ITBA
Instituto Tecnológicc
de Buenos Aires

VIOLETA SAGUIER

CASO DE ESTUDIO

PREDICCION DE VENTAS

82.05 Análisis Predictivo

12/09/2022

AGENDA

01 Introducción

Caso de negocio, objetivo y desafío del trabajo.

02 Tratamiento de la base

Elección, primera inspección y preparación de las variables. Creación de variables nuevas y eliminación de aquellas irrelevantes para el análisis.

03 Análisis exploratorio

Presentación gráfica y analítica de los datos para dar a conocer su relevancia en el modelo.

04 Modelo

Selección del modelo predictivo. Partición de la base de datos en entrenamiento y testeo. Aplicación del modelo.

CASO DE NEGOCIO

Historia de la venta de películas en cines.

OBJETIVO

Necesidad de conocer el comportamiento de los asistentes al cine para poder tomar decisiones dentro de la organización a la hora de disponer películas.

DESAFÍO

Establecer un patrón de comportamiento de la venta de tickets a partir de técnicas de Machine Learning.

PROCESAMIENTO DE LA BASE

KAGGLE - Cinema tickets

La base contiene información de la transmisión de películas en distintos cines. Esta incluye distintos atributos de la misma dispuestos a continuación. La información se despliega entre las fechas 21/02/2018 y 04/11/2018. La base originalmente cuenta con 142,524 registros y 14 variables. Actualización: Quarterly.

Variables originales

```
$ film_code     <dbl>
$ cinema_code   <dbl>
$ total_sales   <dbl>
$ tickets_sold  <dbl>
$ tickets_out   <dbl>
$ show_time     <dbl>
$ occu_perc     <dbl>
$ ticket_price  <dbl>
$ ticket_use    <dbl>
$ capacity      <dbl>
$ date          <date>
$ month         <dbl>
$ quarter       <dbl>
$ day           <dbl>
```

Modificaciones.

- Se crea la variable weekday.
- Se modifica occu_perc a valor sobre 1.
- Se modifican los tipos de dato:

```
$ film_code     <fct>
$ cinema_code   <fct>
$ total_sales   <dbl>
$ tickets_sold  <int>
$ tickets_out   <int>
$ show_time     <dbl>
$ occu_perc     <dbl>
$ ticket_price  <dbl>
$ ticket_use    <int>
$ capacity      <int>
$ date          <date>
$ month         <fct>
$ quarter       <fct>
$ day           <fct>
$ Weekday       <ord>
```

OTRAS MODIFICACIONES.

- Se eliminan las variables *show time* y *quarter*.
- Se modifica el valor ticket_price para un análisis más claro.
- Se crea la variable success que indica “YES” o “NO” de acuerdo al porcentaje de ocupación.

	film_code	cinema_code	total_sales	tickets_sold	tickets_out	occu_perc	ticket_price	ticket_use	capacity	date	month	day	weekday	occu_real	success
	<fct>	<fct>	<dbl>	<int>	<int>	<dbl>	<dbl>	<int>	<int>	<date>	<fct>	<fct>	<ord>	<dbl>	<chr>
1	1492	304	3900000	26	0	0.0426	1500	26	610	2018-05-05	5	5	Sat	0.0426	NO
2	1492	352	3360000	42	0	0.0808	800	42	519	2018-05-05	5	5	Sat	0.0809	NO
3	1492	489	2560000	32	0	0.2	800	32	160	2018-05-05	5	5	Sat	0.2	NO
4	1492	429	1200000	12	0	0.110	1000	12	108	2018-05-05	5	5	Sat	0.111	NO
5	1492	524	1200000	15	0	0.167	800	15	89	2018-05-05	5	5	Sat	0.169	NO

INCONSISTENCIAS Y MODIFICACIONES

Capacity y Porcentaje de ocupación con NAs

Casi un 9% de los registros tienen NA en estos valores. Origen del problema: NA en capacity.

Por ahora no se realiza nada sobre estos.

Capacity negativa

54 registros con capacidad negativa.Se invierten a positivo.

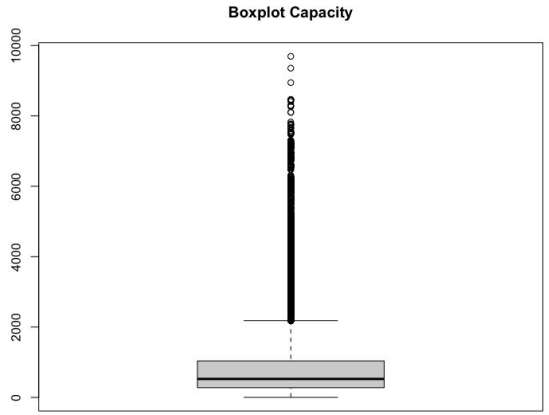
Porcentaje de ocupación mayor a 1

164 registros con un porcentaje mayor al 100% de la capacidad. Se le resta el 1.

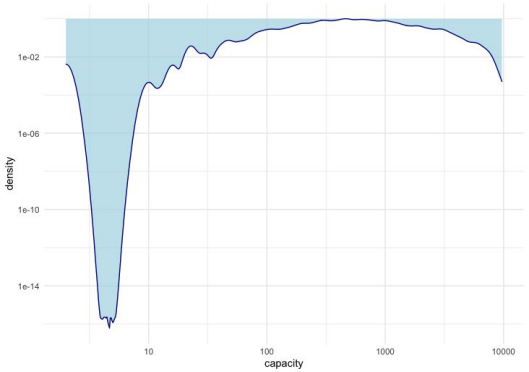
Ticket use menor a 0.

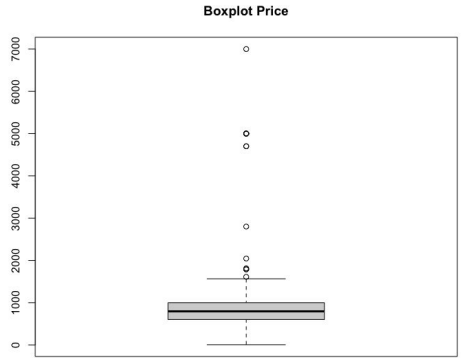
52 registros con un uso de tickets negativos. Es inconsistente. Se eliminan estos registros

ANALISIS DE OUTLIERS

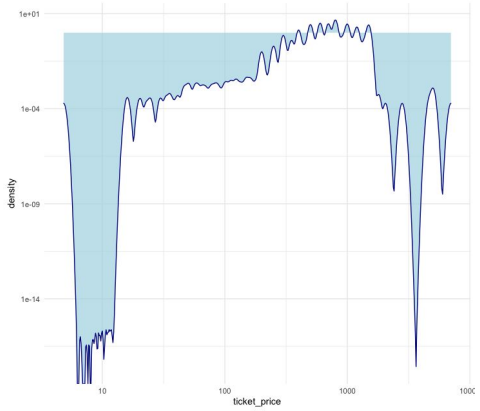


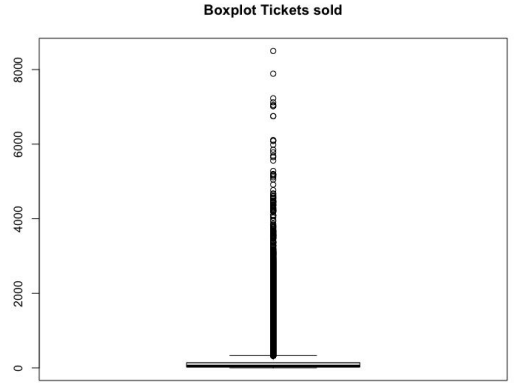

```
   capacity
Min.   :   2.0
1st Qu.: 276.0
Median : 525.0
Mean   : 854.3
3rd Qu.:1038.0
Max.   :9692.0
NA's   :125
```



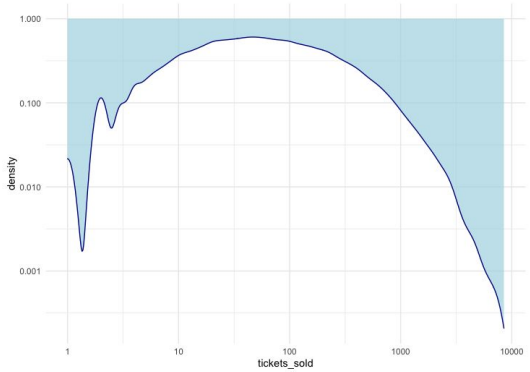


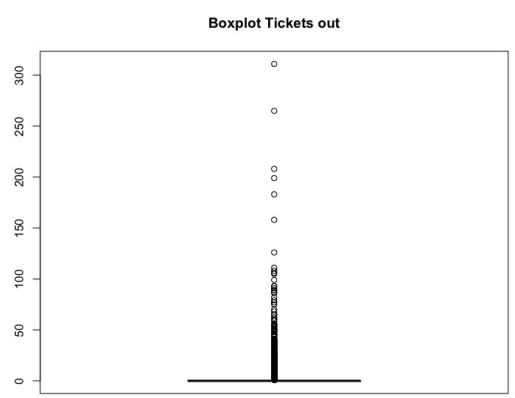

```
  ticket_price
Min.   :   4.839
1st Qu.: 600.000
Median : 794.558
Mean   : 812.335
3rd Qu.:1000.000
Max.   :7000.000
```



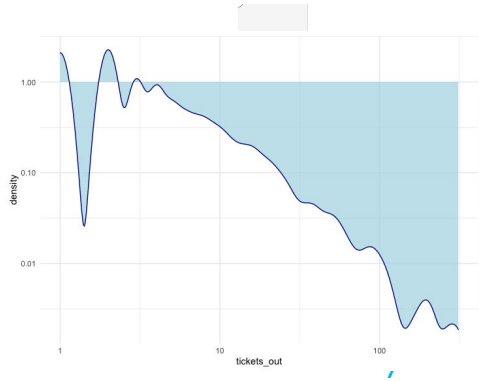


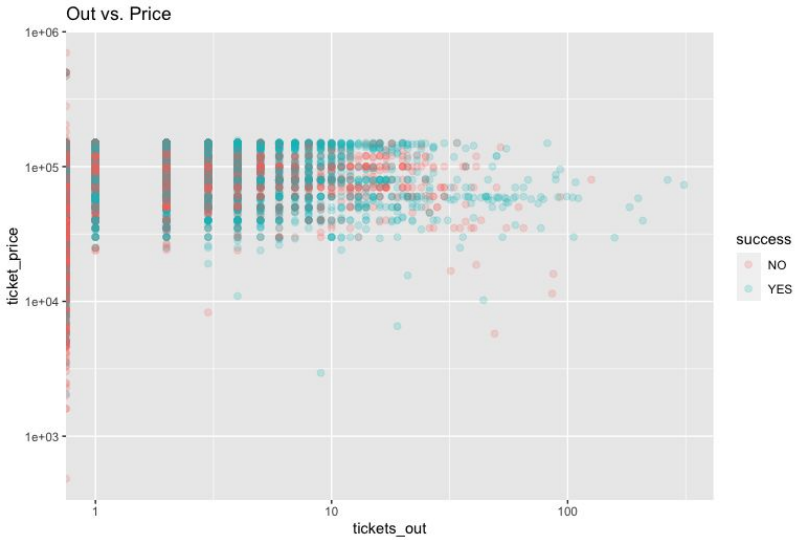

```
  tickets_sold
Min.   :   1.0
1st Qu.:  18.0
Median :  50.0
Mean   : 140.2
3rd Qu.: 143.0
Max.   :8499.0
```

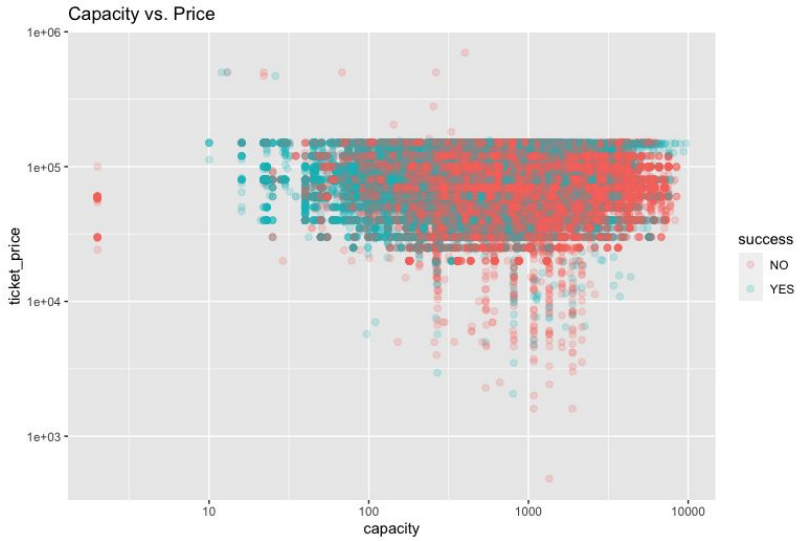





```
  tickets_out
Min.   :  0.000
1st Qu.:  0.000
Median :  0.000
Mean   :  0.213
3rd Qu.:  0.000
Max.   :311.000
```







(muy similar a ticket_use)

En términos generales se detectan tendencias, no correlaciones.

CORRELACIÓN DE SPEARMAN *- ante outliers.*

- La cantidad de tickets no utilizados tiene baja correlación con el resto de las variables.
- El precio influye muy levemente en la cantidad vendida.
- La capacidad no influye en el precio.
- El porcentaje de ocupado real, tiene baja correlación (negativa) con la capacidad.
- El total vendido tiene poca correlación con el precio de ticket.

- Las correlaciones entre las variables categóricas resultan muy pequeñas. (Test V de Cramer)

Modelo: árbol de decisión

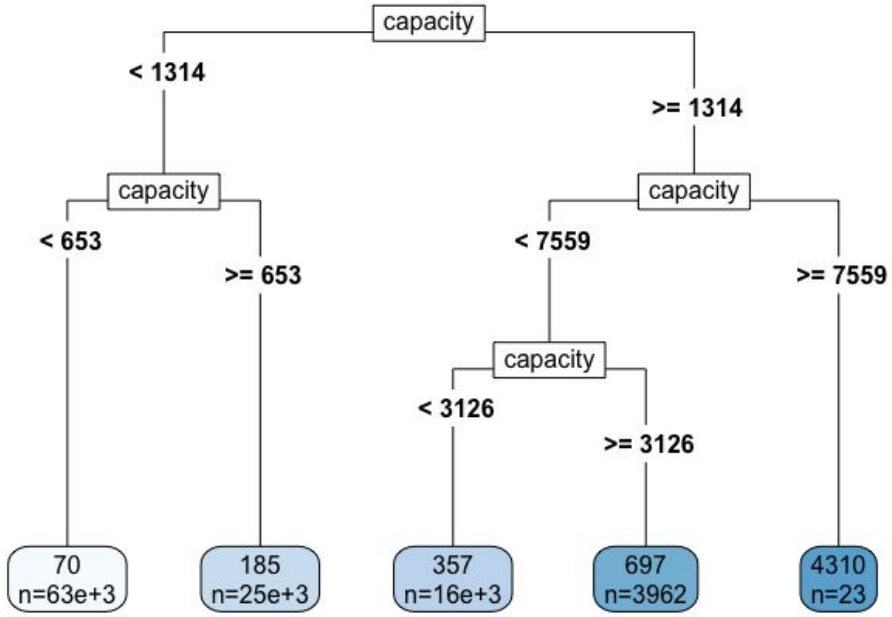
- Se elige un árbol de decisión porque este explícita las variables y condiciones que utiliza para la regresión, lo cual puede ser útil para entender en qué instancias se pueden aplicar distintas campañas o consideraciones.
- Útil en este caso por la falta de linealidades en la información

Criterio de partición: *date*.

La partición se realiza en función de la variable date que indica la cronología. Para esto se asigna al entreno los valores respectivos de los primeros 3Q, y a testeo los valores del último Q.

CREACIÓN DEL MODELO

Árbol de decisión de tipo regresión para predicción de los tickets vendidos.



CONCLUSIONES

POTENCIAL DE LA BASE

- Complemento con el tipo de película a partir de su ID.
- Simulación de la ubicación geográfica (originalmente anónima).
- Su capacidad de actualizarse aumenta su potencial.
- Debería corregirse los errores en distintas variables.
- Diversas aplicaciones: predicción en precios, en cancelaciones, en pérdidas.
- Complemento con resto del negocio: bar, merch, parking.
- Traspolación a eventos de otros venues: estadios, arenas, teatros.

GRACIAS.

Espacio de consulta.