**Prueba de coordinador de datos Vanti**

William Ricardo Fernández Garnica

Para esta prueba vamos a abordar el reto de Optimización de la demanda de gas: Predecir la demanda de gas para mejorar la eficiencia de la producción, el almacenamiento y la distribución, evitando pérdidas y optimizando recursos.

1. **Contexto y beneficios del Negocio:**

Vanti facilita la conexión entre la producción de gas natural y los consumidores finales, asegurando una distribución eficaz que respalda tanto las necesidades residenciales como industriales. En un contexto más amplio, la compañía contribuye activamente a los esfuerzos de descarbonización del país al promover el uso de gas natural, reconocido por su menor impacto ambiental y su capacidad para complementar energías renovables intermitentes como la solar y la eólica.

El proyecto de predecir la demanda de gas se alinea perfectamente con los objetivos estratégicos de Vanti, que incluyen mejorar la eficiencia operativa y reducir las pérdidas derivadas de consumos irregulares como el fraude. Esto no solo optimiza los recursos y mejora la seguridad del suministro, sino que también fortalece la posición de Vanti como un actor clave en la transición energética del país.

Al implementar modelos predictivos avanzados para identificar consumos anormales o fraudulentos, Vanti podrá reducir significativamente las pérdidas operativas y mejorar la precisión en la planificación del suministro. Esto se traduce en una mejor gestión de la capacidad de almacenamiento y distribución, así como en una mayor confianza de los clientes en la estabilidad y fiabilidad del servicio proporcionado.

En un contexto más amplio, la iniciativa de Vanti no solo fortalece su posición competitiva en el mercado energético colombiano, sino que también apoya las políticas gubernamentales de expansión del uso de gas natural como una alternativa más limpia y sostenible. Además, contribuye al desarrollo económico regional y fortalece la seguridad energética nacional, alineándose con los compromisos climáticos internacionales y las metas de sostenibilidad.

El enfoque en la predicción de la demanda de gas no solo mejora la eficiencia operativa de Vanti Gas Natural, sino que también subraya su compromiso con la innovación tecnológica y la mejora continua en la gestión de recursos energéticos. Esta estrategia no solo beneficia a la empresa en términos de reducción de costos y mejora de servicios, sino que también contribuye positivamente al panorama energético y ambiental de Colombia a largo plazo.

1. **Objetivos de negocio:**
   1. **Predicción precisa del consumo de gas:**
      1. Desarrollar modelos predictivos que permitan prever con precisión el consumo de gas en un periodo mensual.
      2. Identificar consumos que se desvíen significativamente de las proyecciones esperadas, lo que puede indicar situaciones anómalas como fraude o problemas operativos.
   2. **Mejora en la planificación y eficiencia operativa:**
      1. Utilizar las predicciones de consumo para optimizar la gestión de recursos, como la capacidad de almacenamiento y la distribución de gas.
      2. Aumentar la eficiencia operativa al alinear la producción y distribución de gas con la demanda proyectada de manera más precisa.
2. **Explorando y ajustando la data**

Podemos observar nuestros datos que en este caso contamos con 108895 registros pertenecientes a 8000 cuentas únicas.

Imagen que contiene Tabla

Descripción generada automáticamente

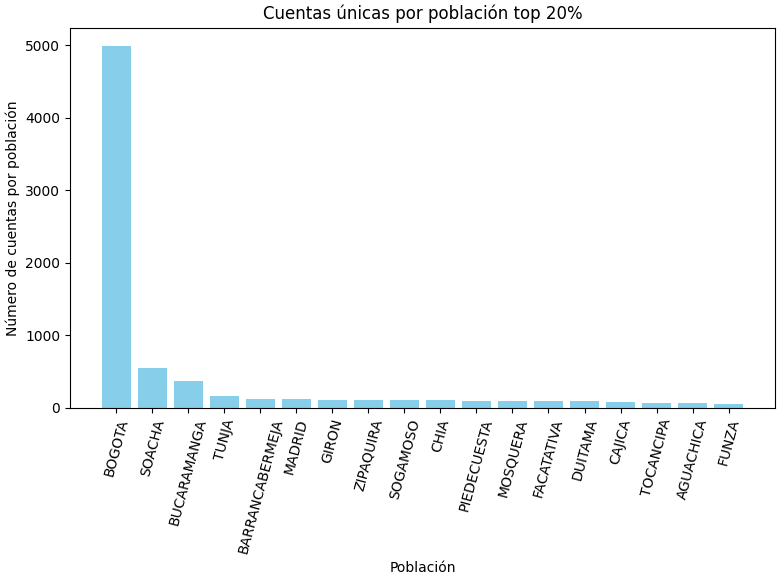
Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamenteTabla

Descripción generada automáticamente Imagen que contiene Interfaz de usuario gráfica

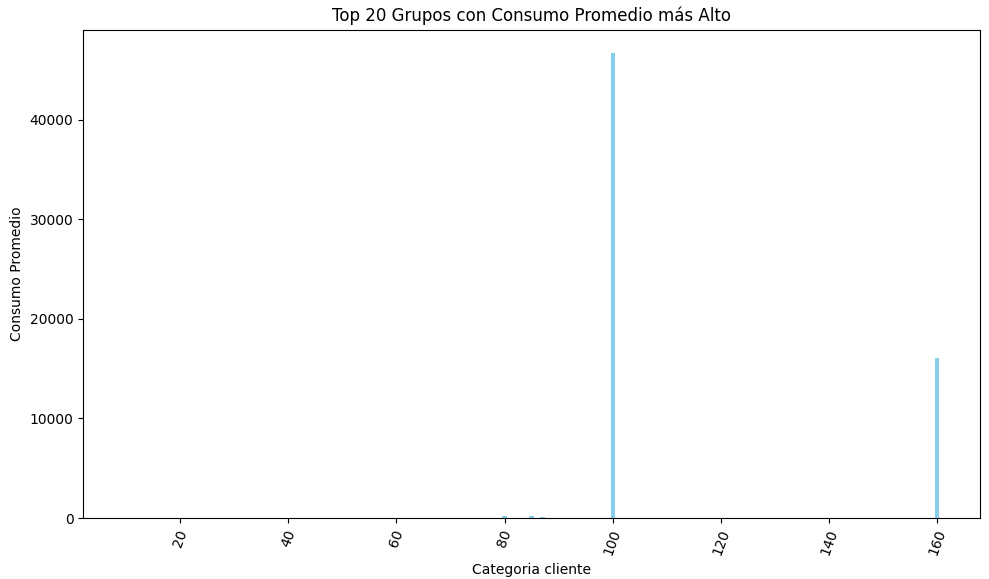
Descripción generada automáticamente

Podemos observar en las siguientes gráficas que tenemos algunos Outlayers a nivel de agrupación, el consumo promedio de las cuentas es muy alto en Zipacon, las porciones P04I y P01I, y para la categoría de cliente 100 y 160.

Gráfico, Gráfico en cascada

Descripción generada automáticamente

Imagen que contiene Gráfico

Descripción generada automáticamente 

Al revisar estos casos nos damos cuenta que en todos los casos corresponden cada uno a una sola cuenta contrato, por lo que pueden ser Outlayers.

Una vez procesados estos casos procedemos a hacer una prueba de chi cuadrado y para las variables categorícas que identificamos, en esta podemos identificar que existe una relación muy significativa entre estas variables, y pueden no ser suficientemente independientes para que sirvan para un modelo de proyección.

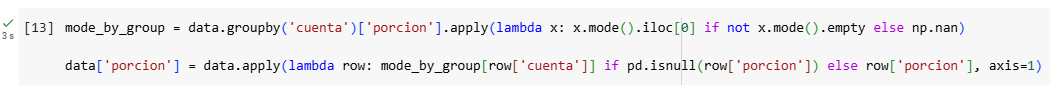


Manejo de valores NULL, podemos identificar 7922 valores NULL en porción, pero al realizar un análisis del problema, pudimos identificar que casi todos los valores NULL se encontraban dentro de registros que pertenecían a la misma cuenta por lo que realizamos un reemplazo de los valores NULL que estaban dentro de la misma cuenta por la moda de la cuenta, lo cuál nos redujo bastante el número de valores NULL hasta llegar a un valor de solo 58 valores NULL, que vamos a aceptar para este ejercicio.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamenteInterfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente



1. **Exploración y manejo de Outliers**

Al realizar una revisión de las cuentas sobre cuantos registros tienen para poder proyectar, encontramos que de las 8000 cuentas que existen, hay 426 cuentas que tienen menos de 14 registros.

Texto

Descripción generada automáticamente

Por lo que vamos a tomar la decisión de no realizar proyecciones sobre estas 426 cuentas, ya que la cantidad de registros no nos lo permite, por lo que las vamos a eliminar. Una vez eliminadas estas cuentas nos quedamos con 7574 cuentas únicas que tienen 14 consumos, que equivalen en nuestra base a 106036 registros. Por lo que solo eliminamos un poco menos del 3% del total de la base.

* 1. **Días de facturación**

Vemos que el promedio de días de facturación es de 29.6 días, el mínimo es de 3 días y el máximo es de 67 días, vamos a explorar si esto es normal dentro de las mismas cuentas o por el contrario estas mediciones son anormales.

Tabla

Descripción generada automáticamente

Para poder identificar si dentro de cada cuenta hay días de facturación anormales vamos a utilizar un rango interpercentil, Escogimos un rango interpercentil y no intercuartilico debido a que como lo observamos en el descriptivo anterior el percentil 25 sacaba días de facturación menores a 29 que no va acorde a lo que en el negocio puede presentarse, que pueden ser facturaciones a 25 días o a 32 días.

Texto

Descripción generada automáticamente con confianza media

Tenemos muy pocas cuentas con más de un valor considerado outlier, pero al analizar más estas cuentas, no parecen ser valores que para este ejercicio nos afecte demasiado, revisaremos con el consumo para comprobar si esto nos afecta o no.

Texto

Descripción generada automáticamente con confianza media

Podemos identificar que, aunque tenemos outliers, estos se presentan muy poco por cada cuenta y los valores se encuentran dentro de los días de facturación normales, el gráfico de caja y bigotes y el QQ nos muestra días bastante extraños, podríamos separar estos casos para entrar a revisar con negocio que sucede, se puede suponer un problema para acceder a la zona. Por el momento para abordar esta situación vamos a tratar estos outliers de días de facturación normalizando el consumo a 30 días.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente Gráfico, Gráfico de dispersión

Descripción generada automáticamente

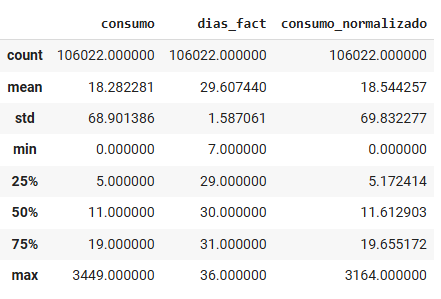
**Estandarización de los consumos**

Normalizamos el consumo acorde a los días facturados del cliente, con esto ya estandarizado y descartando los problemas que observamos con los días facturados procedemos ahora a revisar que outliers tenemos en el consumo y si están relacionados con los días facturados.

Debido a la situación anterior en la que tenemos días de facturación diferentes vamos a realizar una estandarización a 30 días de los consumos, por lo que utilizaremos la siguiente fórmula:



Aun normalizando el consumo vemos consumos bastante altos, por lo que vamos a explorar estos consumos dentro de la misma cuenta contrato para entender si estos consumos son anormales o no.



Exploramos con esta formula en la que nos alineamos con lo exigido por la CREG, utilizando 3 desviaciones estándar por encima de la media para identificar Outlayers, con esto pudimos encontrar 254 outlayers que tienen un comportamiento anormal, por lo que el tratamiento de estos outlayers va a ser eliminarlos de nuestro dataframe.

Imagen que contiene Gráfico de dispersión

Descripción generada automáticamente

Ahora debido a que no vamos a predecir una sola serie de tiempo sino más de 7500 cuentas tenemos que explorar si nuestros datos son aptos para ciertos modelos que nos exigen los principios de normalidad (media 0 y varianza 1), utilizando una prueba de Shapiro Wilks, usando un 5% como Alpha, nos da que del total de nuestra base de datos un 73.25% tiene un comportamiento normal. Por lo que es necesario realizar una normalización. Por lo que utilizaremos un MinMaxScaler() para realizar esta tarea.

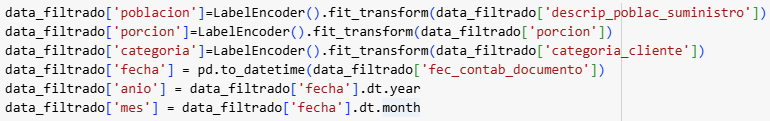


**Preparación de los datos y modelado**

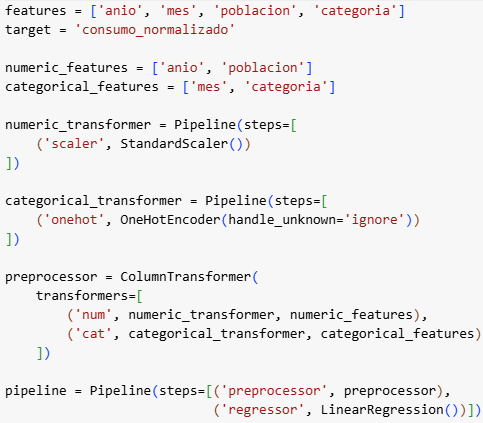
Primero vamos a empezar eliminando todas las cuentas contrato que solo tengan consumos 0.



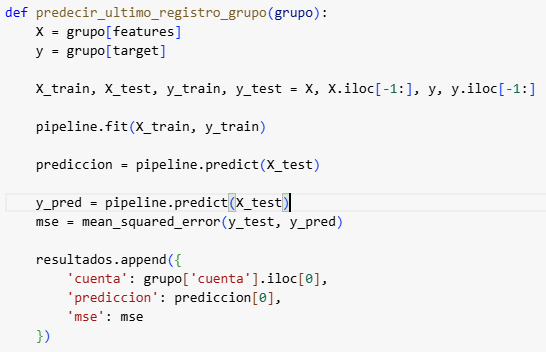
Luego vamos a terminar de procesar nuestras variables de interés realizando un label encoding



Aunque no es lo más optimo decidimos realizar un proceso adicional, para asegurar que nuestras variables quedan en el formato adecuado para nuestro modelo de regresión múltiple.



Realizamos nuestro entrenamiento y testeo de nuestro modelo de regresión múltiple, se utilizará el último registro como testeo y los 13 registros anteriores como entrenamiento.



El rendimiento del modelo terminó siendo muy bueno, nuestro r^2 se ajusta demasiado bien a los datos, y un MSE de 9.76 que es bastante bajo.



También manejamos un modelo ARIMA, aunque no recomendable para este tipo de casos debido a los pocos registros que tiene cada cuenta, nos encontramos con un escenario similar, en el que nos encontramos con un R^2 de 0.97 y un MSE ya más alto de 150. Para este modelo para el manejo de los hiper parámetros manejamos un auto arima, decidí no tunearlos manualmente ya que cada cuenta tenía una tendencia diferente, ninguno de los hiper parámetros p, d, q sería adecuado dejarlo fijo.

Para los hiper parámetros del auto arima:

* Seasonal = False – Esto debido a que no tenemos más de un año de proyecciones y puede tomar cualquier aumento en el consumo como estacional.
* Stepwise = True – Enfoque de búsqueda de los hiper parámetros p, d, q.



Preguntas del problema:

**- ¿Qué problema se quiso resolver con el desarrollo de la solución?**

* Desarrollar un modelo predictivo que permita prever con precisión el consumo de gas en un periodo mensual.
* Identificar consumos que se desvíen significativamente de las proyecciones esperadas, lo que puede indicar situaciones anómalas como fraude o problemas operativos.

**- ¿Qué información buscaste y utilizaste para el desarrollo de la solución? Incluye las hipótesis que te llevaron a descargar dicha información.**

Se utilizó la información entregada por el equipo.

**- Muestra los insights más relevantes que hayas obtenido de tu análisis exploratorio.**

Se presentan con mayor detalle en el cuerpo del documento, pero son:

* Tenemos cuentas con muy pocos registros para poder realizar proyecciones, estas cuentas se pueden manejar con un promedio, que nos permitiría reducir la intensidad computacional para el cálculo de estos modelos.
* Tenemos cuentas con consumos muy elevados, el hecho de que sean 6 cuentas de 8000, con un comportamiento normal, nos lleva a pensar que pueden ser cuentas que están mal clasificadas o pueden ser un caso de fraude, por lo que tenemos que investigarlas.
* Hay una variación importante de los días de facturación, se debe entrar a revisar con el equipo técnico que está sucediendo ya que un periodo de facturación de 3 días o de 68 días es anormal y no debería suceder considerando que hay obligaciones legales de por medio.
* Tenemos muy pocos registros por cuenta para los objetivos ambiciosos que tenemos.

**¿Cómo estos te guiaron a la hora de plantear la solución?**

* Son Insights que me permitieron realizar una priorización sobre los casos que realmente son importantes predecir, y también como poder preprocesar la información para el caso en cuestión.

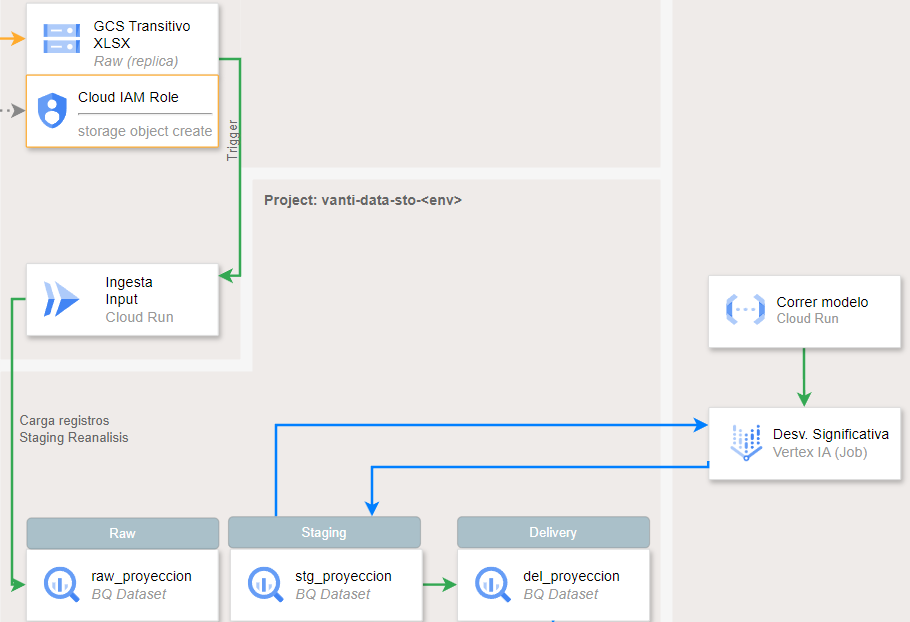
**Danos un contexto de la solución que escogiste, y cómo la abordaste. ¿Qué metodologías aplicaste? ¿Qué métricas de evaluación utilizaste?**

* Busqué utilizar inicialmente una metodología básica que no fuera muy compleja ya que son muchos casos que deben ser procesados, por lo que fui directamente por una regresión múltiple que resultó ser muy buena para predecir.
* Utilicé el MSE y el r^2 para este problema.

**¿Cuál es tu solución final? ¿Cómo se comporta? ¿Cómo soluciona o ayuda a solucionar el problema de negocio?**

* Nos encontramos con que 14 predicciones no son suficientes para poder realizar una predicción adecuada, aunque nuestras métricas nos dan valores muy altos, tanto hasta para pensar que estamos realizando un overfitting, no son confiables ya que no capturan estacionalidad, solo están capturando tendencia. Pero tenemos que evaluar que tanto ahorro puede generarnos una predicción más precisa a nivel de negocio.

**Has un bosquejo del pipeline que se utilizaría para implementar tu solución.**



**Concluye y plantea próximos pasos que harías después de entregar esta solución.**

Necesidad de Datos y Modelos Más Complejos: La disponibilidad limitada de solo 14 predicciones por cada cuenta resalta la importancia de obtener datos históricos más extensos para poder emplear modelos predictivos más avanzados. Estos deben capturar tanto la tendencia como la estacionalidad del consumo de gas para ofrecer predicciones más precisas y confiables.

A pesar de los resultados prometedores según nuestras métricas actuales, es esencial abordar el riesgo potencial de sobreajuste del modelo. Implementar técnicas para mitigar el overfitting asegurará que nuestras predicciones sean robustas y aplicables en diversos contextos operativos.

Es crucial que nuestros modelos puedan capturar de manera efectiva la variabilidad estacional del consumo de gas a diferentes escalas (diaria, semanal, mensual). Esto no solo mejorará la precisión de las predicciones, sino que también nos permitirá identificar desviaciones significativas que podrían indicar problemas operativos o anomalías como fraudes.

La implementación de predicciones más precisas del consumo de gas tiene el potencial de transformar nuestra eficiencia operativa y estratégica de manera significativa. Mejorar la gestión de recursos como la capacidad de almacenamiento y la distribución de gas, alineándolos mejor con la demanda proyectada, no solo optimizará nuestros costos operativos, sino que también fortalecerá nuestra capacidad para responder de manera efectiva a las demandas del mercado y de los clientes.