Cosechando el éxito

Un proyecto de ciencia de datos para analizar el comportamiento de precios de granos y commodities con Python

Aranzazu Salcedo Ferraggine

1 - Contexto / Audiencia
2 - Preguntas de interés
3 - Análisis Exploratorio
4 - Modelado y Predicciones
5 - Resultados
6 - Fuentes

Contexto / Audiencia

El objetivo de este análisis es determinar si a través del estudio de ciertas variables estratégicas es posible establecer un precio esperado para los cereales de mayor relevancia para la Argentina. De este modo el pequeño productor argentino podría estar mejor preparado para los posibles movimientos del mercado.

En un país productor como es la Argentina, acercar herramientas innovadoras y con base científica al sector agropecuario de menor escala, podría revolucionar su desarrollo.

En la actualidad es un sector que aún lucha contra la digitalización y aplicación de tecnología.

Preguntas de interés

El objetivo principal del proyecto es lograr predecir los valores futuros del precio del trigo.

A través de las siguientes inquietudes logramos entender el conjunto de datos

- ¿Influye el precio del barril de crudo en el precio del maíz? Ambos precios tienen una relación positiva?
- ¿En qué proporción la variación del precio del crudo influye y se traslada al precio de la gasolina?
- ¿El precio de la soja guarda una relación con el precio del maíz
- ¿El precio del trigo en los últimos cinco años solo ha ido en aumento?
- ¿El precio del oro tiene relación directa con el precio del barril de crudo?
- ¿El precio del oro tiene relación positiva con el precio del maíz?

Análisis exploratorio

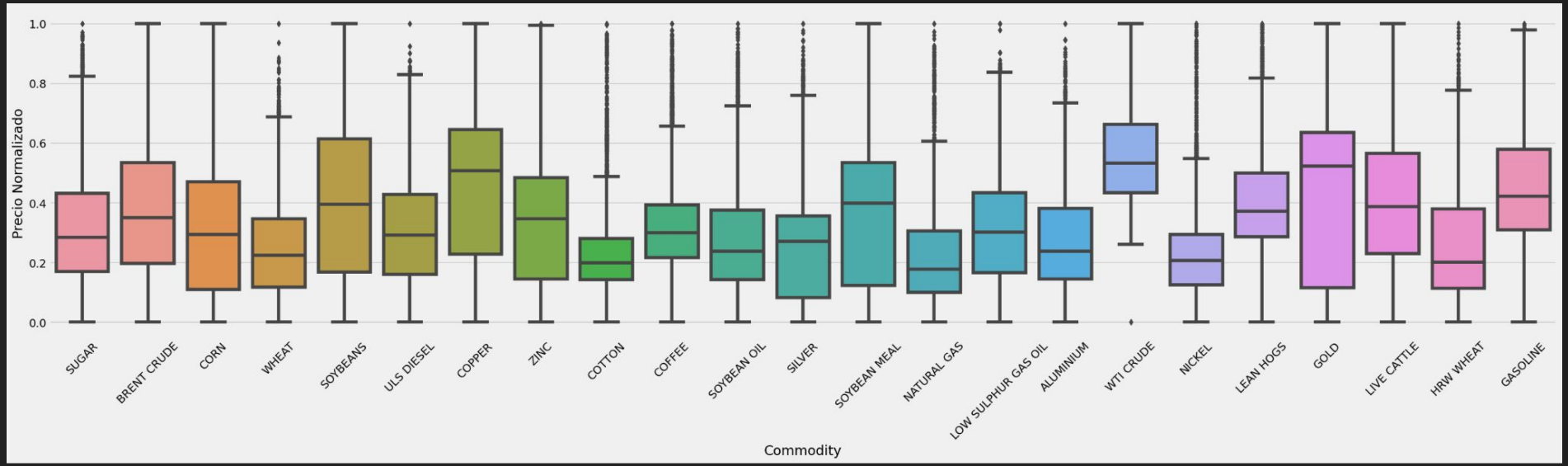
El conjunto de datos principal contiene cotizaciones históricas de granos, metales, combustibles y otros commodities.

Mediante el análisis exploratorio es posible entender mejor la relación entre dichas variables.

Se utilizaron diversas visualizaciones para una mejor comprensión.

¿Cuántos outliers tiene cada variable?

Los outliers son valores que se encuentran fuera de los puntos mínimos y máximos determinados estadísticamente para una distribución normal. Conocer si existen y en donde se ubican nos permitirá crear un mejor modelo.

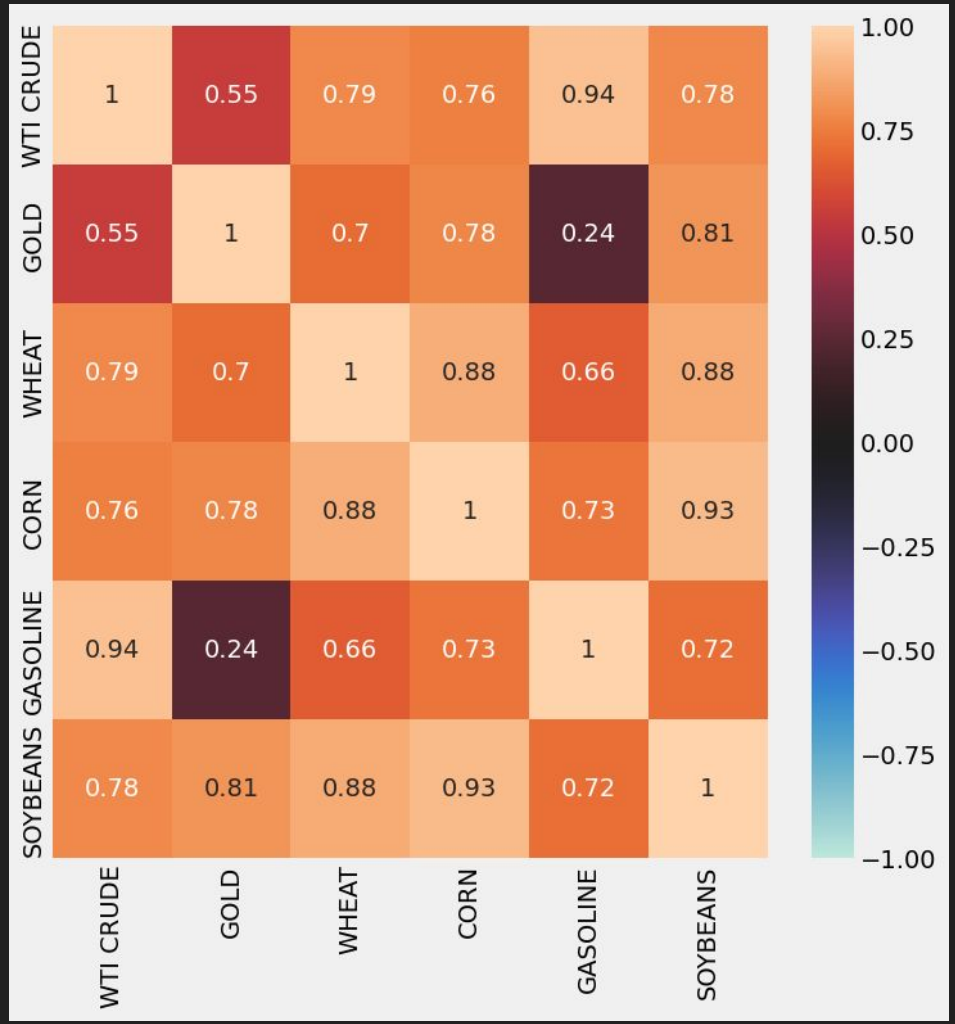


Las variables ‘NATURAL GAS’ y ‘NICKEL’ fueron eliminadas por presentar demasiados valores atípicos.

¿Cómo se relacionan nuestras variables de interés?

Utilizando una matriz de correlación vemos que:

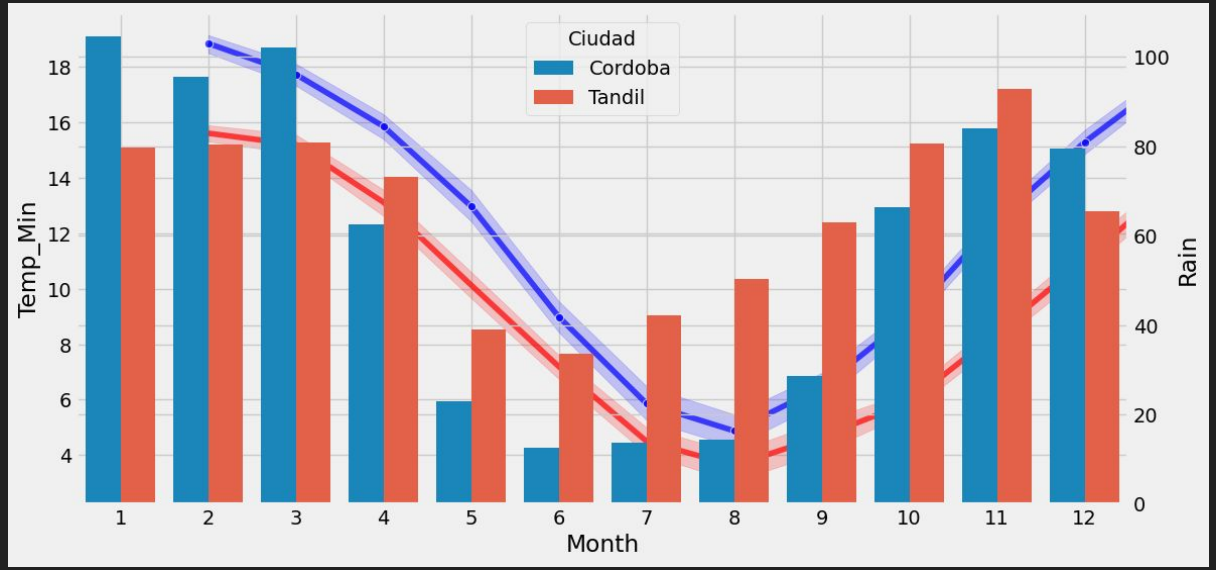
- El precio de la soja guarda relación con el precio del maíz
- El precio del oro tiene relación directa con el precio del barril de crudo.
- El precio del oro tiene relación positiva con el precio del maíz.



Incluimos datos del clima de ciudades con presencia agrícola en Argentina

Para agregar más dimensiones a nuestro conjunto de datos incorporamos nuevas fuentes.

En este caso observamos la estacionalidad de lluvias y temperatura promedio de las ciudades de Córdoba y Tandil.



Se observa un comportamiento similar en ambas ciudades, durante los meses de invierno y menor temperatura se presenta la época de menores lluvias.
La ciudad de Córdoba presenta mayor cantidad de lluvias (mm) de diciembre a marzo, mientras que la temperatura es siempre mayor en relación a Tandil.

Modelado y predicciones

Nuestra incógnita son valores numéricos de una variable continua.

Se seleccionó un algoritmo de regresión lineal que permitiera calcular los valores del precio de trigo en función de los valores que presentaran las otras variables del conjunto.

Luego se realizaron diversas pruebas utilizando herramientas para lograr el mejor modelo posible.

- Creamos nuevas variables a partir de las fechas de las cotizaciones.
- Agregamos variables a partir de nuevos conjuntos de datos: clima, divisas e inflación.
- PCA para simplificar el conjunto de variables.
- Cross validation

Resultados obtenidos

Utilizar el método de cross validation K-fold mejoró el desempeño del modelo. La Iteración Nro 4 es sin duda la que presenta mejores métricas y la simpleza del modelo permite una fácil comprensión. Presenta el mayor R2, mostrando un mejor ajuste a los datos, y los menores MAE y RAE.

	Iteracion_1	Iteracion_2	Iteracion_4	Iteracion_3	Iteracion_5
Mean Squared Error	605.420095	579.125897	535.111905	2719.170470	2957.327553
Root Mean Squared Error	24.605286	24.065035	23.101572	52.145666	54.360194
Mean Absolute Error	18.084065	17.180540	17.221982	41.467664	42.368662
MAPE	3.267969	3.138896	3.092761	7.186163	7.238172
RAE	0.141760	0.134677	0.028941	0.307165	0.071200
Median Absolute Error	13.893842	13.315024	13.487913	33.940120	34.536243
r^2	0.976894	0.977897	0.980203	0.899973	0.891325
r^2 Ajustado	0.976377	0.977245	0.980031	0.898129	0.890930

Fuentes utilizadas

Kaggle Api:

https://www.kaggle.com/datasets/debashish311601/commodity-prices
https://www.kaggle.com/datasets/carlosrearte/kpis-economicos-argentina
https://www.kaggle.com/datasets/carlosrearte/ipc-argentina-19942022

Open Meteo Api:

https://archive-api.open-meteo.com

Python Libraries:

Pandas
Matplotlib
Statsmodels
Scikit-Learn
Seaborn
Numpy
Requests
Json
Locale