimientos!