Tecnológico de Monterrey

ALGORITMOS Y ANÁLISIS DE DATOS

Grupo: 201

Evidencia 1 . Reporte de Situación Problema I

- Elisa Mota - A01752677
- Emiliano Quiroz - A01746310
- Francisco Diaz - A01746158
- Emilio Villacis - A01750148

Instituto Tecnológico y de estudios superiores Monterrey

Campus: Estado de México

1. Lectura y limpieza de la base de datos

Utilizando el archivo “house_sales.xlsx”, los alumnos deberán escribir un algoritmo para limpiar y ordenar la base de datos. Esto incluye la eliminación o reemplazo de valores faltantes, la conversión de variables de texto por categorías numéricas y la creación, si es necesario, de variables dicotómicas para expresar variables categóricas.

La limpieza y el código se puede encontrar aquí:

Reto_Houses_EV1_CLEANING .html

2. Evaluación de tecnologías de información

Investiga la metodología de estimación de un modelo de regresión lineal usando Excel y software de programación (R, Python, etc.). Compara los programas y elabora un reporte de una cuartilla donde se señale las ventajas y desventajas de cada programa, de igual manera plantea una conclusión cuál es el más adecuado usar en finanzas. Se pueden elegir cualesquiera variables para la regresión

Para la generación de modelos predictivos es común , por no mencionar en su totalidad, el uso de un lenguaje de programación. Programas como Excel, mismos que son destinados a la modificación y lectura de datos estructurados, si bien poseen las cualidades necesarias para un análisis predictivo, no poseen ciertas funcionalidades en cuanto a manejo y limpieza de datos. En el uso de grandes cantidades de datos es de importante valor el uso de programas que eficienten la lectura y la modificación de los mismos en donde R y Python son algunos de ellos.

Las ventajas que estos lenguajes poseen involucran una mayor efectividad en cualquier tipo de limpieza y análisis estadístico de mayor nivel, que a un nivel empresarial, son absolutamente necesarias para la toma de decisiones. La creación de logaritmos de índole predictivo van más allá de selección de variables puesto que también existen conceptos teóricos que deben manejarse con anterioridad para su correcto análisis y que en lenguajes como estos son más “fáciles” su detección.

Como bien sabemos, Excel es el programa por excelencia para el uso de Finanzas, nos permite ver estados financieros con claridad, situaciones de empresas , etc. Sin embargo, para la parte de Machine Learning un lenguaje de programación siempre será el más idóneo

para su funcionamiento ya que permitirá generar acciones basadas en estadística y efectividad.

La elaboración del reporte puede elaborarse a partir de las siguientes preguntas:

1. Considerando la base de datos en su estado original, ¿cómo se puede utilizar para hacer un análisis de los precios de las casas en función de sus características?

Tomando la base de datos en su estado original, si es posible realizar un análisis de los precios. Este análisis será mayormente descriptivo, es decir, se podrá entablar queries/solicitudes para estudiar la base de datos y conocer más a detalle el precio de las casas. El análisis descriptivo involucra también un análisis exploratorio por variable en donde podemos realizar conexiones entre variables y estudiar las relaciones entre ellas con la variable objetivo. En otras palabras, generar gráficos , correlaciones, histogramas, líneas de tendencia, etc. Por mencionar un ejemplo, podríamos buscar un tipo de relación entre el precio con la cantidad de metros cuadrados que la casa posee; este análisis nos servirá para observar el comportamiento del precio ante un aumento o decremento de m2.

Es importante mencionar que: un análisis predictivo , dada las condiciones actuales, no es posible, puesto que la base presenta condiciones no favorables para un correcto análisis estadístico. Una base de datos sucia puede generar sesgo en los resultados, por no mencionar que algunas variables no podrán entrar en el modelo por su presencia de valores nulos. Entonces , dadas las características, una predicción no servirá correctamente y se perdería significativamente el valor del objetivo del proyecto.

2. ¿Qué decisiones se pueden tomar acerca de los datos faltantes?

Como se mencionó anteriormente, la base de datos se encontraba sucia por lo que era necesario un proceso de limpieza (Data Cleansing). Entre los procedimientos se encuentran los datos faltantes o nulos entre varias variables. Un dato Nulo involucra que los datos , en algunos casos como el actual, si poseen un valor, sin embargo, tienen un valor carente o de carencia de algo.

Normalmente estos valores son corregidos dependiendo del tipo de dato de la variable. Las decisiones que normalmente se toman se centran en el cambio de dato por uno nuevo , así como en la introducción de un valor totalmente nuevo en las celdas; esto para lograr tener una base llena y en concordancia con su objetivo. En este caso, para las variables categóricas se decidió realizar dos procedimientos: rellenar los valores Na por ‘’None’ y

rellenar Na por su moda. Para las variables numéricas se rellenó por su media. Estas acciones fueron ejecutadas dependiendo de la variable y fueron detalladas en el código de limpieza.

3. ¿Cómo se pueden transformar las variables de texto para usarse en un análisis estadístico sin perder sus características?

Este proceso se lo conoce como creación de variables Dummie's, pues al tener una variable categórica se les asigna un valor de 1 o 0 dependiendo de su característica. El procedimiento se ajusta para cada valor de la variable por lo que su resultado involucra la creación de nuevas variables que a su vez estarán relacionadas con la variable categórica original. De esta forma, es factible el uso de estadísticas en un análisis predictivo.

El proceso fue entablado en la limpieza de la base de datos en donde se crearon 257 variables Dummy provenientes de 43 variables categóricas originales.

4. ¿De qué manera se puede manipular la base de datos, de manera repetitiva, sin tener que hacerlo manualmente? ¿Qué tipo de funciones de programación se pueden utilizar?

Al hablar de manera repetitiva se puede entender en el mundo de la programación como el uso de funciones ya sea creadas por el usuario, así como en el uso mediante paquetes. la manera repetitiva también se puede efectuar con apoyo de Ciclos (while, for, etc) que van a ayudar a eficientizar el código para no efectuar nuevamente una instrucción. En la base de datos se entablaron innumerables funciones y ciclos "for" para el proceso de limpieza. Estos procedimientos fueron de gran ayuda para solventar inquietudes e instrucciones repetitivas.

5. ¿Qué tipo de plataforma tecnológica (hoja de cálculo o lenguaje de programación) presenta más ventajas para manipular grandes bases de datos?

Como *Licenciados en Administración Financiera* estamos acostumbrados a querer resolver toda clase de problemas mediante las hojas de cálculo de *Excel*, las cuales no hay que menospreciar, ya que son una gran herramienta con un sin fin de utilidades. Sin embargo, hay casos como este, en el que manipular tal cantidad de datos y variables tan solo desde *Excel* podría llegar a hacerse bastante complicado y tedioso; debido a esto la mejor opciòn es optar por un lenguaje de programación el cual nos haga màs sencillo, eficaz y óptimo

que queremos realizar. Ejemplos de esto son *Python* y *RStudio*, los cuales son utilizados para desarrollar aplicaciones, códigos y programas de todo tipo. Y aunque *RStudio* suele ser uno de los lenguajes más utilizados a la hora de resolver problemas y situaciones económicas; hemos optado por llevar a cabo nuestra soluciòn problema a través de *Python*, ya que este se caracteriza por ser uno de los lenguajes más utilizados para la realización de distintas aplicaciones y códigos. Además de que es reconocido por lo fácil que puede llegar a ser la lectura de su código, logrando así realizar las mismas funciones que *RStudio*, pero de una manera más sencilla y eficaz. Por lo que podemos concluir que para manejar grandes bases de datos, como con la que estamos trabajando, nuestra mejor opción será *Python*.

3. **Uso de información**

Investiga en la Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros (CONDUSEF), cuales leyes aplica a las instituciones financieras en cuanto al manejo de datos personales y elabora un reporte de máximo una cuartilla sobre la protección de datos personales que le aplican en el desarrollo de la situación problema.

Como ya tenemos conocimiento la *CONDUSEF* se encarga de defender y proteger a los usuarios de las distintas instituciones financieras mexicanas, y parte de este trabajo incluye el garantizar la protección de los distintos datos personales que son empleados al realizar un trabajo como este. Debido a esto la *CONDUSEF* ha creado la *Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados*; la cual cuenta con 167 artículo, mediante los cuales se nos expresa las distintas obligaciones principios y deberes que tienen las distintas instituciones financieras a la hora de hacer uso de cierta información proporcionada por el cliente.

Debido a lo antes mencionado es que se crearon los *Derechos ARCO*; denominados de esta forma debido a que se refieren a los derechos que tenemos para el *Acceso*, *Rectificación*, *Cancelación* y *Oposición* sobre el tratamiento y uso de datos personales. Es decir que toda persona, la cual sea titular de sus datos personales, siempre tendrá el derecho a acceder a estos, al igual que solicitar su rectificación en todo momento; además de que siempre se cuenta con el derecho de cancelarlos o en todo caso oponerse a su utilización.

Cabe mencionar que todo lo anteriormente mencionado se encuentra respaldado por la *Constitución Política de los Estados Unidos Mexicano*, siendo más precisos en el artículo 16 de esta misma. Aunado a esto, es necesario mencionar que la mayoría de compañías al usar datos personales están obligadas a brindar una acuerdo de confidencialidad, por lo que

además de todo lo antes mencionado, tampoco podrán compartir los datos personales personales del cliente mediante ninguna razón.

Por último, es imperativo que todo lo mencionado con anterioridad es aplicable para nuestro caso de estudio, ya que Sales Real State (SRL) está haciendo uso de una gran base de precio de venta de casas incluyendo sus características principales; por lo que estos mismos tienen ciertas responsabilidades y obligaciones sobre estos datos, siempre respetando las leyes antes mencionadas

Bibliografía

Soluciones de aprendizaje automatizado para determinar precios – Centro Internacional de Casos. (2022). Cic.tec.mx. https://cic.tec.mx/casos/index.php/soluciones-de-aprendizaje-automatizado-para-determinar-precios/

.