Aplicación de Machine Learning al análisis de producción y venta de hidrocarburos en la Argentina

Universidad Tecnológica Nacional

Cátedra Ciencia de Datos

Noviembre 2019

Agustina Rosario Elias, Lautaro Rshaid y Nicolás Martínez de Ibarreta

Abstract

Hoy día a nivel país y mundial, nos encontramos en un momento de crisis energética. Dicha crisis tiene causas diversas, cada una de ellas con un peso distinto.

En este proyecto decidimos indagar en la situación de algunos factores de esta crisis: la producción y venta de gas natural y petróleo crudo y sus productos derivados, en un intento de, no solo entender la historia de los mismos, sino también hallar respuestas y/o nuevas incógnitas en lo que concierne a su evolución y desarrollo en los próximos años, cuando la industria deje de ser como hoy por hoy la conocemos.

1 Introducción

En el siguiente informe buscaremos entender el comportamiento tanto de la producción de petróleo crudo y gas natural como de la venta de naftas y gasoil, los principales combustibles líquidos derivados del petróleo, durante los últimos 20 años (1999 – 2019) en la República Argentina.

Esperemos con este análisis obtener información que nos ayude a comprender la dinámica de este insumo vital, como lo es la energía, para el futuro desarrollo de la economía nacional.

Del petróleo crudo derivan muchos subproductos: algunos vitales en el suministro de energía en Argentina y otros inmersos en nuestro día a día de un modo más "simple", como lo son los plásticos, telas, detergentes, jabones, entre otros.

En el presente análisis tendremos en cuenta dos de ellos: la nafta - mezcla de hidrocarburo líquido inflamable - y el gasoil o diésel - hidrocarburo líquido de densidad sobre 850 kg/m³.

2 Dataset

Nuestro dataset fue "Indicadores sectoriales de hidrocarburos" que contiene los registros de producción de hidrocarburos y comercialización de sus principales derivados desde 1999 al 2019.

Está compuesto por 22 features y 11.950 registros tomados de forma mensual, trimestral y anual para todas las provincias de la República Argentina y un consolidado nacional. Todas las features, excepto alcance (distrito al cual pertenece la información), frecuencia temporal (intervalo de tiempo de la muestra) y el índice de tiempo (fecha de la información), son del tipo float.

El primer y principal problema de este dataset fue la cantidad de registros nulos. Existían alrededor de 11.800 nulls en las features relacionados con las reservas de petróleo crudo y gas. Luego de analizar las causas y las posibles implicancias de los nulls antes mencionados optamos por eliminar esas features para enfocar nuestro análisis en la producción de hidrocarburos.

distribution_identifier	indice_tiempo	alcance_id	alcance_nombre	prod_gasnat	prod_petcru	vta_gasoil	vta_nafta	precint_gasnat	precint_petcru	mes	anio
6.1	1999-01-01	26	CHUBUT	75662.974	478924.271	17063.0	15960.0	32.124328	52.053055	1	1999
6.1	1999-02-01	26	CHUBUT	78533.919	464650.702	16852.0	13381.0	31.937054	50.785146	2	1999
6.1	1999-03-01	26	CHUBUT	58142.882	531098.513	15574.0	13479.0	31.358831	65.456304	3	1999
6.1	1999-04-01	26	CHUBUT	56770.974	510881.134	17374.0	13317.0	30.949112	82.804204	4	1999
6.1	1999-05-01	26	CHUBUT	63289.978	527635.096	16487.0	12567.0	34.001200	91.222343	5	1999

Imagen 1

No obstante, seguían existiendo una gran cantidad de registros nulos concentrados en los precios internos de los hidrocarburos. Debido a nuestra decisión de trabajar con provincias productoras (las cuales siempre contaban con precio interno), eliminamos los registros que contenían un null en dicho campo.

Consideramos que la mejor manera de analizar los datos era mediante una toma de registros mensual y por provincia, por ende, también eliminamos los registros anuales y trimestrales correspondiente a las provincias y los nacionales de todo tipo. Además, descartamos el año 2019 ya que se encontraba incompleto.

Finalmente, el dataset queda formado por 2160 samples y 12 features que explican la producción de hidrocarburos y las ventas de sus derivados en las provincias productoras de petróleo con su correspondiente índice de tiempo (ver imágen 1).

3 Background teórico

Para el presente trabajo decidimos trabajar con modelos de regresión.

La regresión es un método para modelar un valor objetivo basado en predictores independientes. Estas técnicas en su mayoría difieren en función del número de variables independientes y el tipo de relación entre las variables independientes y dependientes.

3.1 Regresión lineal simple

Una regresión lineal es un acercamiento para estimar la relación lineal entre dos tipos de variables, una variable de respuesta Y y otra u otras variables explicativas Xi. Es el modelo más usado y es una técnica fundamental a la hora de analizar datos. El modelo es de la forma:

$$y = X\beta + \epsilon$$

Donde Y es un vector n x 1, X es una matriz n x p, β es un vector de coeficientes p x 1 y ϵ es el término de error normal estándar. Normalmente a un modelo donde p = 1 lo llamamos modelo de regresión lineal simple y un modelo donde p > 1 lo llamamos modelo de regresión lineal múltiple o multivariado.

Siempre que construímos un modelo habrá desviaciones entre lo que el modelo predice y lo que se observa en la muestra. La diferencia entre esos valores es conocido como los residuales del modelo

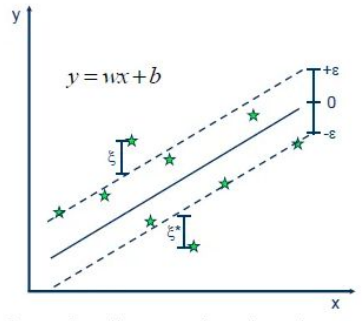
3.2 K-Nearest Neighbors – Regression

K Nearest Neighbors (KNN) es un algoritmo que almacena todos los casos disponibles y predice el objetivo numérico en función de una medida de similitud (por ejemplo, funciones de distancia). El peso (weights) del KNN indica cómo se interpolara cada K vecino: uniforme (todos por igual) o por distancia.

En el entrenamiento se determinan los K vecinos más cercanos por distancia euclídea (distancia par-a-par). En general, un valor de K grande es más preciso ya que reduce el ruido general; pero hay que tener presente que no siempre esto es así. La técnica de cross-validation es otra forma de determinar retrospectivamente un buen valor K mediante el uso de un conjunto de datos independiente para validar su valor K.

3.3 Máquina de vectores de soporte - Regresión SVR

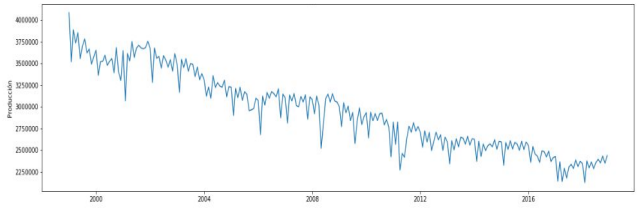
Support Vector Regression (SVR) es un clasificador lineal que busca el hiperplano separador que maximiza el margen entre clases y de esta forma se logra minimizar el error, teniendo en cuenta que se tolera parte del error con un margen épsilon.



En este modelo se utilizan “kernels”. Éstos permiten mapear no linealmente mis muestras a otro espacio donde el hiperplano lineal funcione.

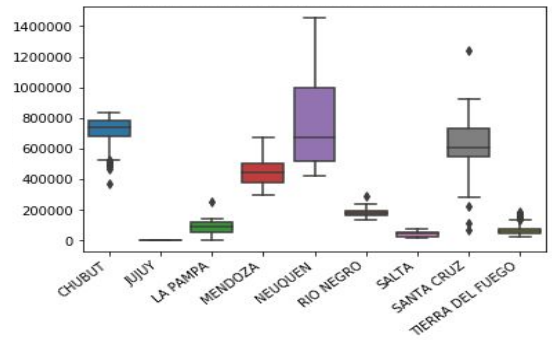
4 Análisis exploratorio de datos

Para comenzar, decidimos observar la producción del petróleo crudo a lo largo de los últimos 20 años. Esto nos dió un panorama amplio de la situación a nivel país y, automáticamente, disparó la pregunta ¿Por qué cayó un 50% la producción desde 1999?



Al ser una industria afectada por tantas variables, no es fácil responder a dicha pregunta, pero observando el comportamiento de las provincias productoras, es posible obtener ciertos insights.

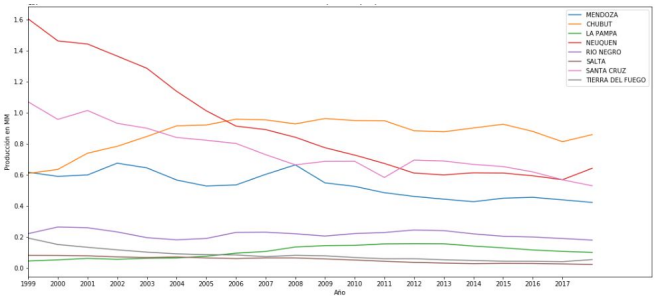
Al hacer un gráfico de boxplots con la producción por provincia, vimos que, mientras que el resto de las provincias se mantenía dentro de cierto rango normal, Neuquén tenía un amplio rango en sus valores.



Cuando graficamos dichos valores en función del tiempo, observamos que el comportamiento de las provincias había sido, en su mayoría, constante.

Chubut (línea color naranja) creció a comienzos del milenio y, manteniendo su producción, hoy aporta el 30,2% de la producción de petróleo del país.

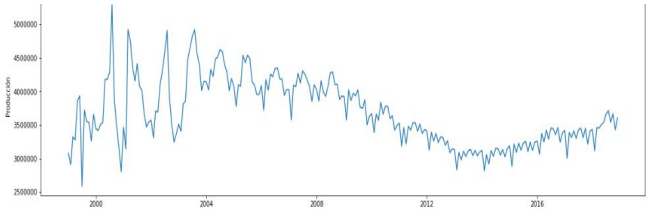
Este % es posible debido a la gran caída de producción en Neuquén (color rojo), cuyas causas no logramos identificar. A pesar del historial desfavorable en la provincia, se estima que la activación de Vaca Muerta llevará la producción a altos niveles.



Luego de entender el contexto del petróleo crudo, continuamos nuestra investigación con la producción de gas natural.

Dicho hidrocarburo tiene como productor principal a Neuquén, con el 84% de los flujos diarios y el 70% de los volúmenes totales aprobados en 2019.

Dicho ésto, la producción que observamos a nivel país, se corresponde con la de Neuquén.



En 1973 se descubre Loma de la Lata, yacimiento que convierte a la provincia en líder productor. Se comienza a explotar hasta 2004 - año en que Repsol toma control de Loma de la Lata -, cuando comienza el declive, el cual se atribuye a falta de inversión y sobreexplotación. En 2013 vuelve a manos de YPF y comienza nuevamente el crecimiento.

Finalmente, en búsqueda de más razones que expliquen el fenómeno de la baja de producción de petróleo crudo, investigamos sobre las ventas de nafta y gasoil.

Mientras que las ventas de gasoil se mantienen casi constantes, las ventas de nafta nos sorprendieron: contrario a lo que pensábamos, vienen en aumento sostenido desde 2005, pasando de 750.000 a 1.750.000

La búsqueda de razones para esta situación no nos fue satisfactoria:

- No puede atribuírse a una provincia en especial, ya que todas vienen aumentando sus ventas.
- Argentina no formó parte del grupo de países exportadores de nafta durante los últimos 10 años, por lo que la exportación no es una de las razones de este aumento.

Uno de los factores que pudo haber contribuído fue el aumento de vehículos en el país. Por ejemplo, en Mendoza en 2009 había un auto cada tres personas, mientras que en 2018 ese número paso a ser un auto cada dos personas. Igualmente, no conocemos con exactitud el peso de este factor en el aumento constante observado.

5 Modelos de regresión

5.1 Introducción

La idea al comenzar la aplicación de los modelos es comprobar nuestra hipótesis:

Utilizando los datos de producción de gas natural, venta de nafta, venta de gasoil, mes y año es posible predecir la producción de petróleo crudo.

Para probar dicho statement, decidimos usar los tres modelos de regresión mencionados en el punto 3, calcular sus errores y analizar cuál es el más conveniente para nosotros.

Definimos la variable **y** como la producción de petróleo crudo y **x** como una matriz con las features mencionadas en la hipótesis.

Luego dividimos nuestras variables en train y test sets, siendo 80% train y 20% test.

Finalmente, escalamos nuestros **x** sets para que nuestros datos se encuentren en un rango acotado y, en consecuencia, la regresión sea más precisa.

Lo que hicimos con cada modelo finalmente fue pedirle que prediga los datos del test set, sin mostrarle las labels correspondientes.

5.2 Regresión lineal

Comenzamos con una regresión lineal simple, ajustando el modelo a los sets creados.

Al hacer la predicción, obtuvimos un resultado con los siguientes errores:

- MSE = 47.446.774.196,5
- RMST = 217.822,8
- MAE = 182.118,3

5.3 KNN Regression

En segunda instancia, probamos un modelo KNN Regression, definiendo como weight = distance.

Para establecer el hiper parámetro neighbors k, hicimos un Grid Search y Cross Validation con 6 folds. Probamos tres valores distintos y finalmente el modelo utilizó k = 20.

Luego de entrenar el modelo, hicimos la predicción de la producción y obtuvimos los siguientes errores:

- MSE = 14.778.102.278,2
- RMST = 121.565,2
- MAE = 79.412,8

5.4 Support Vector Regression

Por último probamos el modelo SVR. En este caso también hicimos Grid Search para los hiperparámetros y Cross Validation con 6 folds.

Probamos el Grid Search por kernels separados (linear y radial basis function) y probando C 1 o 100, obteniendo como mejores parámetros kernel lineal y C = 100.

Finalmente, al entrenar el modelo y hacer la predicción, obtuvimos:

- MSE = 75.281.325.758,9
- RMST = 274.374,4
- MAE = 192.292,6

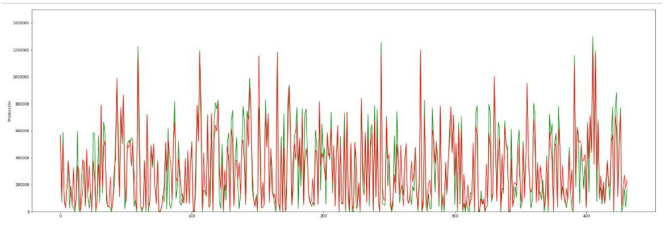
6 CONCLUSIONES

6.1 Resultados de los modelos

Como resumen final podemos comparar los resultados obtenidos:

	MSE	RMSE	MAE
Lineal Regression	47,446,774,196.50	217,822.80	182,118.30
KNN Regression	14,778,102,278.20	121,565.20	79,412.80
SVR	75,281,325,758.90	274,374.40	192,292.60

Concluimos que el modelo que mejor predice la producción de petróleo crudo es KNN Regression, con un MAE de 79,412. Considerando que las producciones llegan hasta 1.2 millones en ciertos casos, estaríamos hablando de un error de aproximadamente el 7%.



Con estos resultados podemos decir que efectivamente nuestra hipótesis inicial es correcta: **es posible predecir la producción de petróleo basándonos en las features producción de gas natural, venta de nafta, venta de gasoil, mes y año.**

6.2 Takeaways

El análisis nos ayudó a entender el efecto que tienen en la producción variables como la inversión, la política de precios que existieron en el país o el declive natural de los yacimientos, además de la relación entre las features, demostrada en la hipótesis planteada y comprobada.

Llegamos a la conclusión que es una industria que necesita una constante inversión buscando generar utilidades en el mediano plazo por lo tanto dar previsibilidad y estabilidad es primordial.

Por otro lado, el consumo de combustibles líquidos continuó en aumento a pesar del declive en la producción. Esto nos lleva a pensar que se debió importar grandes cantidades de crudo para procesar en las refinerías argentinas o el producto final significando una sangría de divisas vía comercio exterior.

El desarrollo de esta industria debería ser parte del plan estratégico nacional, no solo por las ganancias que genera la extracción y venta de hidrocarburos, si no porque produce una sinergia con la industria nacional además de generar las divisas tan necesarias para el desarrollo nacional.

6.3 Future work

La pregunta más grande que nos llevamos es: ¿Es posible predecir qué pasará con la producción en el futuro?

De más está decir que se deberían tener más datos en cuenta para hacer dicho análisis, tales como políticas económicas, nuevos yacimientos, inversiones, aumenta o disminución de demanda, etc. Pero creemos que sería muy interesante y factible llevar este trabajo a ese nivel.

Debido a lo observado, analizado y aprendido en este trabajo, creemos que, mientras continúen las inversiones en la industria, la producción aumentará, al menos al corto plazo.

7 Sources

http://170.210.83.53/htdoc/revele/index.php/cuadernos/article/view/1090/1130

https://www.researchgate.net/profile/Roberto_Kozulj/publication/266074748_La_crisis_energetica_de_la_Argentina_origenes_y_perspectivas/links/562e3bca08ae22b17035d65a/La-crisis-energetica-de-la-Argentina-origenes-y-perspectivas.pdf

https://www.researchgate.net/profile/Adolfo_Giusiano/publication/263236061_Evaluacion_del_Shale_Oil_de_la_Formacion_Vaca_Muerta_Analisis_de_la_declinacion_de_la_produccion/links/00b7d53a34b867d1d2000000.pdf

https://www.saedsayad.com/k_nearest_neighbors_reg.htm

https://www.saedsayad.com/support_vector_machine_reg.htm

https://medium.com/@calaca89/entendiendo-la-regresi%C3%B3n-lineal-con-python-ed254c14c20

https://www.elsol.com.ar/hay-un-auto-cada-dos-personas-en-mendoza

https://www.lanacion.com.ar/economia/despues-decada-argentina-vuelve-exportar-petroleo-alta-nid2239700

https://inneuquen.info/nota-principal/un-oasis-de-gas-natural-neuquen-se-convirtio-en-la-mayor-exportadora-del-recurso-en-el-pais

https://www.petrolnews.net/noticia.php?ID=3a25ea&r=12139