Assur'aimant
Insurance Charges Prediction

Using `Python` with `Sklearn`

Brief : Prédire une prime d'assurance grâce l'IA

Emad DARWICH **Cédric DUROISIN**

January 17, 2023

Table of Contents

Assur'aimant proposal

- Perform data analysis to better understand Assur'aimant's customers
- Create a solution that would allow Assur'aimant to estimate the insurance premiums of its subscribers in the US market

Insurance Charges Prediction

Our objective is twofold:

Data analysis

Conduct an exploratory data analysis to better understand Assur'aimant's customers

Modeling

Create a machine learning model that estimates customers' insurance charges based on their demographic data.

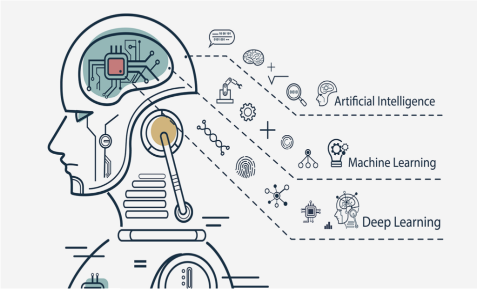


Table of Contents

2 Data analysis

Dataset

Descriptive statistics

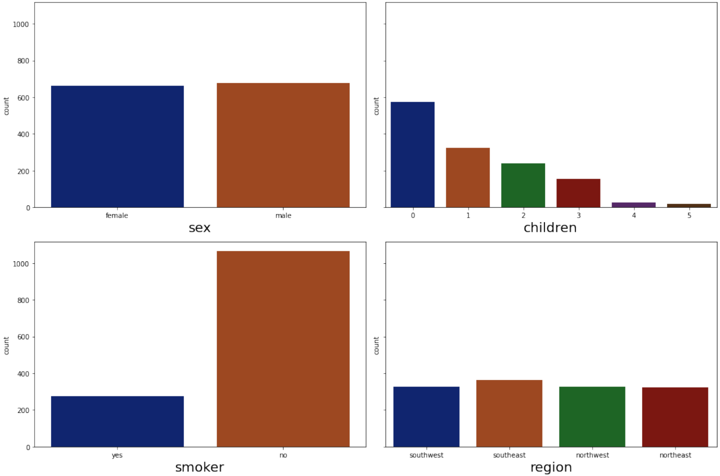
The file Dataset contains:

- 1338 observations
- 7 features
  4 numerical features (age, bmi, children and charges)
  and 3 categorical features (sex, smoker and region)
- no missing values
- and only one duplicate Rows

age	sex	bmi	children	smoker	region	charges
19	male	30.59	0	no	northwest	1639.5631
19	male	30.59	0	no	northwest	1639.5631

Dataset

Categorical features



Dataset
Categorical features

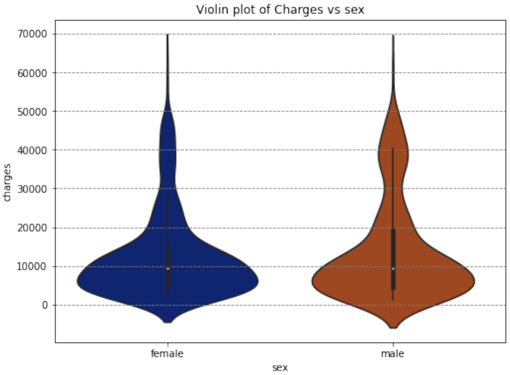
Frequency tables :

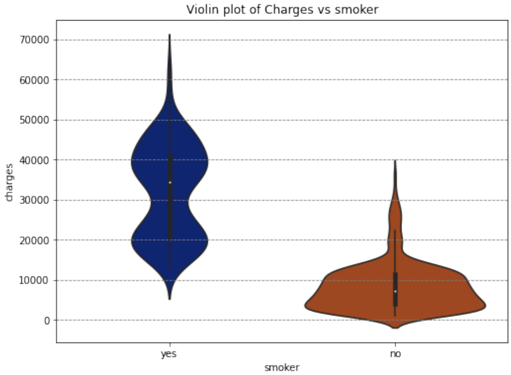
children	n	f	F
0	574	0.428999	0.428999
1	324	0.242152	0.671151
2	240	0.179372	0.850523
3	157	0.117339	0.967862
4	25	0.018685	0.986547
5	18	0.013453	1.000000

smoker	no	yes
sex		
female	547	115
male	517	159

Dataset

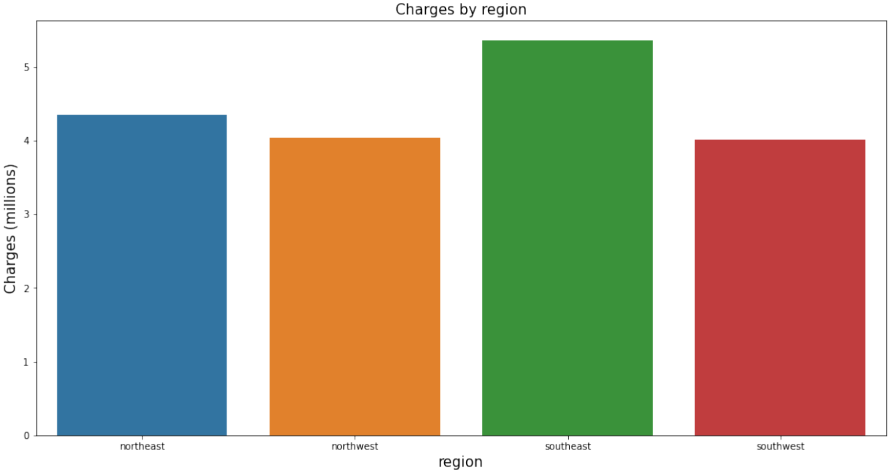
Categorical feature by charges





Dataset

Categorical feature by charges

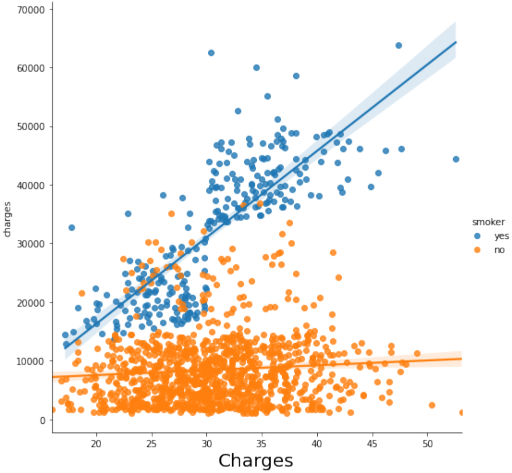


Dataset

Categorical features by charges

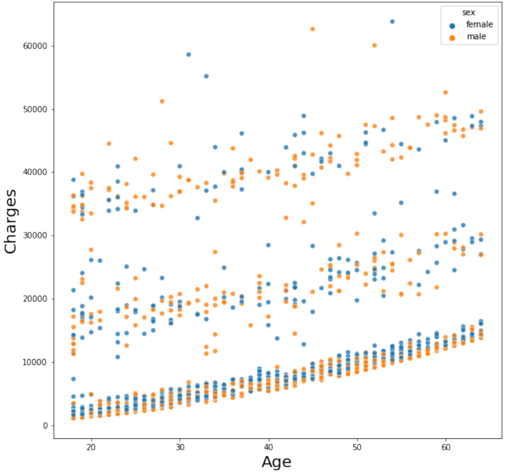
Dataset

Categorical features by charges



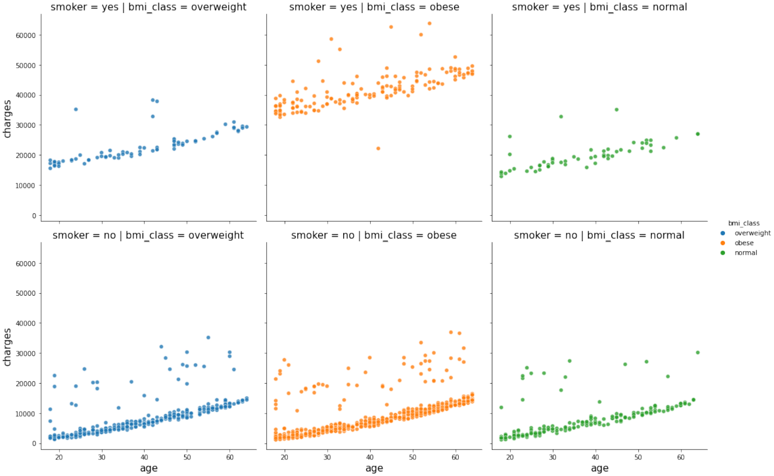
Dataset

Categorical feature by charges



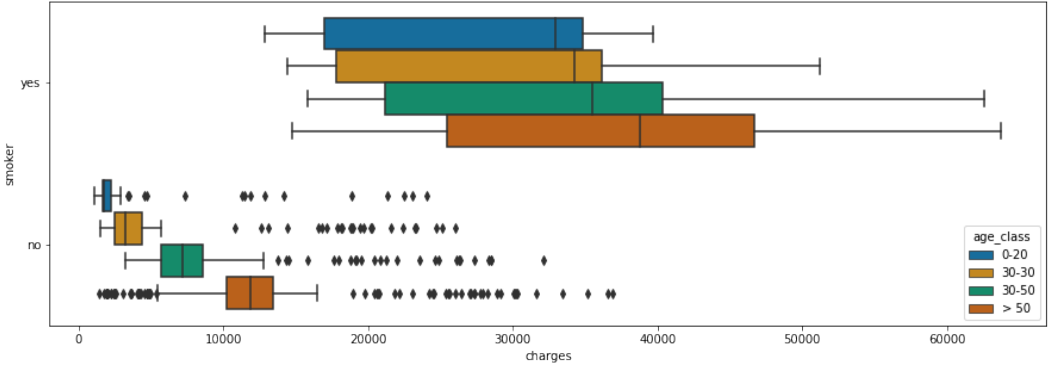
Dataset

Categorical features by charges



Dataset

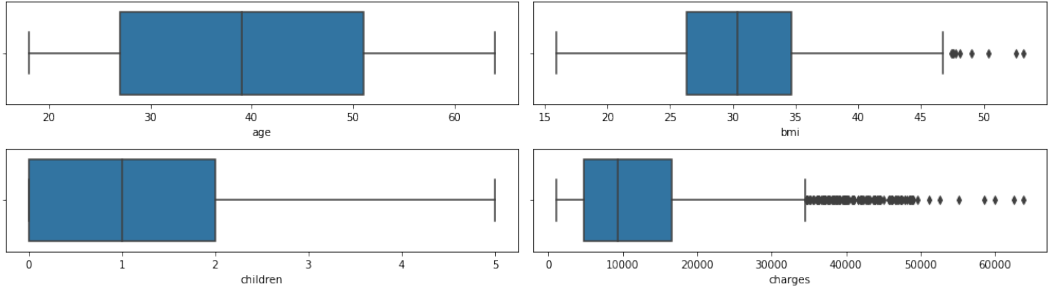
Categorical feature by charges



Dataset

Numerical features

Only bmi and charges have outliers



- bmi has 9 outliers
- charges has 139 outliers

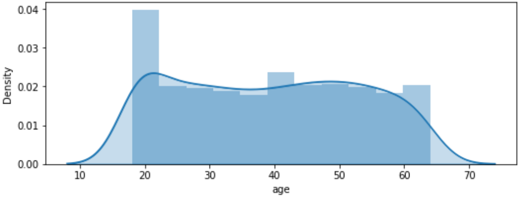
Univariate analysis

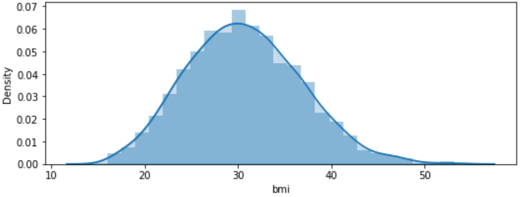
2 Data analysis

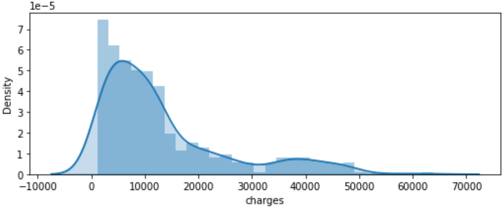
- Testing normality graphically
- Apply log-transformations
- Normality test

Univariate analysis

Numerical features



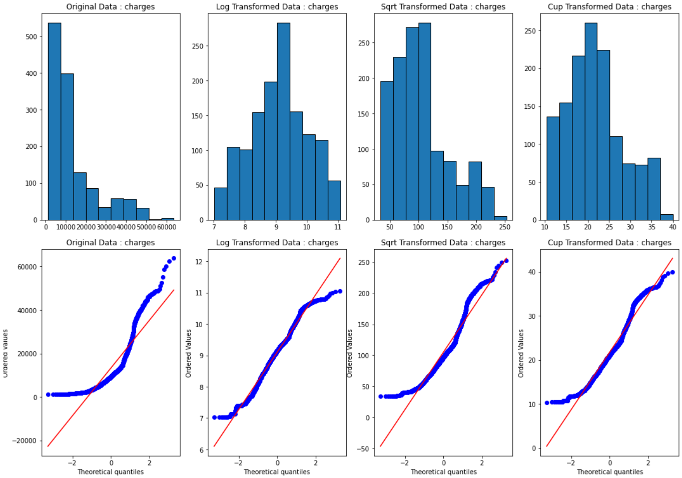




	age	bmi	charges
skew	0.055673	0.284047	1.515880
kurt	-1.245088	-0.050732	1.606299

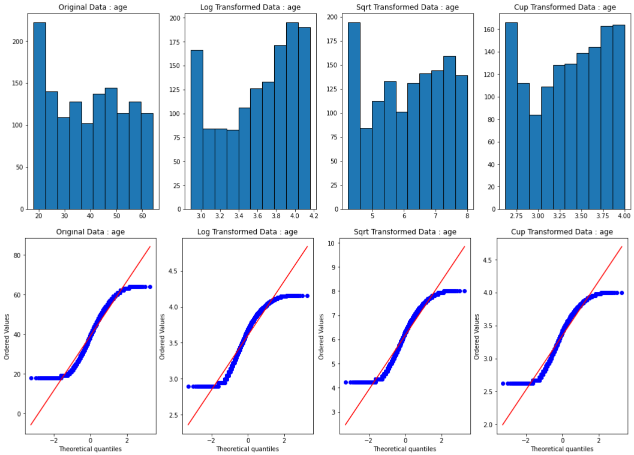
Univariate analysis

Numerical features : charges



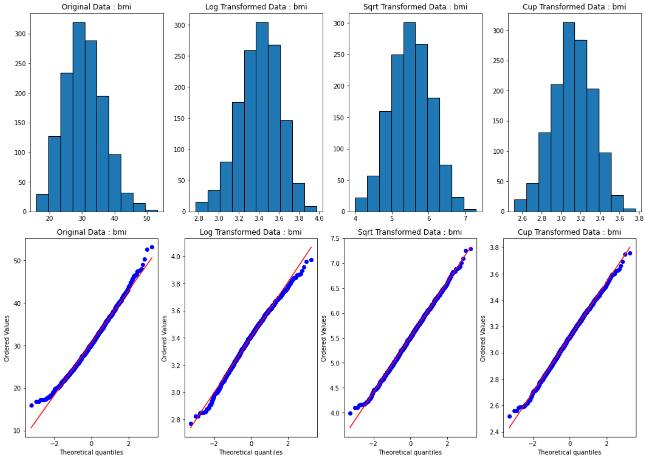
Univariate analysis

Numerical features : age



Univariate analysis

Numerical features : bmi



Univariate analysis

Numerical features : bmi

charges	ORIGIN.		log transf.		sqrt transf.		Cube R transf.	
	stat	pvalue (5%)	stat	pvalue (5%)	stat	pvalue (5%)	stat	pvalue (5%)
Agostino and Pearson	336.885	0.000	52.717	0.000	112.461	0.000	69.040	0.000
Shapiro-Wilk	0.815	0.000	0.983	0.000	0.934	0.000	0.962	0.000
Kolmogorov-Smirnov	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000

age	ORIGIN.		log transf.		sqrt transf.		Cube R transf.	
	stat	pvalue (5%)	stat	pvalue (5%)	stat	pvalue (5%)	stat	pvalue (5%)
Agostino and Pearson	1557.821	0.000	593.282	0.000	1349.019	0.000	1067.809	0.000
Shapiro-Wilk	0.945	0.000	0.930	0.000	0.942	0.000	0.939	0.000
Kolmogorov-Smirnov	1.000	0.000	0.998	0.000	1.000	0.000	0.996	0.000

bmi	ORIGIN.		log transf.		sqrt transf.		Cube R transf.	
	stat	pvalue (5%)	stat	pvalue (5%)	stat	pvalue (5%)	stat	pvalue (5%)
Agostino and Pearson	17.581	0.000	15.400	0.000	2.851	0.240	4.127	0.127
Shapiro-Wilk	0.994	0.000	0.995	0.000	0.999	0.345	0.998	0.169
Kolmogorov-Smirnov	1.000	0.000	0.997	0.000	1.000	0.000	0.994	0.000

sqrt transf. (y) = $y^{\frac{1}{2}}$ Cube Root transformation (y) = $y^{\frac{1}{3}}$

Here we test the null hypothesis that a sample comes from a normal distribution

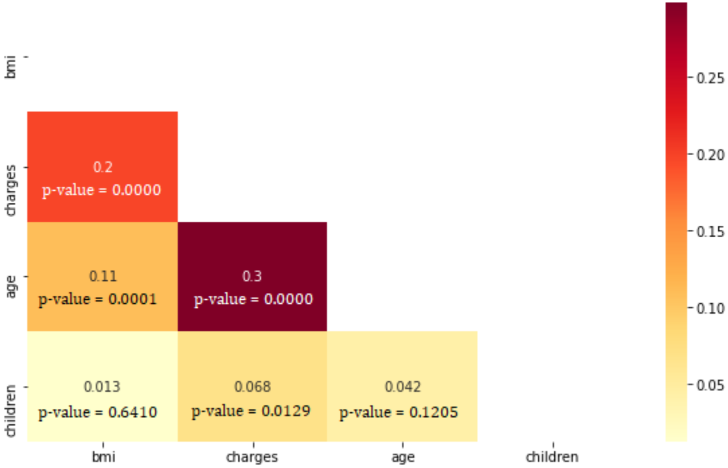
Correlation

2 Data analysis

- Pearson correlation
- Point Biserial test for correlation
- χ^2 test for correlation
- ANOVA

Correlation

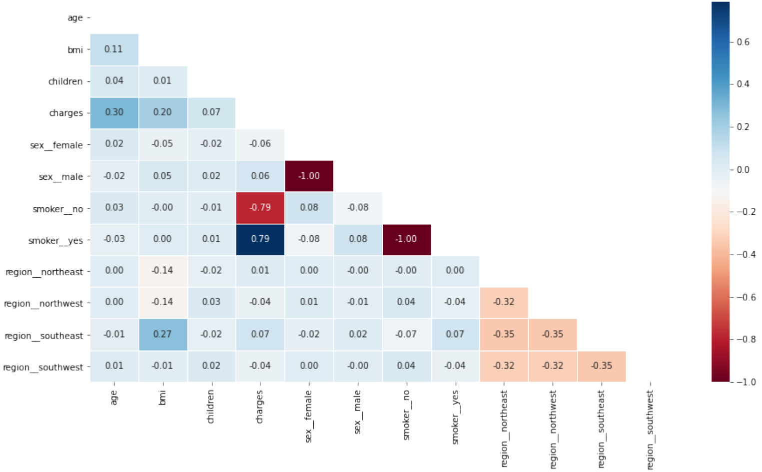
Pearson correlation



Here we test the null hypothesis that there is a statistically significant association between groups

Correlation

Pearson correlation



Correlation

Point Biserial correlation

Corr. : "charges" and "sex"

```
Point Biserial corr. : 0.057
t-value: -2.098
p-value: 0.036
```

Corr. : "charges" and "smoker"

```
Point Biserial corr. : -0.787
t-value: 46.665
p-value: 0.000
```

Here we test the null hypothesis that the correlation is statistically significant.

Correlation

χ^2 test for correlation

p-value	sex	region	smoker	bmi class
sex				
region	0,9239			
smoker	0.0062	0,063548		
bmi class	0.2251	0.00005	0,609511	

H0: The variables are not correlated with each other (Independent).

Correlation
ANOVA

```
--------------------------------------------
| ANOVA between "charges" and "children" |
--------------------------------------------
F-value: 3.297                 p-value: 0.006
```

```
H0:there is no difference in means
H1: at least two means differ by comparing two groups.

        ASSUMPTION CHECK (Normality - Shapiro)
==================================================
The assumption of normality is tested on the residuals
Residuals are not normal (stat (W) =0.812, p=0.000)


             Kruskal-Wallis test
==================================================
statistic : 29.4871                  pvalue : 0.0
No significant differences between categories
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	365.1962	0.9	-2026.0598	2756.4522	False
0	2	2707.5881	0.0413	62.3365	5352.8398	True
0	3	2989.3428	0.0662	-109.9976	6088.6831	False
0	4	1484.6807	0.9	-5546.1043	8515.4657	False
0	5	-3579.9404	0.7925	-11817.2428	4657.3621	False
1	2	2342.3919	0.2025	-588.3514	5273.1352	False
1	3	2624.1465	0.2211	-722.1665	5970.4596	False
1	4	1119.4845	0.9	-6023.6128	8262.5817	False
1	5	-3945.1366	0.7286	-12278.5064	4388.2332	False
2	3	281.7546	0.9	-3250.5338	3814.0431	False
2	4	-1222.9074	0.9	-8454.9949	6009.18	False
2	5	-6287.5285	0.2705	-14697.3027	2122.2458	False
3	4	-1504.6621	0.9	-8914.9012	5905.577	False
3	5	-6569.2831	0.2432	-15132.7437	1994.1775	False
4	5	-5064.6211	0.7242	-15702.2375	5572.9954	False

Reject = True, means statistically significant difference.

Correlation
ANOVA

```
-----------------------------------------
| ANOVA between "charges" and "children" |
-----------------------------------------
F-value: 27.952            p-value: 0.000
```

```
H0:there is no difference in means
H1: at least two means differ by comparing two groups.

        ASSUMPTION CHECK (Normality - Shapiro)
==================================================
The assumption of normality is tested on the residuals
Residuals are not normal (stat (W) =0.862, p=0.000)


             Kruskal-Wallis test
==================================================
statistic : 15.8076               pvalue : 0.0004
No significant differences between categories
```

```
     Multiple Comparison of Means - Tukey HSD, FWER=0.05
================================================================
group1   group2    meandiff  p-adj    lower      upper    reject
----------------------------------------------------------------
normal      obese   5270.111    0.0  3204.8498  7335.3722   True
normal overweight   705.2854 0.7474 -1570.2691  2980.8399  False
 obese overweight -4564.8256    0.0 -6327.8405 -2801.8107   True
----------------------------------------------------------------

Reject = True, means statistically significant difference.
```

Correlation

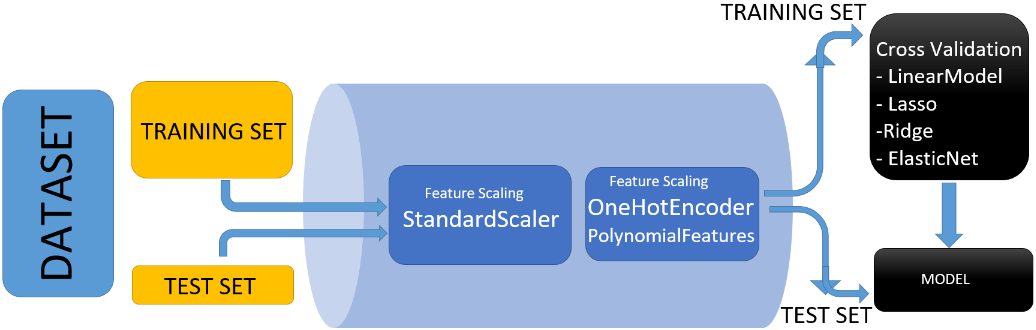
ANOVA

```
---------------------------------------
| ANOVA between "charges" and "region" |
---------------------------------------
F-value: 2.970            p-value: 0.031


H0:there is no difference in means
H1: at least two means differ by comparing two groups.

        ASSUMPTION CHECK (Normality - shapiro)
==================================================
The assumption of normality is tested on the residuals
Residuals are not normal (stat (W) =0.827, p=0.000)


             Kruskal-Wallis test
==================================================
statistic : 4.7342              pvalue : 0.1923
No significant differences between categories
```

```
       Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================================
  group1    group2    meandiff  p-adj    lower      upper   reject
-----------------------------------------------------------------
northeast northwest  -988.8091 0.7245 -3428.9343 1451.3161  False
northeast southeast  1329.0269 0.4745 -1044.9417 3702.9955  False
northeast southwest -1059.4471 0.6792 -3499.5723 1380.6781  False
northwest southeast  2317.8361 0.0583   -54.1994 4689.8716  False
northwest southwest    -70.638 0.9999 -2508.8826 2367.6066  False
southeast southwest -2388.4741 0.0477 -4760.5096  -16.4386   True
-----------------------------------------------------------------

Reject = True, means statistically significant difference.
```

Table of Contents

Machine Learning
Pipeline



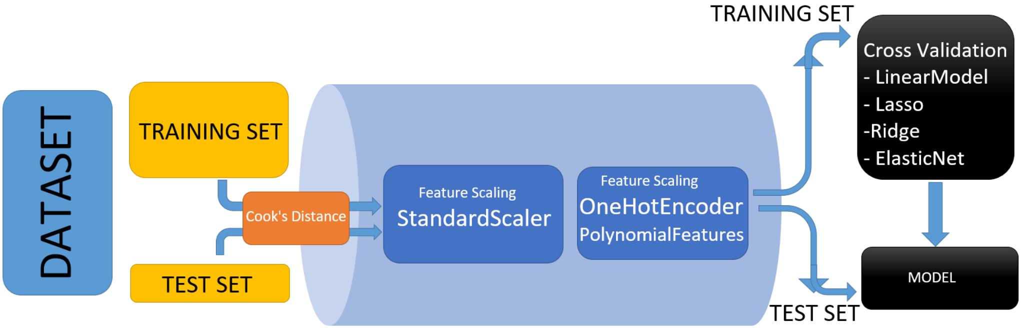
Machine Learning

Models

Model	R2	MAE	RMSE	Score (test)	Score (trainging)	Score (CV)
LASSO (Polynomial=2)	0.9124	1957.294489	3340.68227	0.92258252	0.858246217	0.8488 (+/- 0.04)
ElasticNet (Polynomial=2)	0.9124	1960.457382	3335.03664	0.92284396	0.857777099	0.8488 (+/- 0.04)
Ridg (Polynomial=2)	0.9088	2029.465675	3419.33535	0.91889416	0.861147147	0.8432 (+/- 0.04)
LR (Polynomial=2)	0.908	2059.129706	3441.72265	0.91782865	0.861231494	0.8395 (+/- 0.04)
LASSO (Polynomial=1)	0.7717	3655.59983	5092.51496	0.8200991	0.739238658	0.7286 (+/- 0.04)
ElasticNet (Polynomial=1)	0.7724	3662.001424	5094.07174	0.81998909	0.739370553	0.7285 (+/- 0.04)
LR (Polynomial=1)	0.7759	3708.067361	5107.96534	0.81900583	0.739694812	0.7282 (+/- 0.04)
Ridg (Polynomial=1)	0.7743	3706.798635	5107.54383	0.8190357	0.739682407	0.7282 (+/- 0.04)

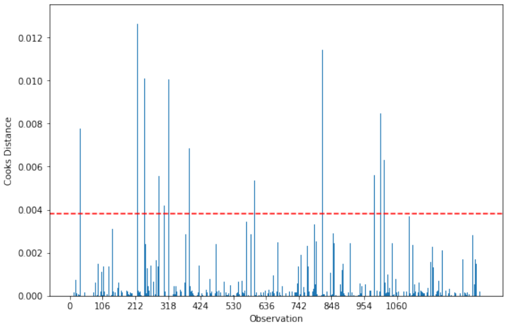
Machine Learning

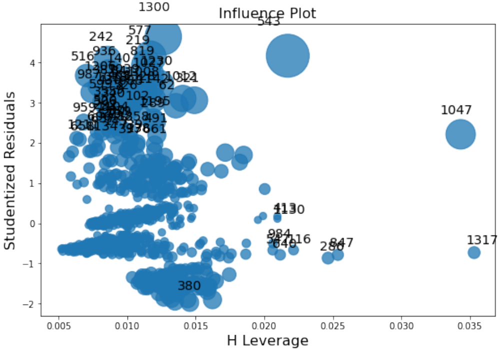
Pipeline



Machine Learning

Cook's distance





Machine Learning
Models

Model	R2	MAE	RMSE	Score (test)	Score (trainging)	Score (CV)
LASSO (Polynomial degree=2)	0.9167	1406.28958	3284.14811	0.92518061	0.93174224	0.9289 (+/- 0.03)
ElasticNet (Polynomial degree=2)	0.9163	1418.11301	3286.18929	0.92508757	0.93147991	0.9289 (+/- 0.03)
Ridg (Polynomial degree=2)	0.9156	1405.07355	3316.22355	0.92371199	0.93386043	0.9260 (+/- 0.03)
LR (Polynomial degree=2)	0.9156	1422.81265	3316.64098	0.92369278	0.93393518	0.9076 (+/- 0.05)
LASSO (Polynomial degree=1)	0.7717	3862.59891	5146.51752	0.81626342	0.82742769	0.8217 (+/- 0.02)
ElasticNet (Polynomial degree=1)	0.7723	3870.75378	5149.93451	0.81601936	0.82752003	0.8217 (+/- 0.02)
Ridg (Polynomial degree=1)	0.774	3909.13092	5166.05096	0.81486604	0.8277796	0.8211 (+/- 0.02)
LR (Polynomial degree=1)	0.7756	3916.18802	5168.88797	0.81466265	0.82779611	0.8210 (+/- 0.02)

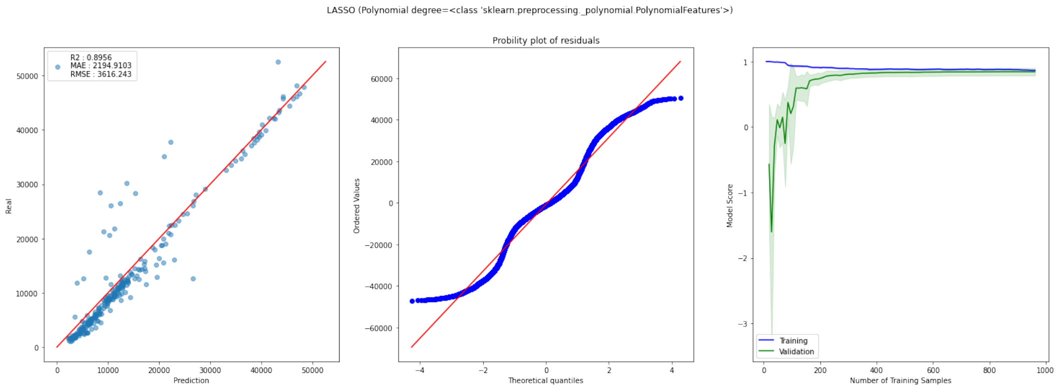
Machine Learning

Models

Model	Score (CV)	Score (CV Cook)	
LASSO (Polynomial degree=2)	0.8488	0.9289	+0.0801
ElasticNet (Polynomial degree=2)	0.8488	0.9289	+0.0801
Ridg (Polynomial degree=2)	0.8432	0.926	+0.0828
LR (Polynomial degree=2)	0.8395	0.9076	+0.068
LASSO (Polynomial degree=1)	0.7286	0.8217	+0.0931
ElasticNet (Polynomial degree=1)	0.7285	0.8217	+0.0932
Ridg (Polynomial degree=1)	0.7282	0.8211	+0.0929
LR (Polynomial degree=1)	0.7282	0.821	+0.0928

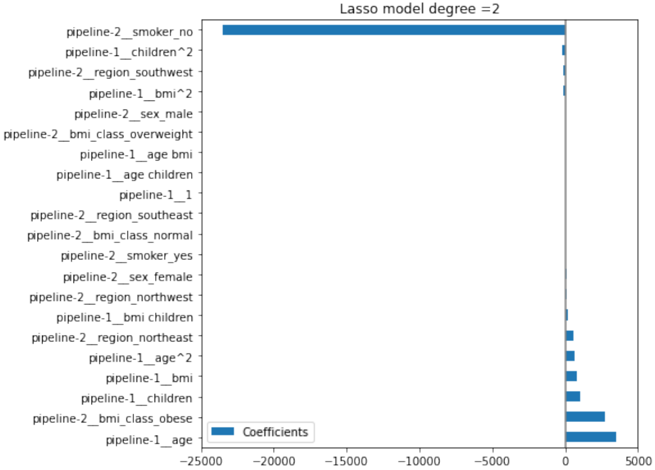
Machine Learning

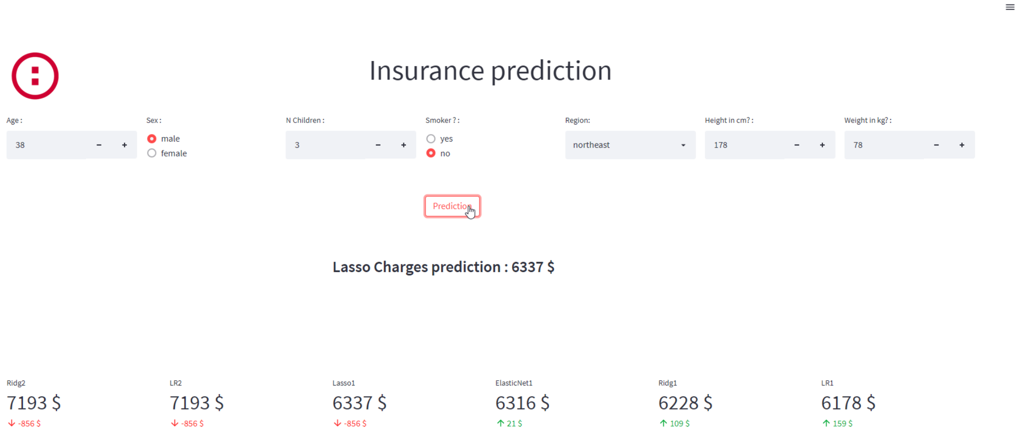
Results analysis



Machine Learning

Results analysis





Insurance prediction
Age :
38
Sex :
male
female
N Children :
3
Smoker ? :
yes
no
Region:
northeast
Height in cm? :
178
Weight in kg? :
78
Prediction
Lasso Charges prediction : 6337 $
Ridg2
7193 $
-856 $
LR2
7193 $
-856 $
Lasso1
6337 $
-856 $
ElasticNet1
6316 $
21 $
Ridg1
6228 $
109 $
LR1
6178 $
159 $

Table of Contents

4 Conclusion

- Smoking impact on health
- More observations
- More features (Alcool consumption, ...)

Assur'aimant
Insurance Charges Prediction

Thank you for listening!
Any questions?