Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2115 – Programación como Herramienta para la Ingeniería (I/2021)

Actividad 4

Objetivos

- Aplicar los contenidos de análisis de datos con Python para procesar y visualizar información, y realizar predicciones.

Entrega

- **Lenguaje a utilizar:** Python 3.6 o superior
- **Lugar:** repositorio privado en GitHub. Recuerde incluir todo en una carpeta de nombre **A4**.
- **Entrega: lunes 14 de junio a las 18:30 hrs.**
- **Formato de entrega:** archivo Python Notebook (**A4.ipynb**) y archivo Python (**A4.py**) con la solución de este enunciado. Los archivos deben estar ubicados en la carpeta **A4**. No se debe subir ningún otro archivo a la carpeta. Utilice múltiples celdas de texto y código para facilitar la revisión de su programa.
- **NO SE ADMITEN ENTREGAS FUERA DE PLAZO**
- **Entregas con errores de sintaxis y/o que generen excepciones serán calificadas con nota 1.0.**

Introducción

Con el fin de evaluar los contenidos de análisis de datos con Python, en esta actividad deberá realizar una serie de procesamientos y visualizaciones de datos para finalmente predecir las ventas de una serie de tiendas de la cadena Rossmann.

Descripción de los datos

La fuente primaria de datos para será un subconjunto del set "Rossmann Store Sales", disponible en el sitio del curso. El set contiene información sobre las ventas de tiendas de la cadena Rossmann. Específicamente, el archivo `data.csv`: contiene alrededor de doscientos mil registros de ventas para distintos días, donde cada uno incluye 8 características y 1 variable a predecir (`Sales`).

IMPORTANTE

Para cumplir las misiones de esta actividad, es su responsabilidad explorar inicialmente el contenido de los archivos y familiarizarse con el formato en que está almacenada la información.
Recuerde además codificar numéricamente los valores de las columnas categóricas y normalizar las numéricas cuando corresponda. La actividad no considera puntaje por hacer esto, pero sí descuentos cuando no es realizado o es realizado en una variable o momento incorrecto.

Misión 1: conociendo e importando los datos

Importe los datos contenidos en el archivo `data.csv` a un `DataFrame`. Visualice cada variables y consulte algunos estadísticos generales con los métodos revisados en clases, tanto para variables numéricas como categóricas. Finalmente, identifique la cantidad de datos faltantes para cada variable y comente brevemente sobre los estadísticos observados. **(1 pto.)**

Misión 2: limpieza y depuración

Analice la existencia de valores por incompletos para las variables. Complete o elimine registros en base a la naturaleza y distribución de cada variable y justifique sus decisiones, utilizando visualizaciones y/o argumentos analíticos. Finalmente, analice la existencia de *outliers*, corrigiéndolos en caso de ser necesario. Para ninguno de los análisis anteriores, considere la variable a predecir. **(2 ptos.)**

Misión 3: entrenamiento de modelos

Entrene 3 regresores para predecir el valor de la variables `Sales`, siendo 1 de estos una regresión lineal. Utilice 2 conjuntos de prueba distintos: i) el tercio **más reciente** de los registros y ii) un tercio elegido aleatoriamente. Compare el rendimiento entre ambos conjuntos de prueba y entre los regresores, justificando las diferencias en base a las características de los datos y modelos utilizados, considerando también la existencia de *overfitting* y *underfitting*.
Finalmente, visualice y transforme la variable `Sales` usando una transformación logarítmica, y repita el proceso anterior. Comente sobre las diferencias en el rendimiento. **(3 ptos.)**

Política de Integridad Académica

"Como miembro de la comunidad de la Pontificia Universidad Católica de Chile me comprometo a respetar los principios y normativas que la rigen. Asimismo, prometo actuar con rectitud y honestidad en las relaciones con los demás integrantes de la comunidad y en la realización de todo trabajo, particularmente en aquellas actividades vinculadas a la docencia, el aprendizaje y la creación, difusión y transferencia del conocimiento. Además, velaré por la integridad de las personas y cuidaré los bienes de la Universidad."

En particular, se espera que mantengan altos estándares de honestidad académica. Cualquier acto deshonesto o fraude académico está prohibido; los alumnos que incurran en este tipo de acciones se exponen a un procedimiento sumario. Ejemplos de actos deshonestos son la copia, el uso de material o equipos no permitidos en las evaluaciones, el plagio, o la falsificación de identidad, entre otros. Específicamente, para los cursos del Departamento de Ciencia de la Computación, rige obligatoriamente la siguiente política de integridad académica en relación a copia y plagio: Todo trabajo presentado por un alumno (grupo) para los efectos de la evaluación de un curso debe ser hecho individualmente por el alumno (grupo), sin apoyo en material de terceros. Si un alumno (grupo) copia un trabajo, se le calificará con nota 1.0 en dicha evaluación y dependiendo de la gravedad de sus acciones podrá tener un 1.0 en todo ese ítem de evaluaciones o un 1.1 en el curso. Además, los antecedentes serán enviados a la Dirección de Docencia de la Escuela de Ingeniería para evaluar posteriores sanciones en conjunto con la Universidad, las que pueden incluir un procedimiento sumario. Por "copia" o "plagio" se entiende incluir en el trabajo presentado como propio, partes desarrolladas por otra persona. Está permitido usar material disponible públicamente, por ejemplo, libros o contenidos tomados de Internet, siempre y cuando se incluya la cita correspondiente.