DATA MINING – G6E

Modelo: Regressão Linear

Previsão: Número de gostos de publicações

Eduardo Brito – up201806271
Hugo Guimarães – up201806490
Paulo Ribeiro – up201806505
Pedro Ferreira – up201806506

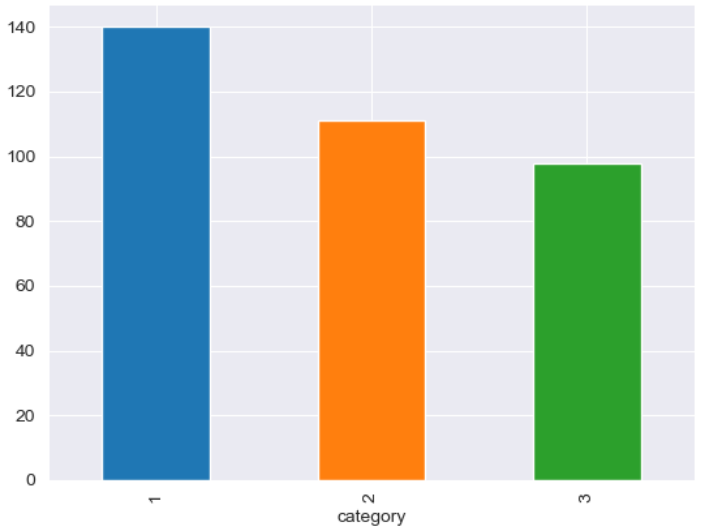
page_total_likes	type_of_post	category	month	weekday	hour	paid	no_likes	ID
139441	Photo	2	12	4	3	0	79	POST_1
139441	Status	2	12	3	10	0	130	POST_2
139441	Photo	3	12	3	3	0	66	POST_3
139441	Photo	2	12	2	10	1	1572	POST_4
139441	Photo	2	12	2	3	0	325	POST_5

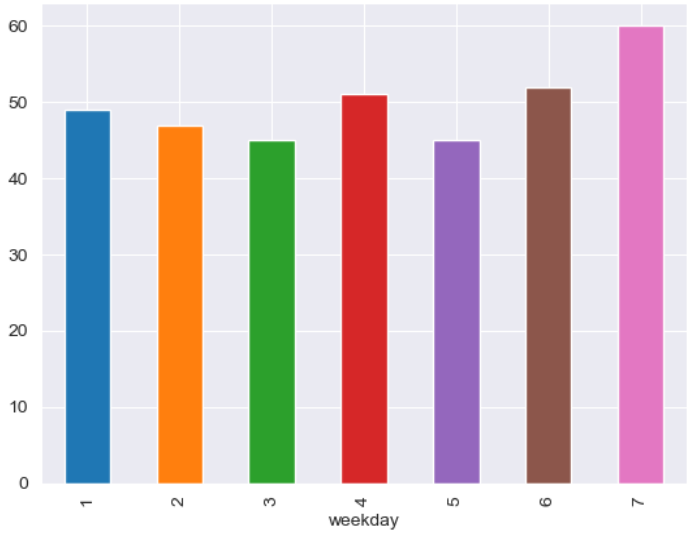
page_total_likes	type_of_post	category	month	weekday	hour	paid	no_likes	ID
116435	Photo	2	5	5	9	0	-1	POST_351
116435	Photo	3	5	4	13	1	-1	POST_352
116435	Status	1	5	4	3	1	-1	POST_353
116435	Photo	3	5	3	7	0	-1	POST_354
116435	Photo	2	5	2	14	0	-1	POST_355

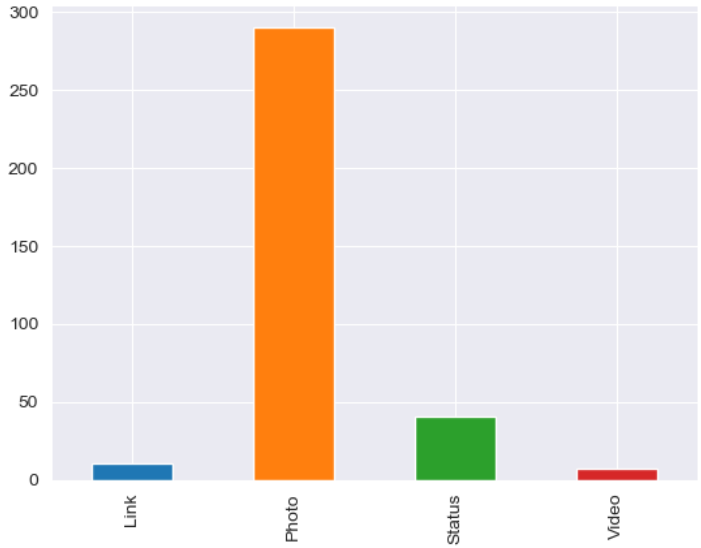
OVERVIEW DOS DADOS

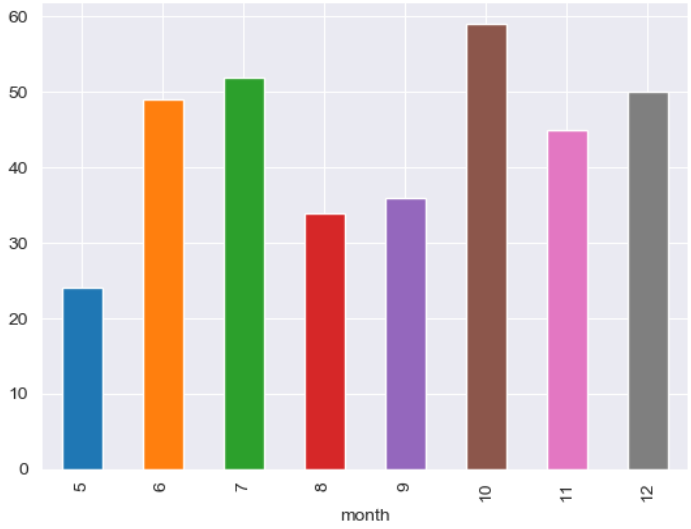
Objetivo: Prever o número de likes de posts desconhecidos

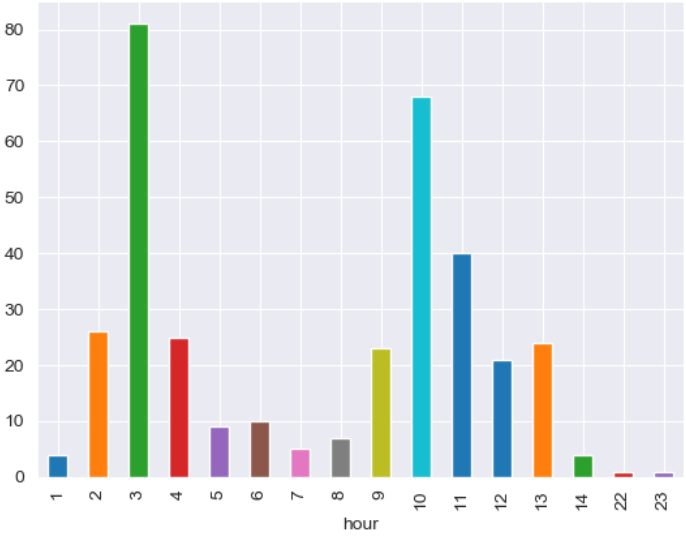
ALGUNS GRÁFICOS: FREQUÊNCIAS



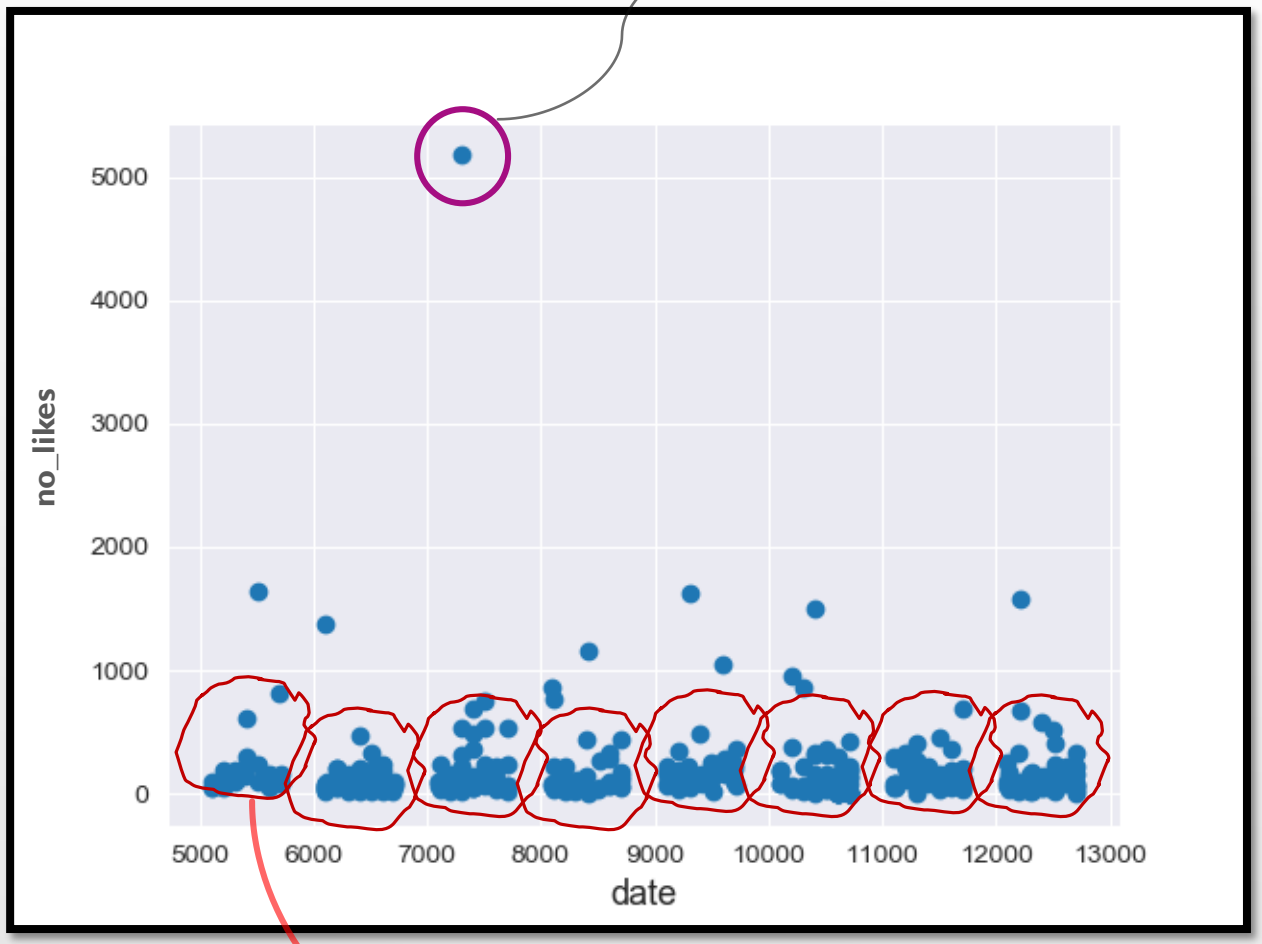








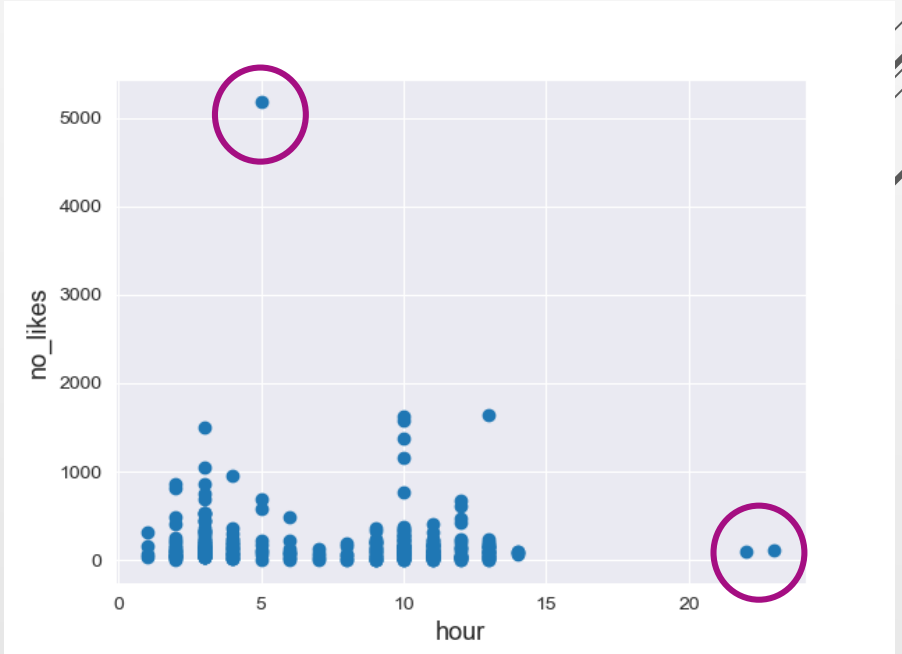
Possível Outlier



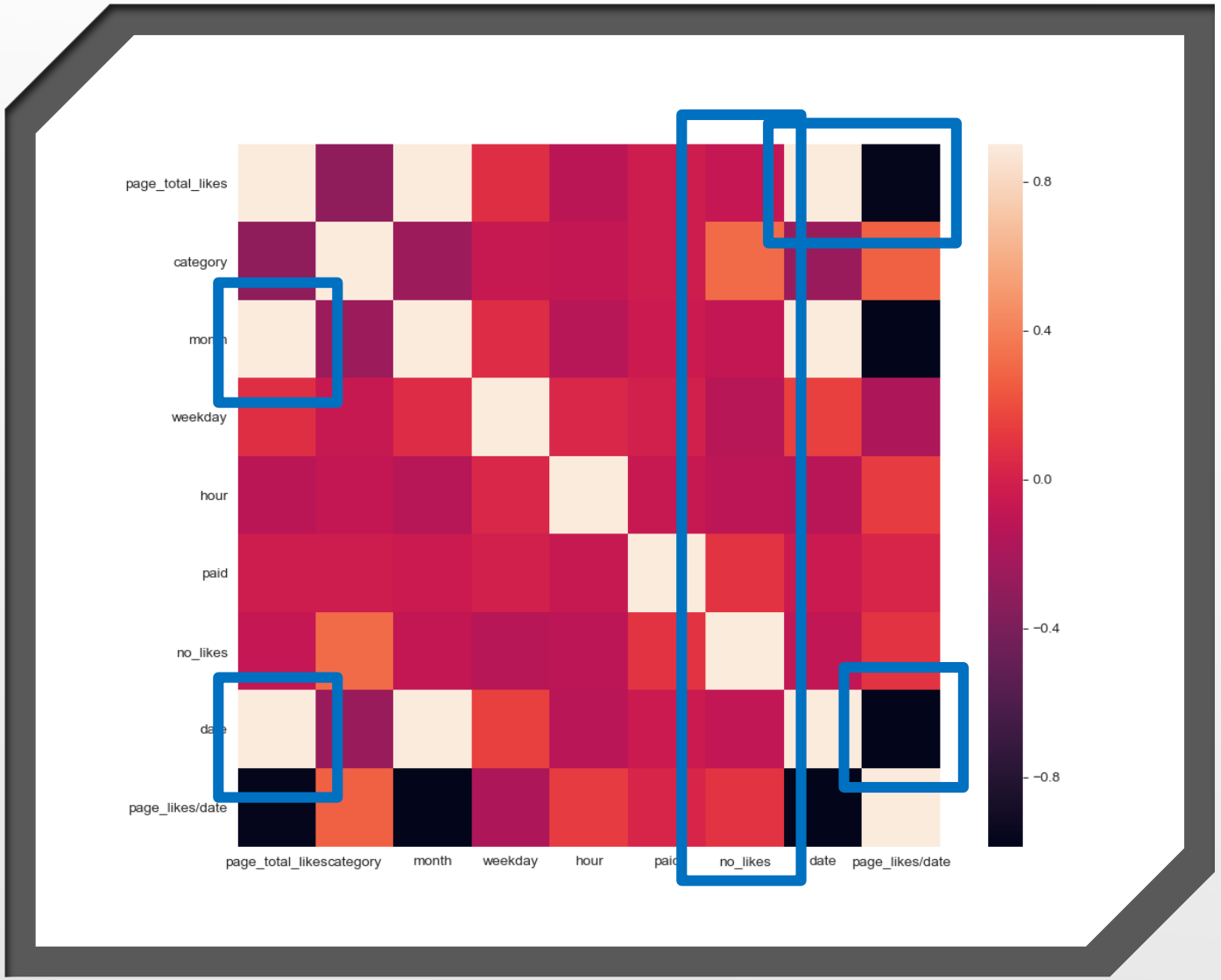
Dados identicamente concentrados

NOVA VARIÁVEL: DATA

```
train['date'] = 1000*train['month'] +
                100*train['weekday'] +
                train['hour']
```

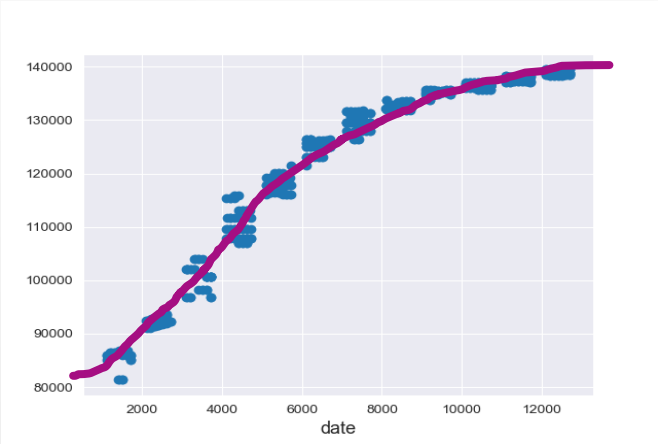


MAPA DE CORRELAÇÕES



Não há correlação visível entre no_likes e qualquer outra variável.

Só foi possível estabelecer uma relação notória entre date|month e page_total_likes:

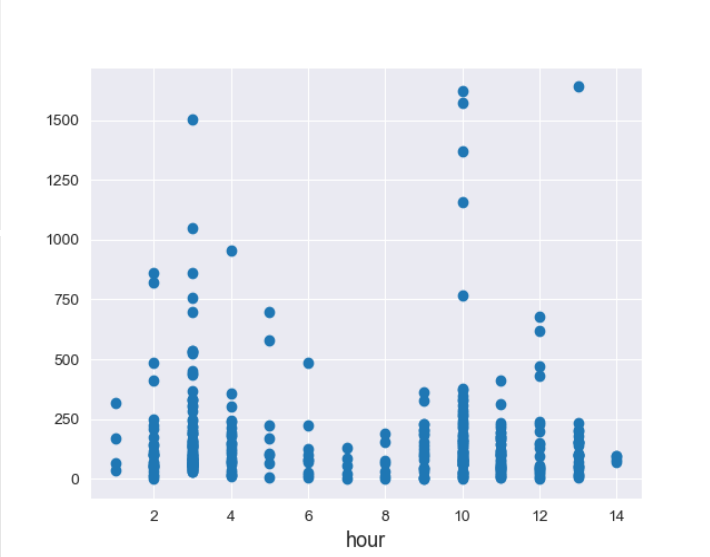


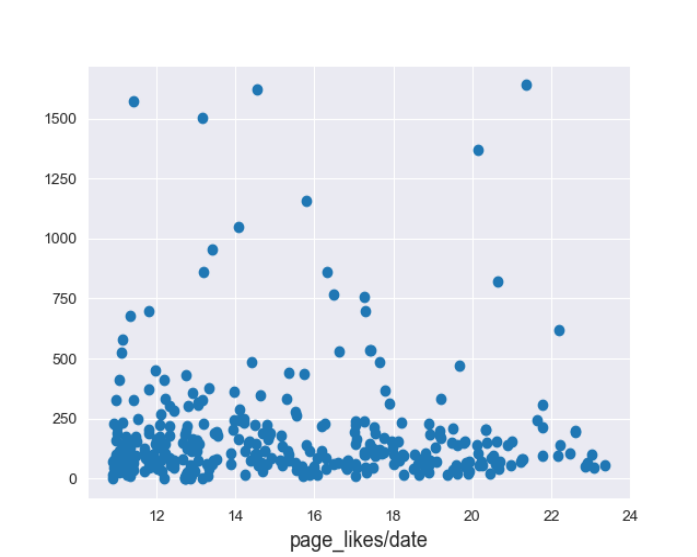
Attributes	type_of...	page_to...	category	month	weekday	hour	paid	no_likes
type_of_post	1	0.160	-0.174	0.193	0.023	-0.037	0.051	-0.028
page_total_likes	0.160	1	-0.301	0.943	0.069	-0.123	-0.032	-0.025
category	-0.174	-0.301	1	-0.256	-0.065	-0.080	-0.035	0.160
month	0.193	0.943	-0.256	1	0.066	-0.135	-0.048	-0.025
weekday	0.023	0.069	-0.065	0.066	1	0.036	-0.005	-0.056
hour	-0.037	-0.123	-0.080	-0.135	0.036	1	-0.067	-0.057
paid	0.051	-0.032	-0.035	-0.048	-0.005	-0.067	1	0.060
no_likes	-0.028	-0.025	0.160	-0.025	-0.056	-0.057	0.060	1

attribute	weight ↓
category	0.160
paid	0.060
hour	0.057
weekday	0.056
type_of_post	0.028
month	0.025
page_total_likes	0.025

Suspect outliers:

page_total_likes	type_of_post	category	month	weekday	hour	paid	no_likes
131630	Photo	3	7	2	23	0	113
130791	Photo	2	7	3	5	1	5172
126141	Photo	1	6	4	22	1	102

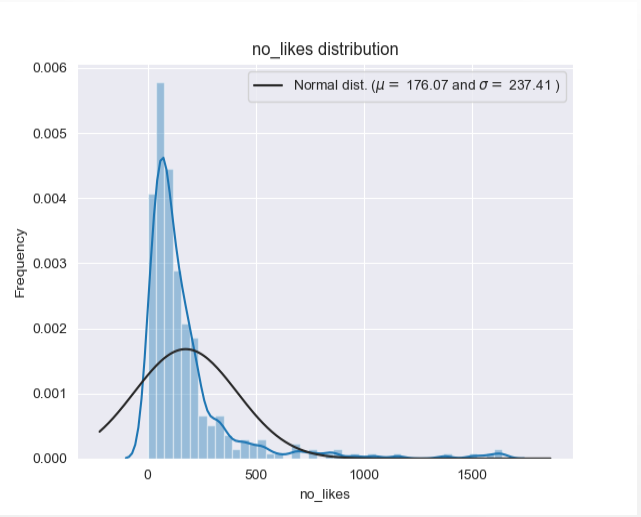


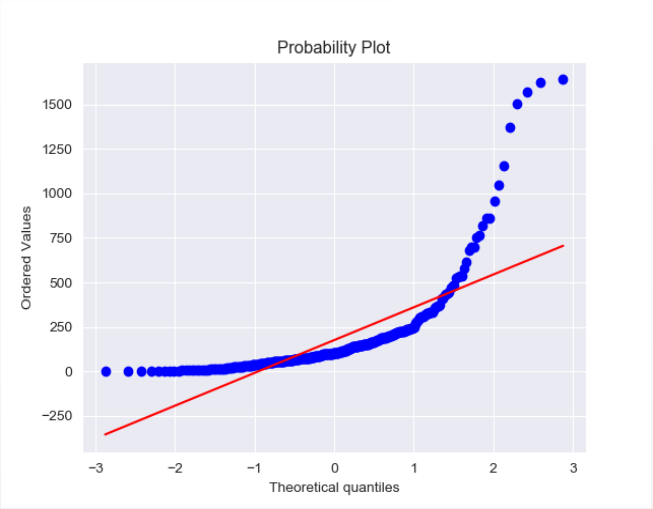


REMOVENDO OUTLIERS

```
#Visualizing 'outlier' suspects
print("Suspect outliers:\n", train[(train['hour']>20)
                                 | (train['no_likes']>3000)
                                 | (train['month']<5)] )

#Deleting outliers
train = train.drop(train[(train['hour']>20)
                         | (train['no_likes']>3000)
                         | (train['month']<5)].index)
```





INTERVALO DE CONFIANÇA A 80% :

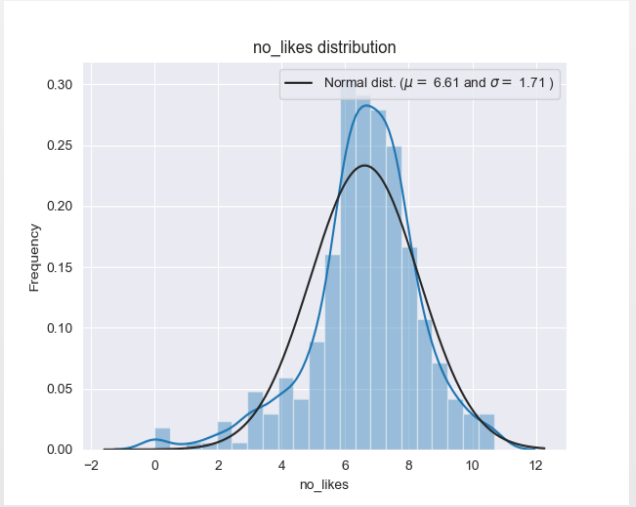
80% Confidence Interval:
175.67528735632183 ± 16.3
(159 to 192)

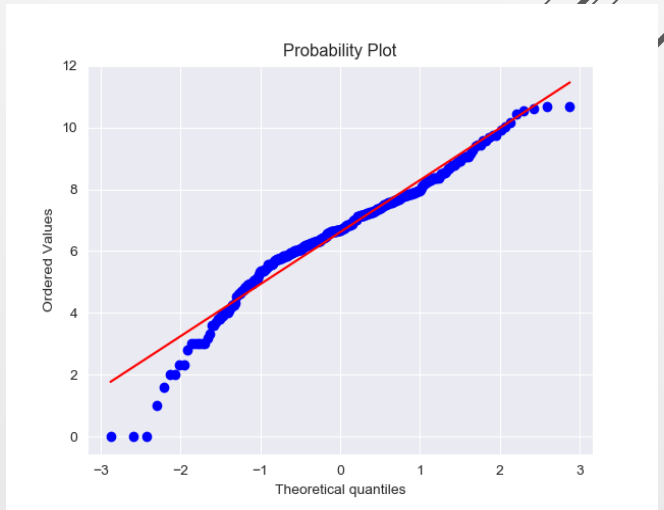
DISTRIBUIÇÃO DA VARIÁVEL: NO_LIKES

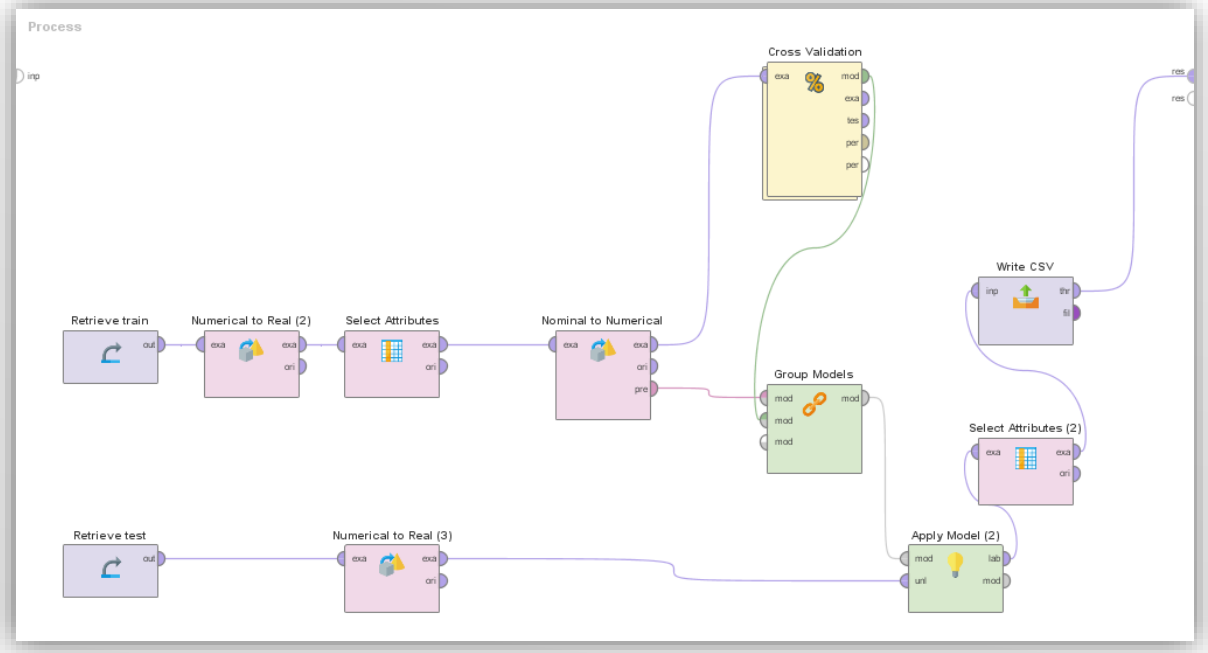
Tentativa de normalização com a aplicação da transformação:

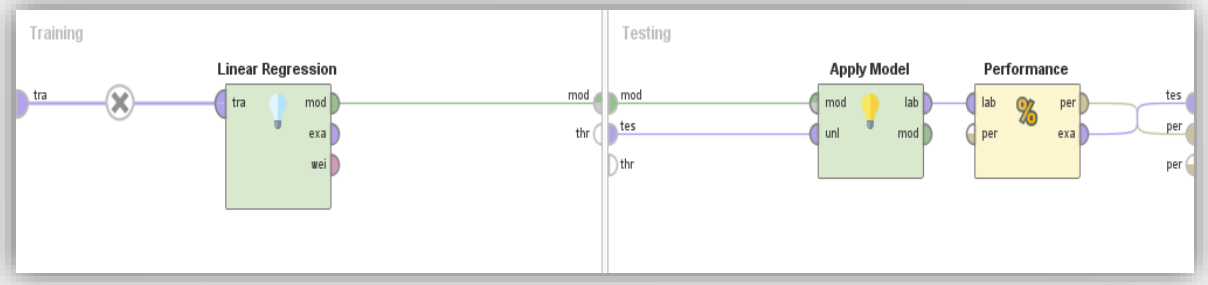
train["no_likes"] = np.**log2(train["no_likes"]+1)**

Por forma a conseguir aproximar de uma distribuição normal.









PARTES DO ALGORITMO (RAPIDMINER E PYTHON)

python code:

```
from sklearn.preprocessing import LabelEncoder
cols = ("type_of_post","paid")
# process columns, apply LabelEncoder to categorical features
for c in cols:
    lbl = LabelEncoder()
    lbl.fit(list(all_data[c].values))
    all_data[c] = lbl.transform(list(all_data[c].values))

train = all_data[:ntrain]
test = all_data[ntrain:]

from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

def rmsle(y, y_pred):
        return np.sqrt(mean_squared_error(y, y_pred))

model = LinearRegression().fit(train.values, y_train)
stacked_train_pred = model.predict(train.values)
stacked_pred = np.exp2(model.predict(test.values))-1
print("Coef: ",model.coef_)
print("RMSE(Train): ", rmsle(stacked_train_pred,y_train ))

train['no_likes'] = np.exp2(y_train)-1
train['no_likes_pred'] = np.exp2(stacked_train_pred)-1
print(train.head(20))
```

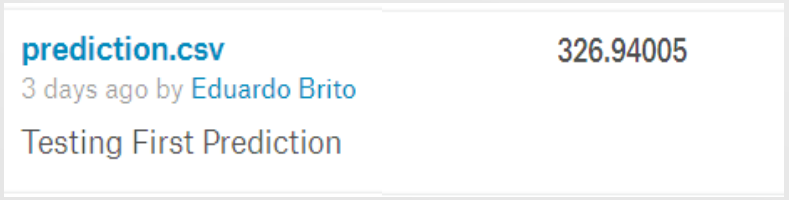
MAU RESULTADO

Row No.	ID	prediction(n...
74	POST_424	-67.740
75	POST_425	-67.740
76	POST_426	25.430
77	POST_427	-131.473
78	POST_428	-126.207
79	POST_429	-115.675
80	POST_430	-115.675
81	POST_431	-194.461
82	POST_432	-99.877
83	POST_433	-118.123

Resultados negativos

Vários posts com resultados extremos

Previsões irrealistas



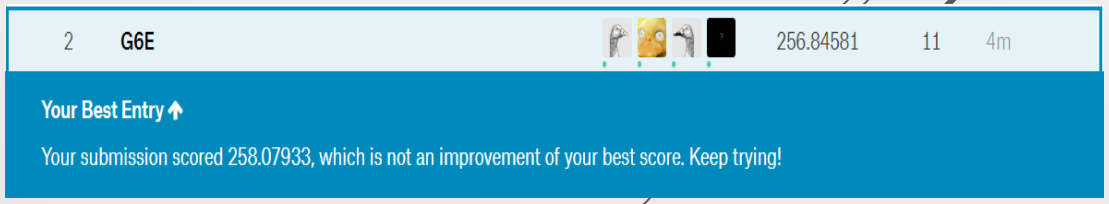
BOM RESULTADO

Row No.	ID	no_likes
33	POST_383	120.086
34	POST_384	190.001
35	POST_385	125.311
36	POST_386	295.529
37	POST_387	130.790
38	POST_388	201.017
39	POST_389	136.537
40	POST_390	142.564
41	POST_391	388.930

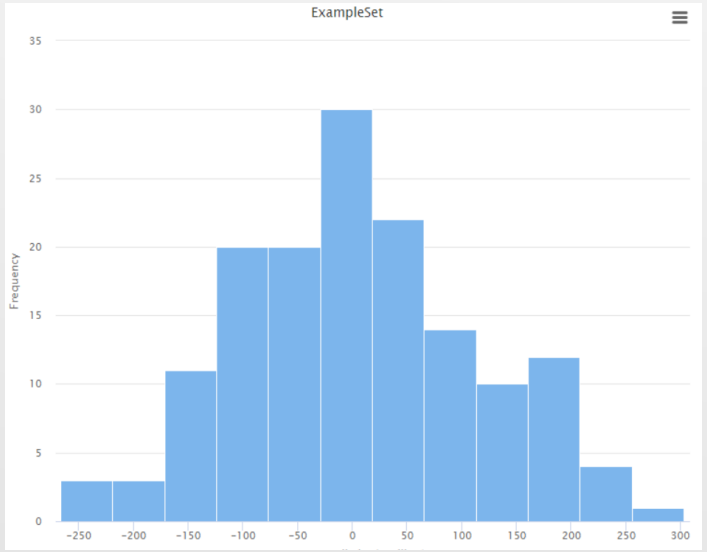
Resultados mais normalizados

Média mais próxima da prevista

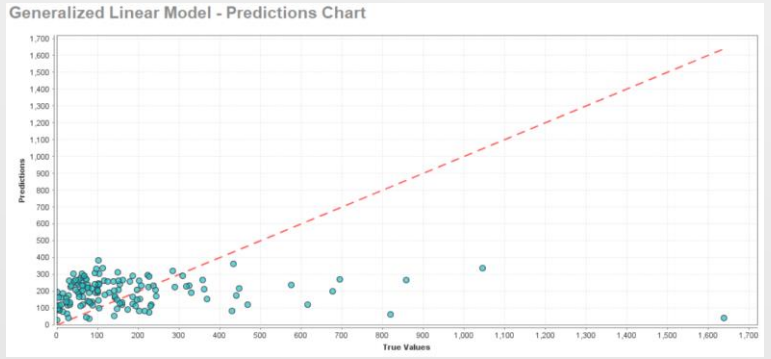
Previsões menos distantes



Através da medida de erro usada, consegue-se facilmente observar uma enorme diferença entre as previsões e o valor real dos dados, o que pode significar o falhanço total da aplicação de um modelo linear na previsão destes resultados. Isto leva a crer que não existe uma relação linear notória que permita descrever a evolução da variável em estudo, tornando-se praticamente aleatória a sua existência a par com as outras variáveis recolhidas.



$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Predicted_i - True_i)^2}$$



CONCLUSÃO

Eduardo Brito – up201806271
Hugo Guimarães – up201806490
Paulo Ribeiro – up201806505
Pedro Ferreira – up201806506