

# ANÁLÍTICA DE DATOS

Estudiante: Danny Sebastián Díaz Padilla

## Taller

Auto Model

Load Data

Select Task

Prepare Target

Select Inputs

Model Types

Results

RESTART

BACK

NEXT

Predict

Want to predict the values of a column?

Clusters

Want to identify groups in your data?

Outliers

Want to detect outliers in your data?

house_sqft	num_of_bedrooms	num_of_bathrooms	year_built	tax_assessed_value	last_sold_price
Number	Number	Number	Number	Number	Number
1770	3	2	1990	195000	196358
1770	3	2	1990	195000	197715
1770	3	2	1990	195000	197816
1772	3	2	1990	200000	198011
1850	3	2.500	1990	200000	200530
1850	3	2.500	1990	200000	201805
1850	3	2.500	1990	205000	206175
1850	3	2.500	1990	205000	207027

120 rows - 6 columns (6 numerical)

Auto Model

Load Data

Select Task

Prepare Target

Select Inputs

Model Types

Results

RESTART

BACK

NEXT

Predict

Want to predict the values of a column?

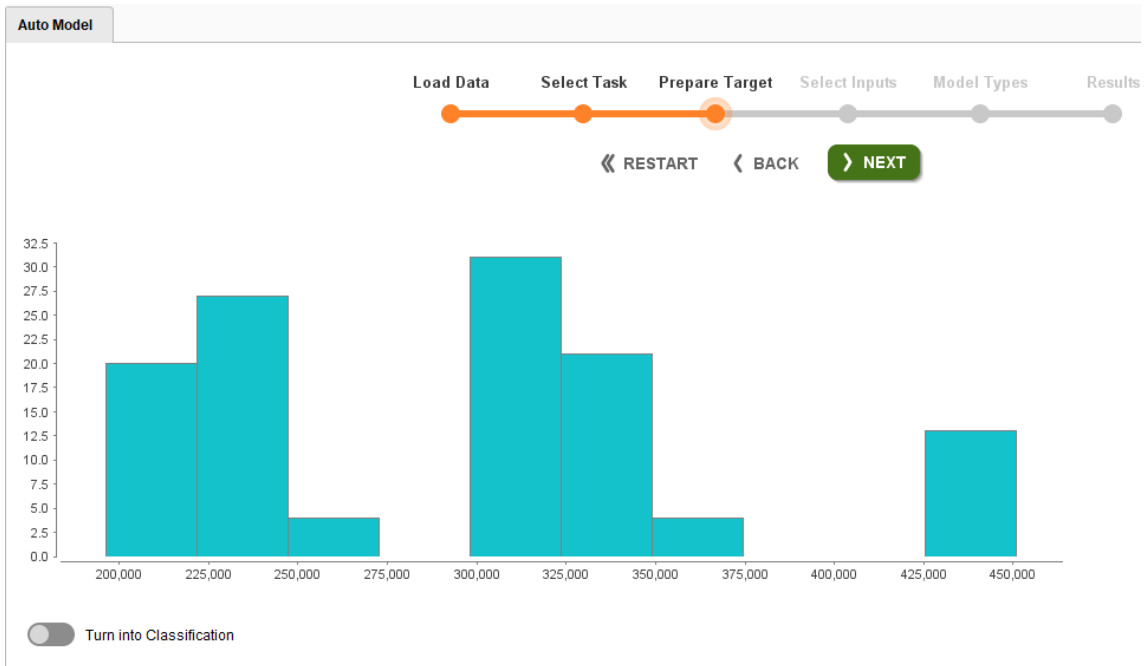
Clusters

Want to identify groups in your data?

Outliers

Want to detect outliers in your data?

# ANALÍTICA DE DATOS



Auto Model

Load Data Select Task Prepare Target Select Inputs Model Types Results

« RESTART < BACK > NEXT

Selected: 4 / Total: 5

☐ Deselect Red ☐ Deselect Yellow ☒ Select All ☒ Deselect All

Selected	Status ↑	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
<input type="checkbox"/>	<span style="color: red;">●</span>	<div><div></div><div>C</div><div>S</div><div>M</div><div>T</div></div>	tax_assessed_value	99.93%	23.33%	8.33%	0.00%	0.00%
<input checked="" type="checkbox"/>	<span style="color: yellow;">●</span>	<div><div></div><div>C</div><div>S</div><div>M</div><div>T</div></div>	house_sqft	41.82%	20.00%	16.67%	0.00%	0.00%
<input checked="" type="checkbox"/>	<span style="color: yellow;">●</span>	<div><div></div><div>C</div><div>S</div><div>M</div><div>T</div></div>	year_built	77.76%	8.33%	16.67%	0.00%	0.00%
<input checked="" type="checkbox"/>	<span style="color: green;">●</span>	<div><div></div><div>C</div><div>S</div><div>M</div><div>T</div></div>	num_of_bedrooms	32.99%	2.50%	65.00%	0.00%	0.00%

Auto Model

Load Data Select Task Prepare Target Select Inputs Model Types Results

« RESTART < BACK ▶ RUN

Execute on: Local Computer (this machine)

Queue: No queues available

Select Folder for Storing Results

The results of this run will be stored in the folder selected below. We recommend to use an empty folder in the selected server repository.

Local Repository (COMPANY)

☒ Generalized Linear Model

☒ Use Regularization ☐ Calculate p-Values

☒ Deep Learning

☒ Decision Tree

☒ Automatically Optimize Maximal Depth: 20

☒ Random Forest

☒ Automatically Optimize Number of Trees: 20 Maximal Depth: 20

☒ Gradient Boosted Trees

☒ Automatically Optimize Number of Trees: 20 Maximal Depth: 20 Learning Rate: 0.01

☒ Support Vector Machine

☐ Remove Columns with Too Many Values

Maximum Number of Values: 50

☐ Extract Date Information

☐ Extract Text Information

Select Text Columns (0)

Number of Extracted Features: 1,000

☐ Automatic Feature Selection

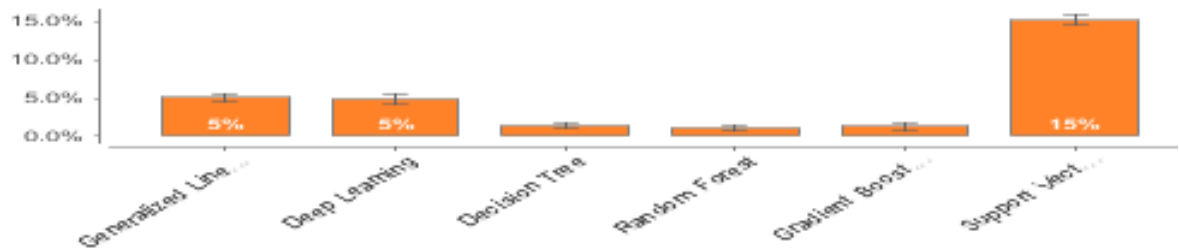
Additional Minutes (Maximum): 60

Final Feature Set should be Accurate

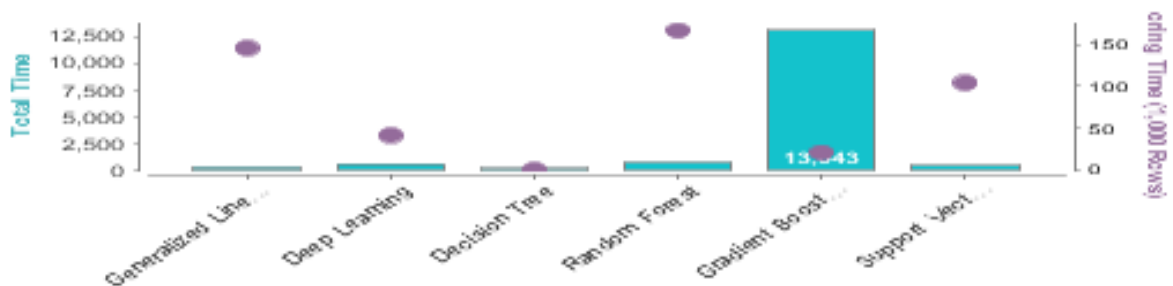
☐ Automatic Feature Generation

## ANALÍTICA DE DATOS

### Relative Error



### Runtimes (ms)



Model	Relative Error	Standard Deviation	Gains	Total Time	Training Time (1,000 Rows)
Generalized Linear Model	5.1%	± 0.4%	?	254 ms	425 ms
Deep Learning	4.9%	± 0.7%	?	596 ms	2 s
Decision Tree	1.4%	± 0.3%	?	136 ms	8 ms

Model	Relative Error	Standard Deviation	Gains	Total Time	Training Time (1,000 Rows)
Random Forest	1.1%	± 0.2%	?	697 ms	100 ms
Gradient Boosted Trees	1.3%	± 0.5%	?	13 s	3 s
Support Vector Machine	15.1%	± 0.6%	?	406 ms	92 ms

### Importar data de casas en venta de una Inmobiliaria

**house\_sqft** – superficie de la casa

**num\_of\_bedrooms** - Número de habitaciones

**num\_of\_bathrooms** - Número de baños

**year\_built** - El año en que se construyó la casa

**tax\_assessed\_value** - Valor de acceso fiscal de la casa

**last\_sold\_price** - Último precio de venta de la casa

**rate\_per\_sqfoot** – Tasa por pie cuadrado.


**home\_type** – (value apartment, townhouse or single family home)

**school\_rating\_1to10** – Calificación de la escuela que le corresponde al lugar


## ANÁLÍTICA DE DATOS

### Identificar los algoritmos con mejor resultados

El modelo con mayor velocidad al otorgar un score y en general es el Árbol de decisiones con solo un 1.4% de error.

Decision Tree		1.4%	± 0.3%	?	136 ms	8 ms
---------------	---	------	--------	---	--------	------

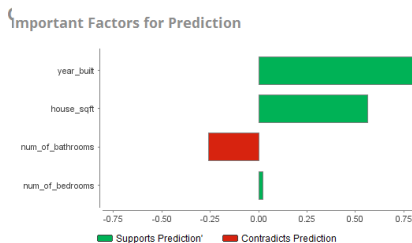
El mejor modelo en cuestión de predicción es “Random Forest” con una menor desviación que el árbol de decisiones pero se demora 5.125 (o mejor dicho 697/136) veces más al entrenar y 12.5 (o mejor dicho 100/8) veces más al otorgar scores.

Random Forest		1.1%	± 0.2%	?	697 ms	100 ms
---------------	---	------	--------	---	--------	--------

### Predecir 10 distintos escenarios de venta de casas con distintos algoritmos

#### A. Algoritmo: Generalized Linear Model

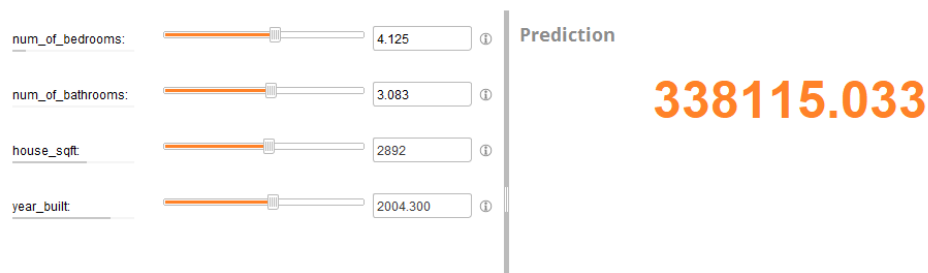
Factores:



Escenarios:

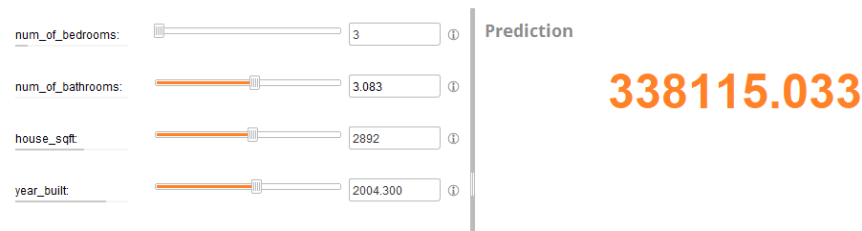
1. Con número promedio de datos.

#### Generalized Linear Model - Simulator



2. Con bajo número de cuartos

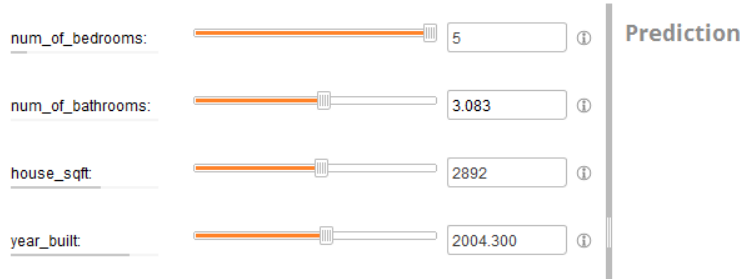
#### Generalized Linear Model - Simulator



## ANALÍTICA DE DATOS

### 3. Con alto número de cuartos

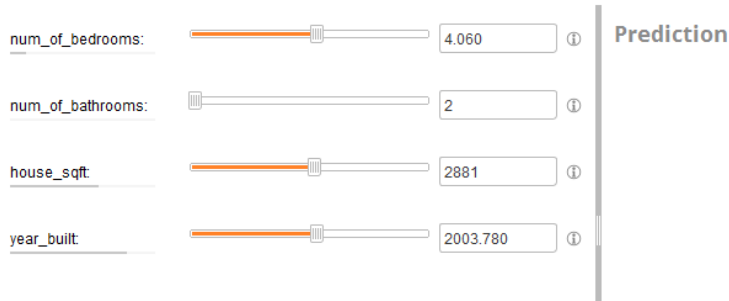
#### Generalized Linear Model - Simulator



338115.033

### 4. Con bajo número de baños

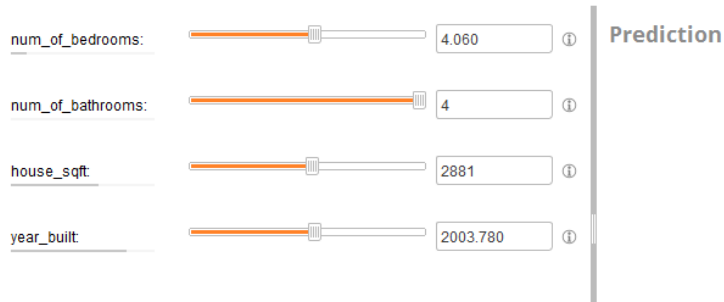
#### Generalized Linear Model - Simulator



362211.067

### 5. Con alto número de baños

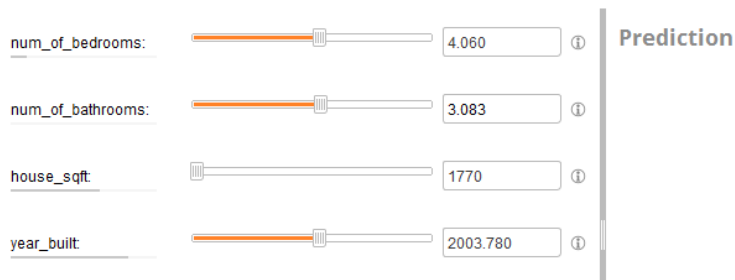
#### Generalized Linear Model - Simulator



310537.660

### 6. Con una superficie pequeña

#### Generalized Linear Model - Simulator

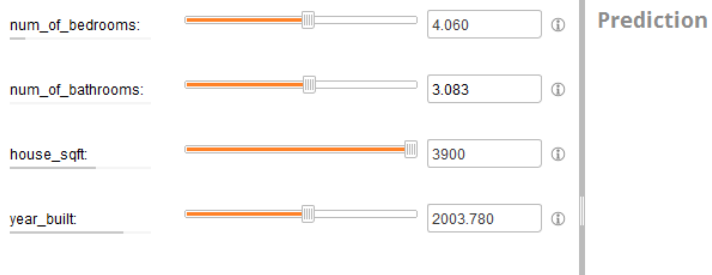


273077.495

## ANALÍTICA DE DATOS

### 7. Con una superficie enorme

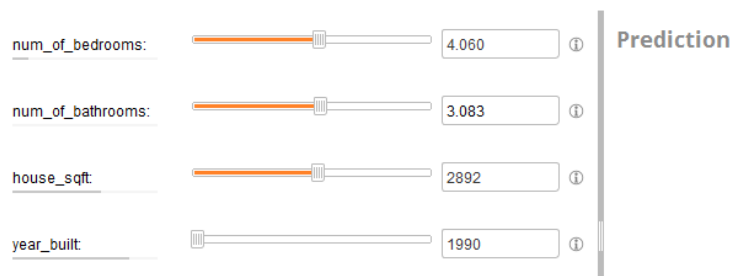
#### Generalized Linear Model - Simulator



390301.901

### 8. Casa antigua

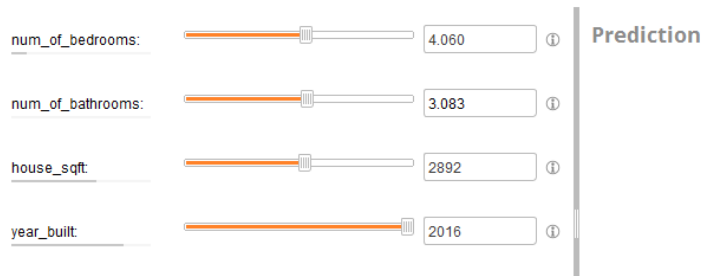
#### Generalized Linear Model - Simulator



247685.578

### 9. Casa moderna

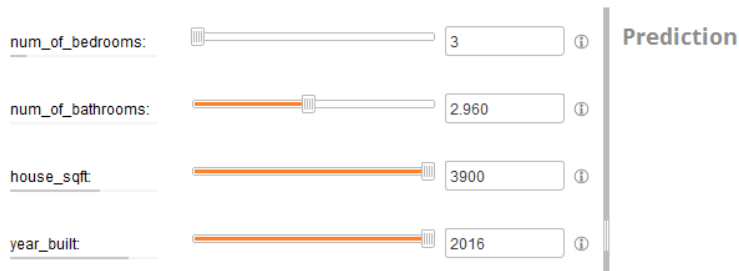
#### Generalized Linear Model - Simulator



412102.769

### 10. Con alta superficie, bajo número de cuartos pero construida recientemente

#### Generalized Linear Model - Simulator



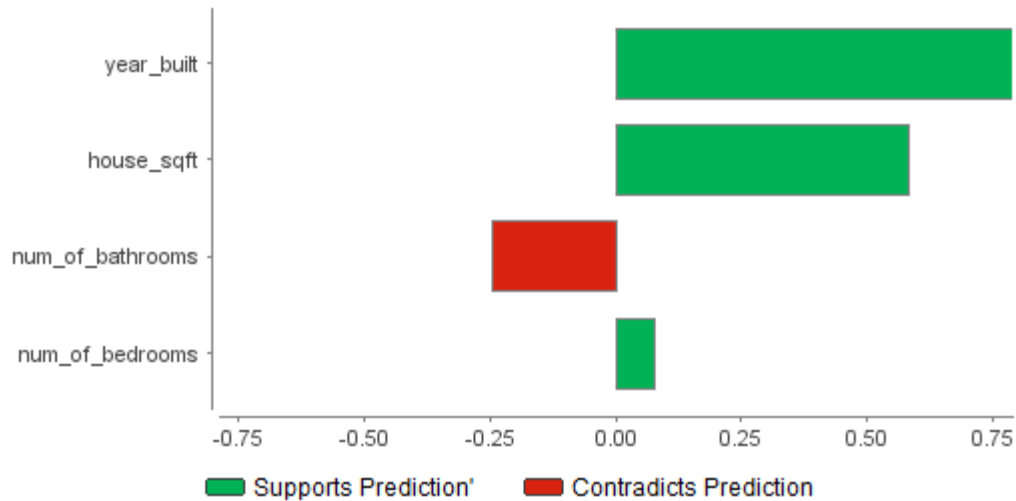
470764.508

## ANALÍTICA DE DATOS

### B. Algoritmo: Deep Learning

Factores:

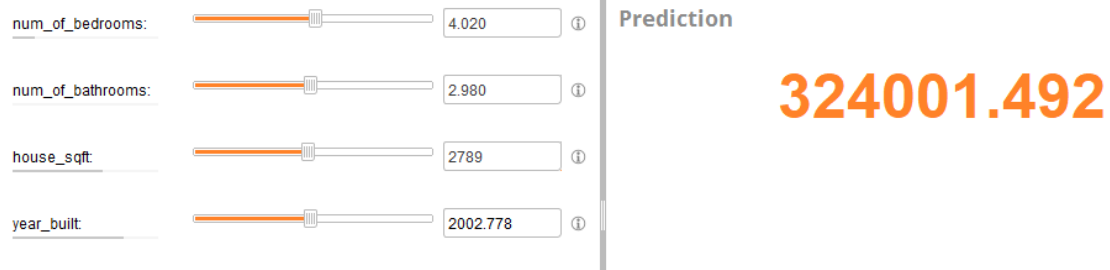
#### Important Factors for Prediction



Escenarios:

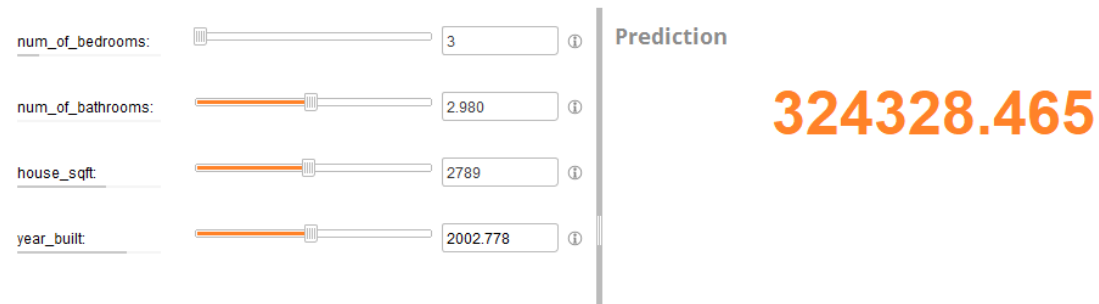
1. Con número promedio de datos

#### Deep Learning - Simulator



2. Con bajo número de cuartos

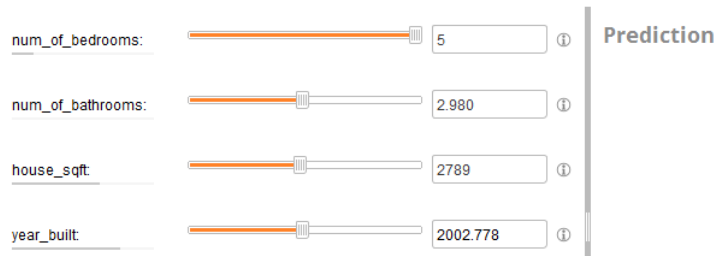
#### Deep Learning - Simulator



## ANALÍTICA DE DATOS

### 3. Con alto número de cuartos

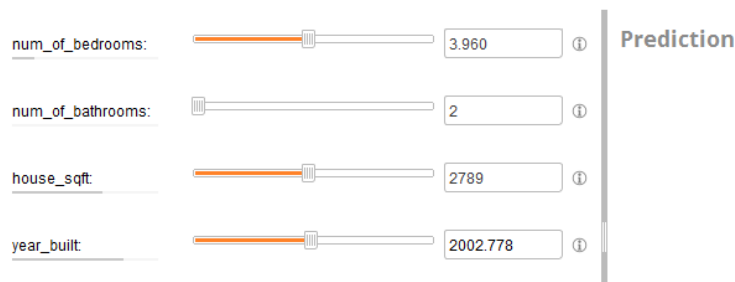
#### Deep Learning - Simulator



329832.858

### 4. Con bajo número de baños

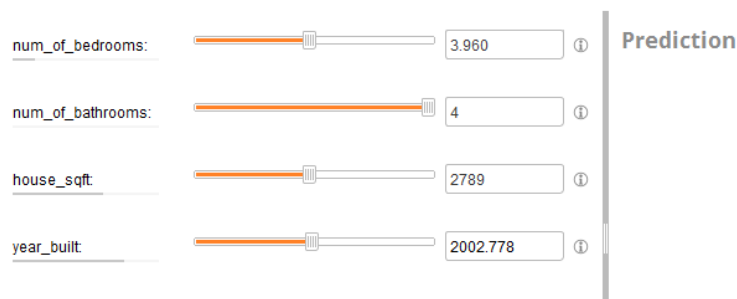
#### Deep Learning - Simulator



349464.358

### 5. Con alto número de baños

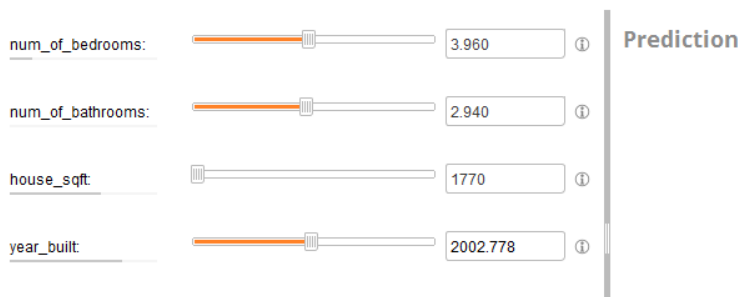
#### Deep Learning - Simulator



301744.643

### 6. Con poca superficie

#### Deep Learning - Simulator



260952.603



## 7. Con gran superficie

### Deep Learning - Simulator

num_of_bedrooms:	<input type="range"/>	<input type="text" value="3.960"/>	①
num_of_bathrooms:	<input type="range"/>	<input type="text" value="2.940"/>	①
house_sqft:	<input type="range"/>	<input type="text" value="3900"/>	①
year_built:	<input type="range"/>	<input type="text" value="2002.778"/>	①

Prediction

382543.379

## 8. Casa antigua

### Deep Learning - Simulator

num_of_bedrooms:	<input type="range"/>	<input type="text" value="3.960"/>	①
num_of_bathrooms:	<input type="range"/>	<input type="text" value="2.940"/>	①
house_sqft:	<input type="range"/>	<input type="text" value="2789"/>	①
year_built:	<input type="range"/>	<input type="text" value="1990"/>	①

Prediction

246634.639

## 9. Casa moderna

### Deep Learning - Simulator

num_of_bedrooms:	<input type="range"/>	<input type="text" value="3.960"/>	①
num_of_bathrooms:	<input type="range"/>	<input type="text" value="2.940"/>	①
house_sqft:	<input type="range"/>	<input type="text" value="2789"/>	①
year_built:	<input type="range"/>	<input type="text" value="2016"/>	①

Prediction

422240.036

## 10. Casa con pocos cuartos, construida actualmente y con gran superficie

### Deep Learning - Simulator

num_of_bedrooms:	<input type="range"/>	<input type="text" value="3"/>	①
num_of_bathrooms:	<input type="range"/>	<input type="text" value="2.940"/>	①
house_sqft:	<input type="range"/>	<input type="text" value="3900"/>	①
year_built:	<input type="range"/>	<input type="text" value="2016"/>	①

Prediction

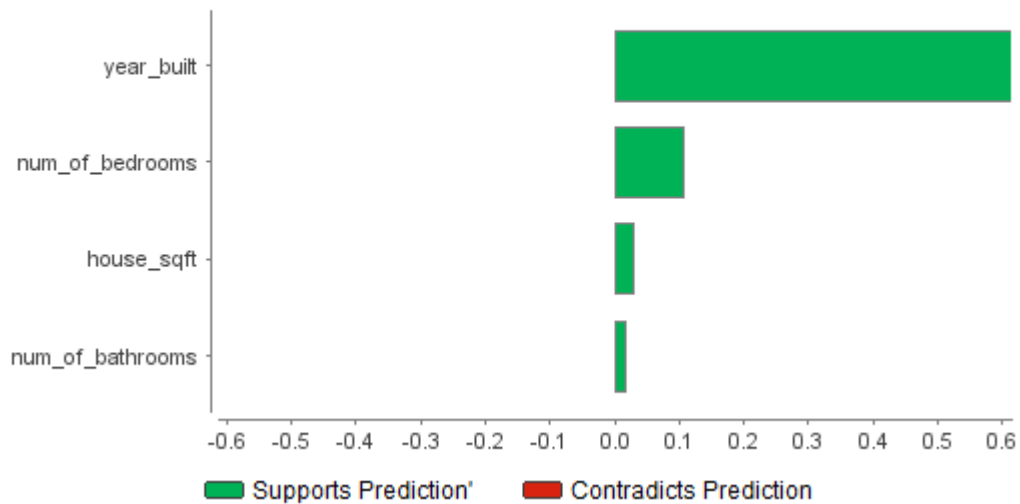
437065.867

## ANALÍTICA DE DATOS

### C. Algoritmo: Decision Tree

Factores:

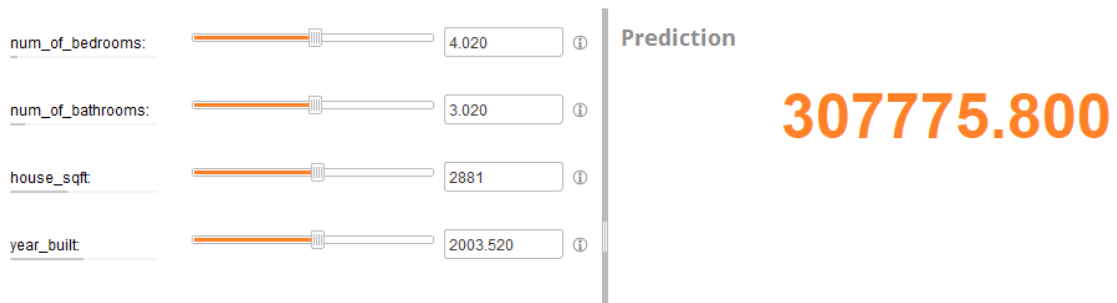
#### Important Factors for Prediction



Escenarios:

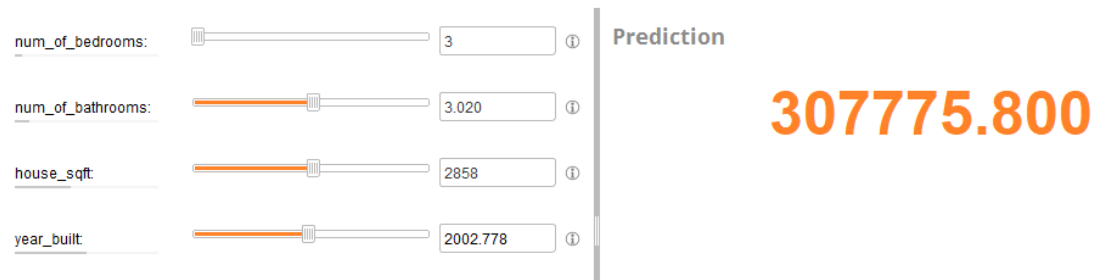
1. Con número promedio de datos

#### Decision Tree - Simulator



2. Con bajo número de cuartos

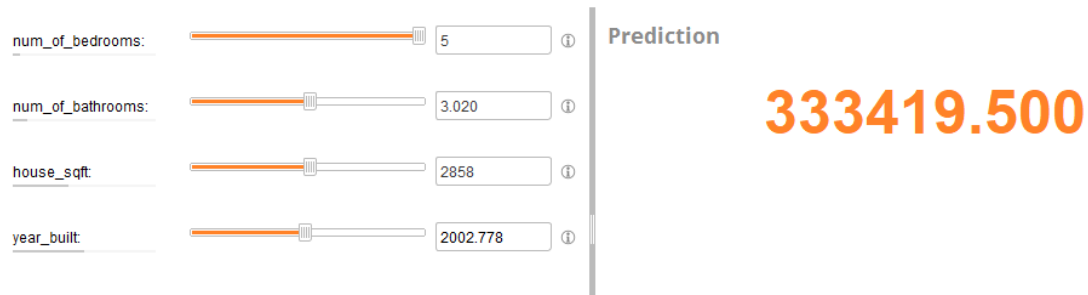
#### Decision Tree - Simulator



## ANALÍTICA DE DATOS

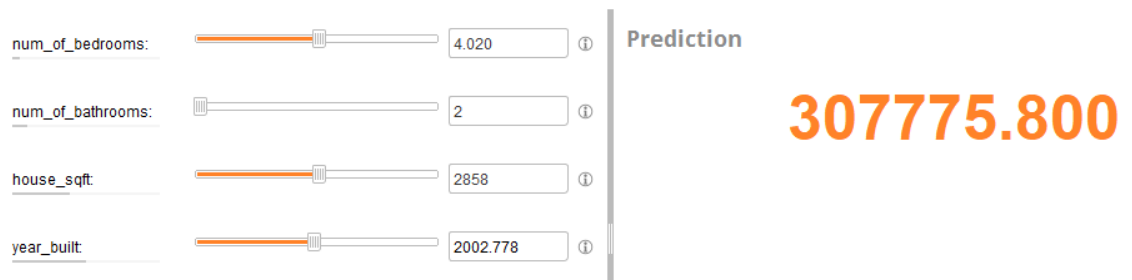
### 3. Con alto número de cuartos

#### Decision Tree - Simulator



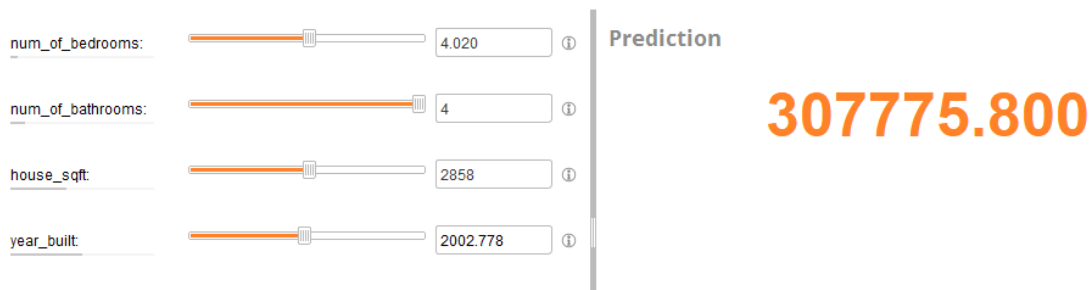
### 4. Con bajo número de baños

#### Decision Tree - Simulator



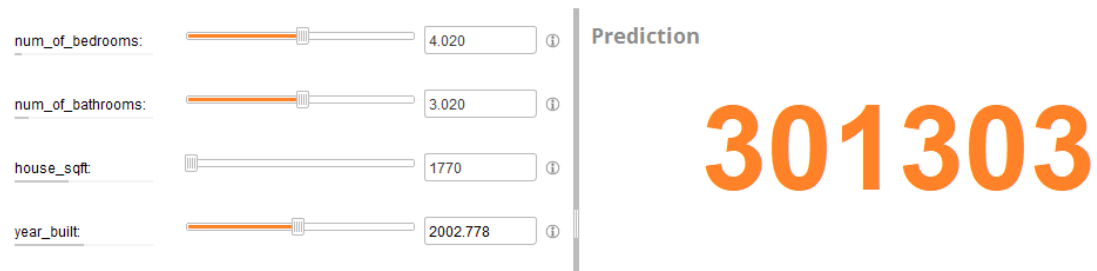
### 5. Con alto número de baños

#### Decision Tree - Simulator



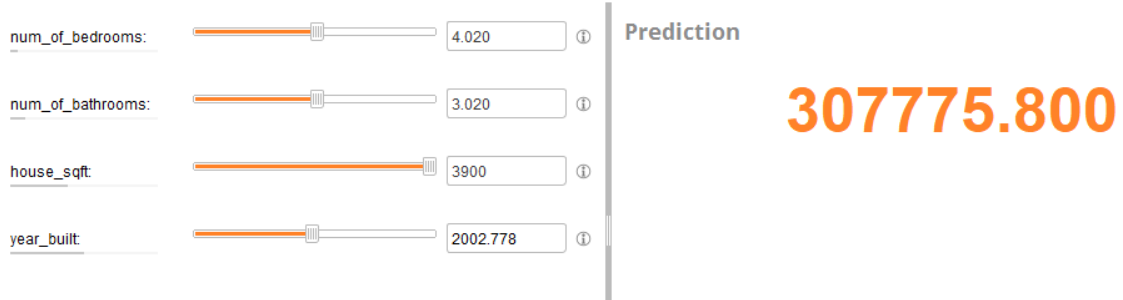
### 6. Con poca superficie

#### Decision Tree - Simulator



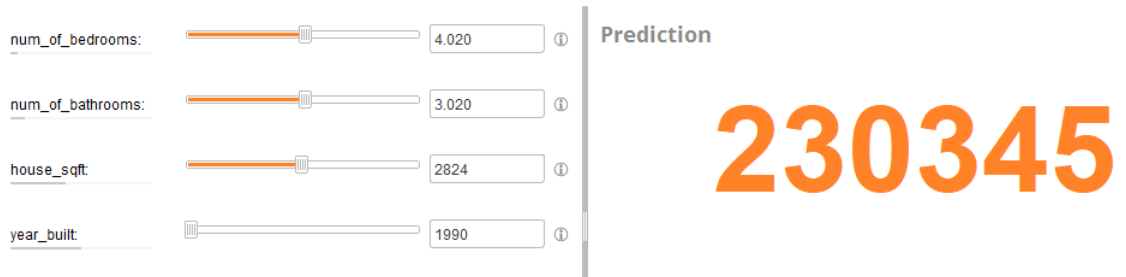
7. Con gran superficie

Decision Tree - Simulator



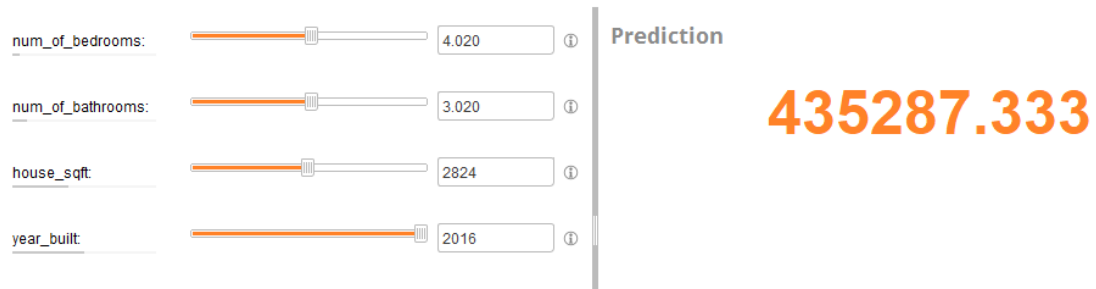
8. Casa antigua

Decision Tree - Simulator



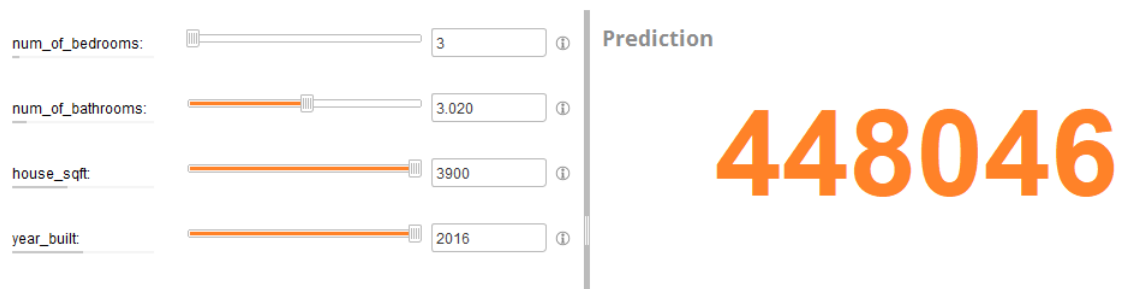
9. Casa moderna

Decision Tree - Simulator



10. Casa con pocos cuartos, construida actualmente y con gran superficie

Decision Tree - Simulator

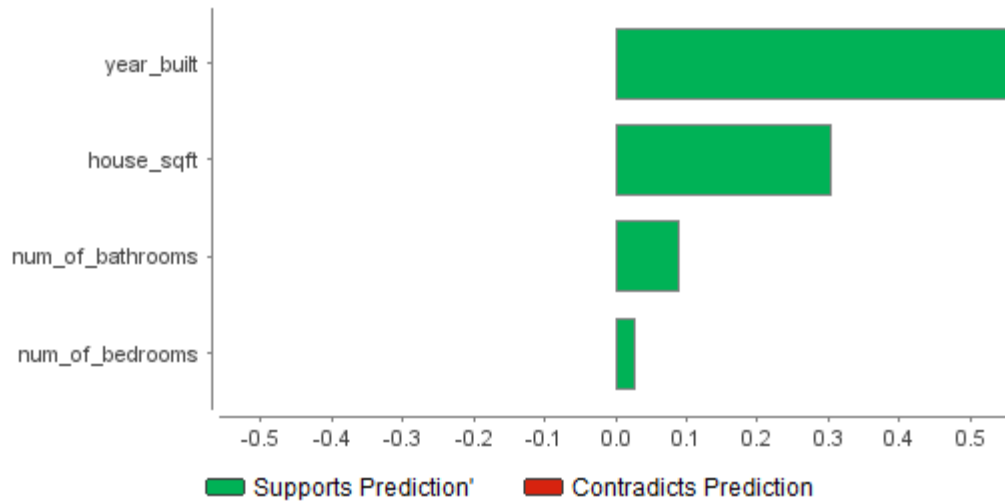


## ANALÍTICA DE DATOS

D. Algoritmo: Random Forest

Factores:

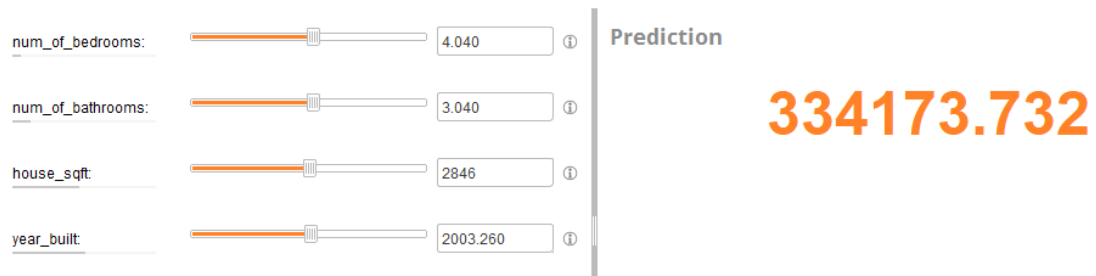
### Important Factors for Prediction



Escenarios:

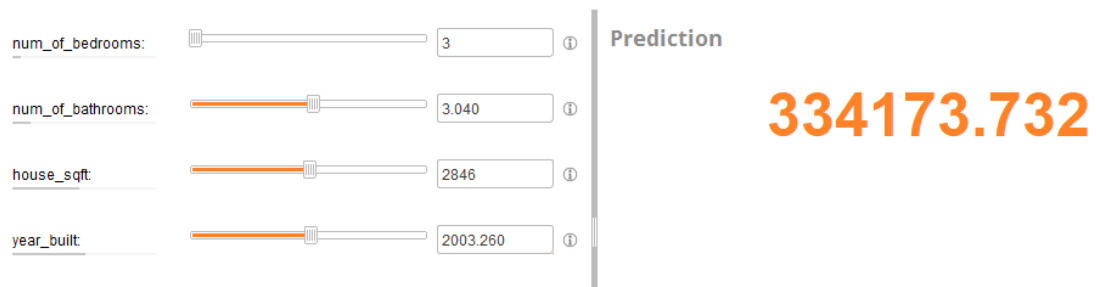
1. Con número promedio de datos

### Random Forest - Simulator



2. Con bajo número de cuartos

### Random Forest - Simulator



## ANALÍTICA DE DATOS

### 3. Con alto número de cuartos

#### Random Forest - Simulator

num_of_bedrooms:	<input type="range" value="5"/>	5	ⓘ
num_of_bathrooms:	<input type="range" value="3.040"/>	3.040	ⓘ
house_sqft:	<input type="range" value="2846"/>	2846	ⓘ
year_built:	<input type="range" value="2003.260"/>	2003.260	ⓘ

Prediction

342996.860

### 4. Con bajo número de baños

#### Random Forest - Simulator

num_of_bedrooms:	<input type="range" value="4.040"/>	4.040	ⓘ
num_of_bathrooms:	<input type="range" value="2"/>	2	ⓘ
house_sqft:	<input type="range" value="2846"/>	2846	ⓘ
year_built:	<input type="range" value="2003.260"/>	2003.260	ⓘ

Prediction

334156.437

### 5. Con alto número de baños

#### Random Forest - Simulator

num_of_bedrooms:	<input type="range" value="4.040"/>	4.040	ⓘ
num_of_bathrooms:	<input type="range" value="4"/>	4	ⓘ
house_sqft:	<input type="range" value="2846"/>	2846	ⓘ
year_built:	<input type="range" value="2003.260"/>	2003.260	ⓘ

Prediction

342698.070

### 6. Con poca superficie

#### Random Forest - Simulator

num_of_bedrooms:	<input type="range" value="4.040"/>	4.040	ⓘ
num_of_bathrooms:	<input type="range" value="3.020"/>	3.020	ⓘ
house_sqft:	<input type="range" value="1770"/>	1770	ⓘ
year_built:	<input type="range" value="2003.260"/>	2003.260	ⓘ

Prediction

304227.007

7. Con gran superficie

Random Forest - Simulator

num_of_bedrooms:	<input type="range"/>	<input type="text" value="4.040"/>	①
num_of_bathrooms:	<input type="range"/>	<input type="text" value="3.020"/>	①
house_sqft:	<input type="range"/>	<input type="text" value="3900"/>	①
year_built:	<input type="range"/>	<input type="text" value="2003.260"/>	①

Prediction

331807.489

8. Casa antigua

Random Forest - Simulator

num_of_bedrooms:	<input type="range"/>	<input type="text" value="4.040"/>	①
num_of_bathrooms:	<input type="range"/>	<input type="text" value="3.020"/>	①
house_sqft:	<input type="range"/>	<input type="text" value="2824"/>	①
year_built:	<input type="range"/>	<input type="text" value="1990"/>	①

Prediction

247992.377

9. Casa moderna

Random Forest - Simulator

num_of_bedrooms:	<input type="range"/>	<input type="text" value="4.040"/>	①
num_of_bathrooms:	<input type="range"/>	<input type="text" value="3.020"/>	①
house_sqft:	<input type="range"/>	<input type="text" value="2824"/>	①
year_built:	<input type="range"/>	<input type="text" value="2016"/>	①

Prediction

436980.420

10. Casa con pocos cuartos, construida actualmente y con gran superficie

Random Forest - Simulator

num_of_bedrooms:	<input type="range"/>	<input type="text" value="3"/>	①
num_of_bathrooms:	<input type="range"/>	<input type="text" value="3.020"/>	①
house_sqft:	<input type="range"/>	<input type="text" value="3900"/>	①
year_built:	<input type="range"/>	<input type="text" value="2016"/>	①

Prediction

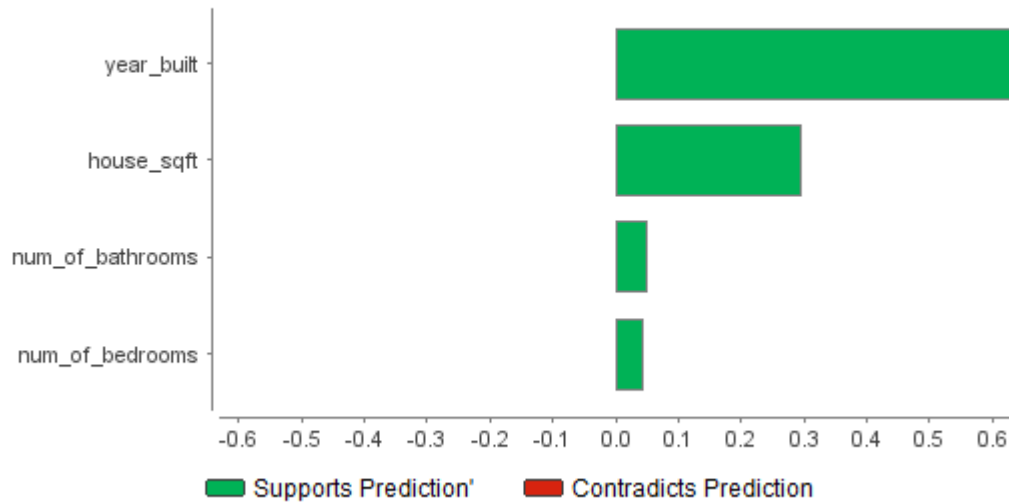
420910.115

## ANALÍTICA DE DATOS

### E. Algoritmo: Gradient Boosted Trees

Factores:

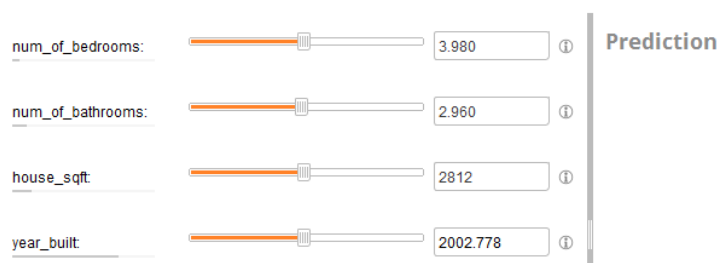
#### Important Factors for Prediction



Escenarios:

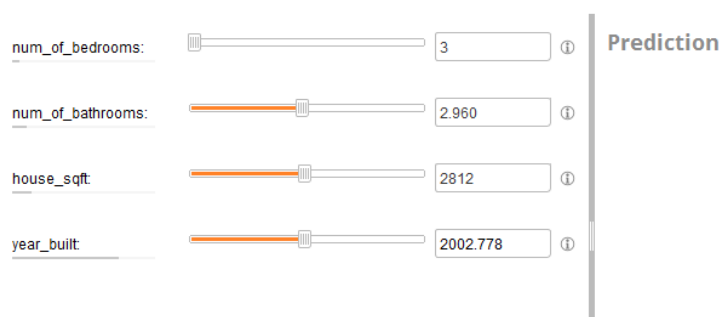
1. Con número promedio de datos

#### Gradient Boosted Trees - Simulator



2. Con bajo número de cuartos

#### Gradient Boosted Trees - Simulator





## ANALÍTICA DE DATOS

### 3. Con alto número de cuartos

#### Gradient Boosted Trees - Simulator

num_of_bedrooms:	<input type="range" value="5"/>	<input type="text" value="5"/>	①
num_of_bathrooms:	<input type="range" value="2.960"/>	<input type="text" value="2.960"/>	①
house_sqft:	<input type="range" value="2812"/>	<input type="text" value="2812"/>	①
year_built:	<input type="range" value="2002.778"/>	<input type="text" value="2002.778"/>	①

Prediction

312337.478

### 4. Con bajo número de baños

#### Gradient Boosted Trees - Simulator

num_of_bedrooms:	<input type="range" value="3.980"/>	<input type="text" value="3.980"/>	①
num_of_bathrooms:	<input type="range" value="2"/>	<input type="text" value="2"/>	①
house_sqft:	<input type="range" value="2812"/>	<input type="text" value="2812"/>	①
year_built:	<input type="range" value="2002.778"/>	<input type="text" value="2002.778"/>	①

Prediction

303783.169

### 5. Con alto número de baños

#### Gradient Boosted Trees - Simulator

num_of_bedrooms:	<input type="range" value="3.980"/>	<input type="text" value="3.980"/>	①
num_of_bathrooms:	<input type="range" value="4"/>	<input type="text" value="4"/>	①
house_sqft:	<input type="range" value="2812"/>	<input type="text" value="2812"/>	①
year_built:	<input type="range" value="2002.778"/>	<input type="text" value="2002.778"/>	①

Prediction

310880.605

### 6. Con poca superficie

#### Gradient Boosted Trees - Simulator

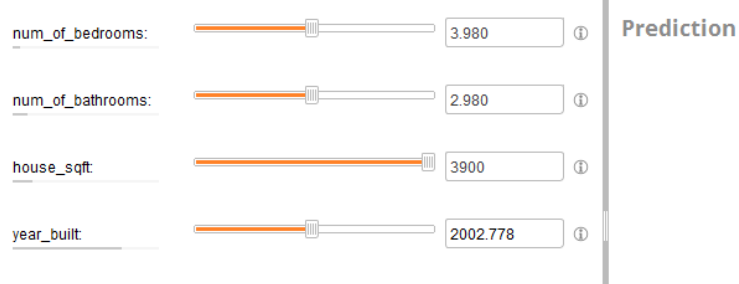
num_of_bedrooms:	<input type="range" value="3.980"/>	<input type="text" value="3.980"/>	①
num_of_bathrooms:	<input type="range" value="2.980"/>	<input type="text" value="2.980"/>	①
house_sqft:	<input type="range" value="1770"/>	<input type="text" value="1770"/>	①
year_built:	<input type="range" value="2002.778"/>	<input type="text" value="2002.778"/>	①

Prediction

306830.077

7. Con gran superficie

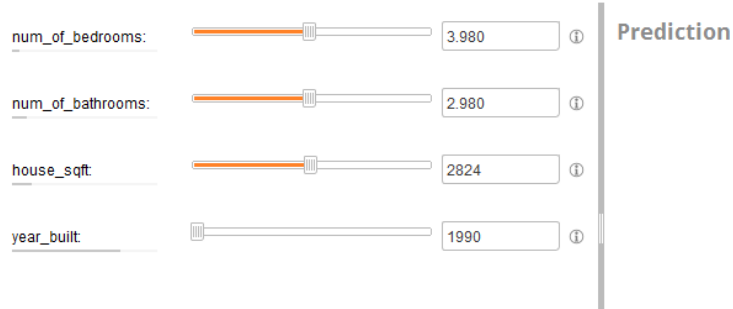
Gradient Boosted Trees - Simulator



327689.165

8. Casa antigua

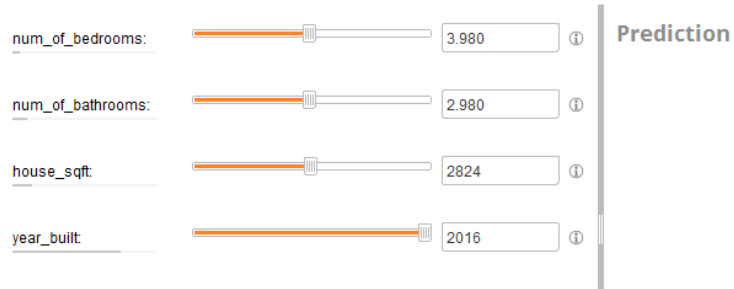
Gradient Boosted Trees - Simulator



213431.249

9. Casa moderna

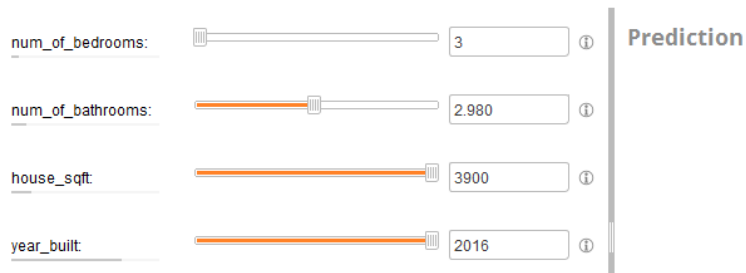
Gradient Boosted Trees - Simulator



437775.922

10. Casa con pocos cuartos, construida actualmente y con gran superficie

Gradient Boosted Trees - Simulator



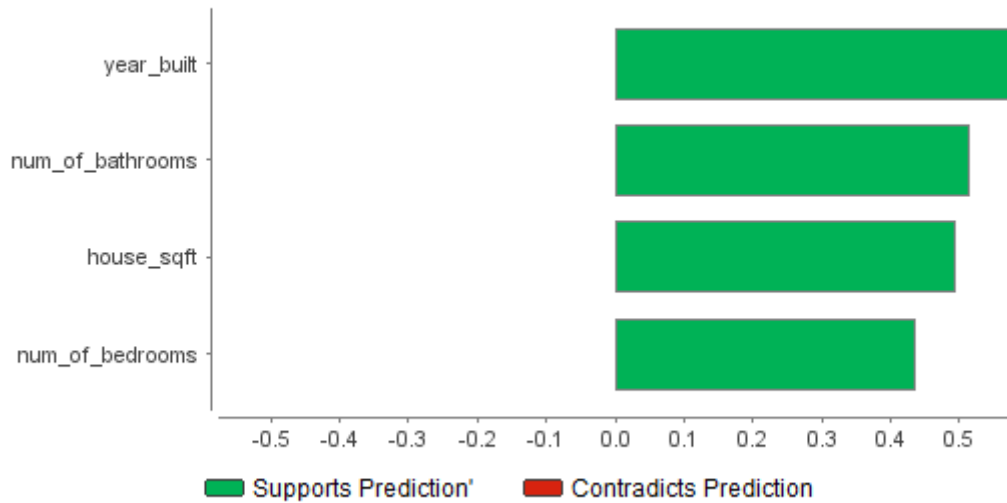
441310.257

## ANALÍTICA DE DATOS

### F. Algoritmo: Support Vector Machine

Factores:

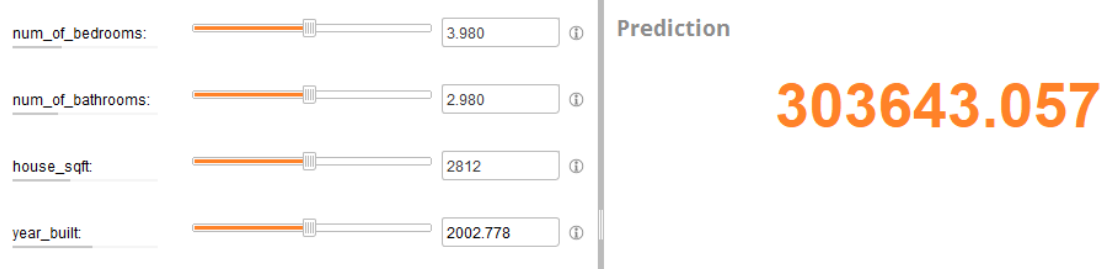
#### Important Factors for Prediction



Escenarios:

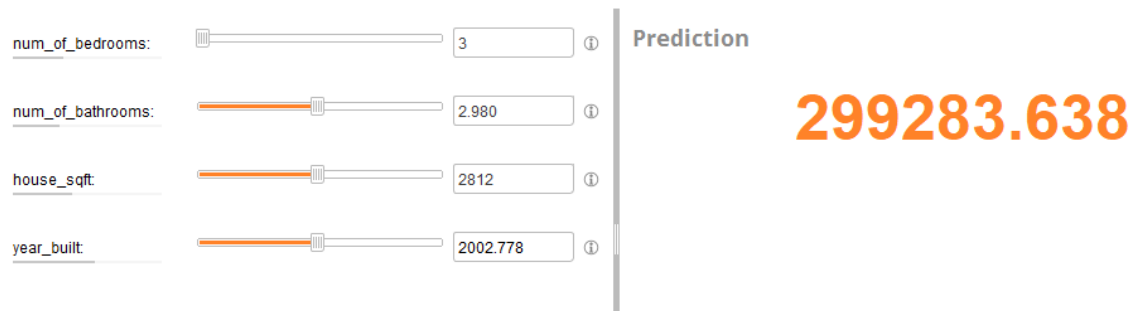
1. Con número promedio de datos

#### Support Vector Machine - Simulator



2. Con bajo número de cuartos

#### Support Vector Machine - Simulator



3. Con alto número de cuartos

Support Vector Machine - Simulator

num_of_bedrooms:	<input type="range" value="5"/>	<input type="text" value="5"/>	ⓘ
num_of_bathrooms:	<input type="range" value="2.980"/>	<input type="text" value="2.980"/>	ⓘ
house_sqft:	<input type="range" value="2812"/>	<input type="text" value="2812"/>	ⓘ
year_built:	<input type="range" value="2002.778"/>	<input type="text" value="2002.778"/>	ⓘ

Prediction

310760.366

4. Con bajo número de baños

Support Vector Machine - Simulator

num_of_bedrooms:	<input type="range" value="3.980"/>	<input type="text" value="3.980"/>	ⓘ
num_of_bathrooms:	<input type="range" value="2"/>	<input type="text" value="2"/>	ⓘ
house_sqft:	<input type="range" value="2812"/>	<input type="text" value="2812"/>	ⓘ
year_built:	<input type="range" value="2002.778"/>	<input type="text" value="2002.778"/>	ⓘ

Prediction

299340.701

5. Con alto número de baños

Support Vector Machine - Simulator

num_of_bedrooms:	<input type="range" value="3.980"/>	<input type="text" value="3.980"/>	ⓘ
num_of_bathrooms:	<input type="range" value="4"/>	<input type="text" value="4"/>	ⓘ
house_sqft:	<input type="range" value="2812"/>	<input type="text" value="2812"/>	ⓘ
year_built:	<input type="range" value="2002.778"/>	<input type="text" value="2002.778"/>	ⓘ

Prediction

311715.984

6. Con poca superficie

Support Vector Machine - Simulator

num_of_bedrooms:	<input type="range" value="3.980"/>	<input type="text" value="3.980"/>	ⓘ
num_of_bathrooms:	<input type="range" value="2.980"/>	<input type="text" value="2.980"/>	ⓘ
house_sqft:	<input type="range" value="1770"/>	<input type="text" value="1770"/>	ⓘ
year_built:	<input type="range" value="2002.778"/>	<input type="text" value="2002.778"/>	ⓘ

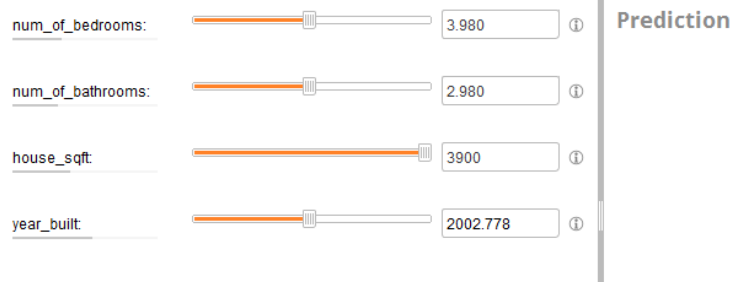
Prediction

296460.738

## ANALÍTICA DE DATOS

### 7. Con gran superficie

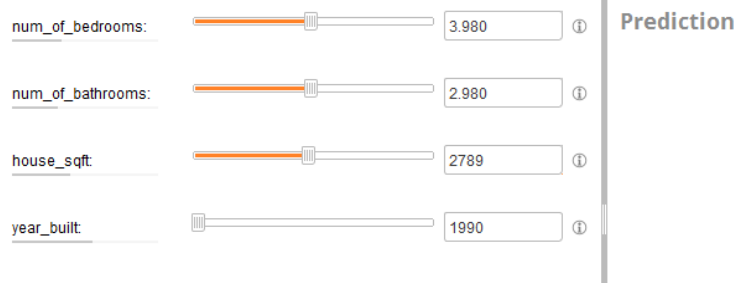
#### Support Vector Machine - Simulator



**310759.009**

### 8. Casa antigua

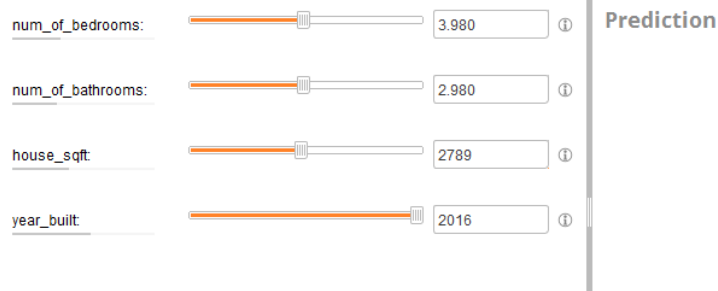
#### Support Vector Machine - Simulator



**295893.752**

### 9. Casa moderna

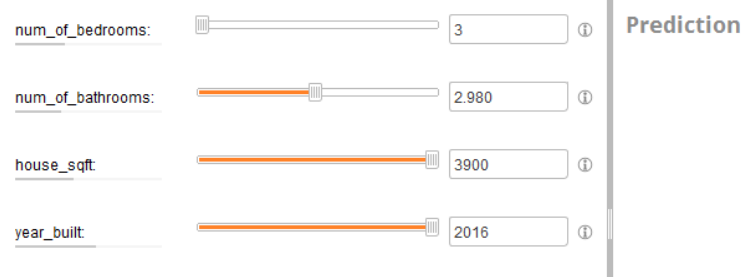
#### Support Vector Machine - Simulator



**312007.469**

### 10. Casa con pocos cuartos, construida actualmente y con gran superficie

#### Support Vector Machine - Simulator



**309162.915**