



CHICAGO BICYCLE

ÉRAMOS QUATRO

O problema proposto é o de Compartilhamento de Bicicletas. Neste conjunto, temos uma série de atributos e, como target, devemos prever a quantidade de bicicletas que serão alugadas em um determinado período.

QUEM SOMOS

GRUPO: ÉRAMOS QUATRO

Felipe Pereira - Engenheiro de Dados
Alexandre Guidin - Analista de Sistemas
Anderson Rocha - Analista de Sistemas



ANÁLISE DE DADOS

O dataset utilizado passou por 4 fases:

- **Raw Data (fornecido)**
- **Clean Data (fornecido)**
- **Transformation**
- **Prediction Data**

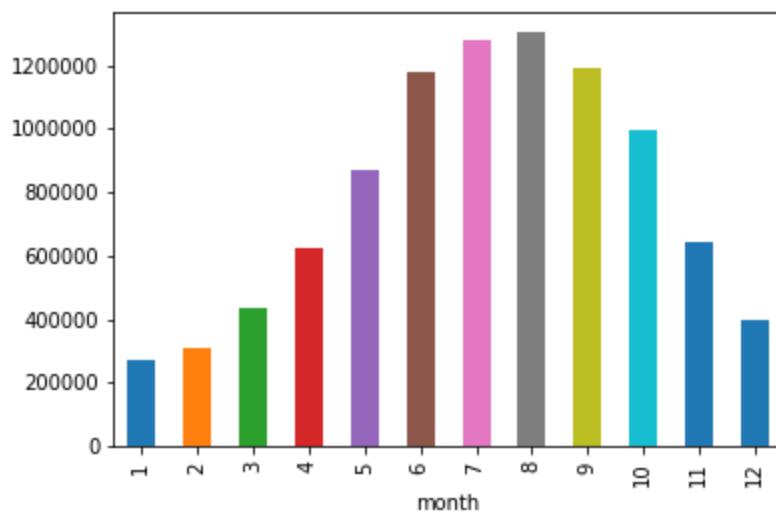
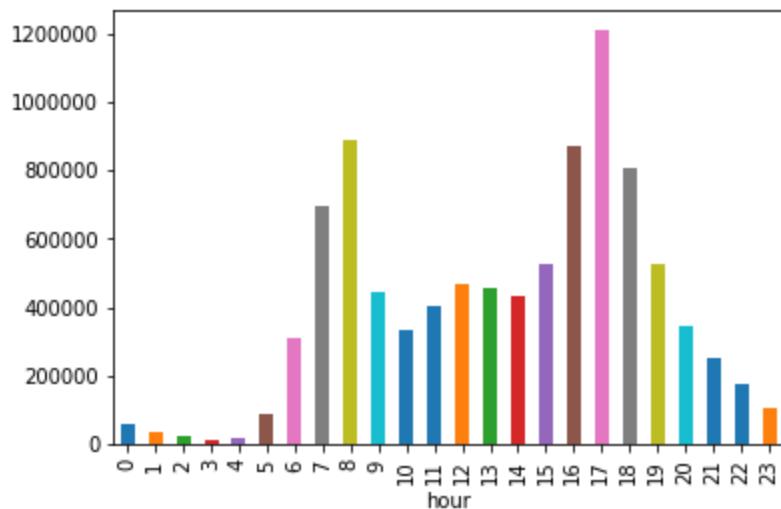
ANÁLISE DE DADOS

Correlação

median_temperature	tstorms	unknown	cloudy	rain_or_snow	not_clear	clear	rentals	
0.076302	0.003549	-0.003490	0.023567	-0.012664	-0.032328	-0.009926	0.168986	year
0.257745	-0.001537	-0.003441	0.041586	-0.061673	0.002691	0.004487	0.147929	month
0.255064	-0.001183	-0.002941	0.036959	-0.058947	0.000403	0.008848	0.144008	week
-0.016198	-0.012698	0.001871	-0.001572	-0.005611	-0.004731	0.013531	-0.134087	day
0.084455	0.017825	-0.001143	0.047427	0.003428	-0.029590	-0.065340	0.252474	hour
-0.009158	-0.010013	-0.004803	0.006108	0.006926	-0.004740	-0.010734	-0.002065	is_weekend
1.000000	0.075140	0.003980	0.174433	-0.152806	-0.019605	-0.105052	0.475935	mean_temperature
1.000000	0.075140	0.003980	0.174433	-0.152806	-0.019605	-0.105052	0.475935	median_temperature
0.075140	1.000000	-0.000701	-0.179072	-0.030399	-0.009348	-0.029537	-0.015352	tstorms
0.003980	-0.000701	1.000000	-0.014758	-0.002505	-0.000770	-0.002434	-0.000146	unknown
0.174433	-0.179072	-0.014758	1.000000	-0.639909	-0.196783	-0.621764	0.193332	cloudy
-0.152806	-0.030399	-0.002505	-0.639909	1.000000	-0.033406	-0.105551	-0.147542	rain_or_snow
-0.019605	-0.009348	-0.000770	-0.196783	-0.033406	1.000000	-0.032459	-0.007363	not_clear
-0.105052	-0.029537	-0.002434	-0.621764	-0.105551	-0.032459	1.000000	-0.112553	clear
0.475935	-0.015352	-0.000146	0.193332	-0.147542	-0.007363	-0.112553	1.000000	rentals

ANÁLISE DE DADOS

Hora e Mês



TRANSFORMAÇÃO

- Dimensionalidade:
 - 9.495.235 registros
 - 23 atributos
- Qualidade dos dados:
 - Nenhum valor nulo
 - Nenhum valor corrompido
- Agrupamento:
 - Os dados foram agrupados para se tornarem predizíveis
- Atributos categóricos:
 - Foram transformados em atributos discretos
- Redução de dimensionalidade:
 - Atributos não utilizados no modelo foram retirados
- Normalização:
 - Os dados foram normalizados para o input nos modelos

PREPARAÇÃO DO DATASET

```
#Modelos a serem testados
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
from sklearn.linear_model import LinearRegression, BayesianRidge, LogisticRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import AdaBoostRegressor, RandomForestRegressor

#Retirada da variável target das features de predição
X = data.drop('rentals',1)
y = data.rentals

#Separação de conjunto de testes
X_model, X_test, y_model, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

#Separação de conjunto de validação
X_train, X_val, y_train, y_val = train_test_split(X_model, y_model, test_size=0.2, random_state=0)
```

PREPARAÇÃO DOS MODELOS

```
#Treinamento de modelos
lr = LinearRegression(n_jobs=5, fit_intercept=True)
logr = LogisticRegression(penalty='l2', C=1.0, fit_intercept=True, random_state=0, solver='liblinear', n_jobs=5)
dt = DecisionTreeRegressor(max_depth=10, criterion='mse', splitter='best', random_state=0, presort=True)
dtr = AdaBoostRegressor(dt, n_estimators=500, learning_rate=0.1, random_state=0)
rf = RandomForestRegressor(n_estimators=100, criterion='mse', max_features='auto', random_state=0, n_jobs=5)
blr = BayesianRidge(n_iter=1000, fit_intercept=True)

#Criação de vetor de modelos
algs = []
algs.append(lr)
algs.append(logr)
algs.append(dt)
algs.append(dtr)
algs.append(rf)
algs.append(blr)

#Fit dos modelos
for alg in algs:
    print('Fitting: ', type(alg).__name__)
    alg.fit(X_model, y_model)
```

RESULTADOS

	Name	Type	R2	MAE	MSE
0	LinearRegression	Train	0.323172	187.213935	72950.151816
1	LogisticRegression	Train	-0.083556	200.826578	116788.336282
2	DecisionTreeRegressor	Train	0.893314	57.073984	11498.862146
3	AdaBoostRegressor	Train	0.969952	40.341926	3238.639710
4	RandomForestRegressor	Train	0.991510	14.508301	915.028547
5	BayesianRidge	Train	0.323140	187.200312	72953.635601
6	LinearRegression	Validation	0.320537	186.327603	71867.391533
7	LogisticRegression	Validation	-0.081194	201.124391	114358.864777
8	DecisionTreeRegressor	Validation	0.881600	59.083956	12523.314164
9	AdaBoostRegressor	Validation	0.968995	40.807981	3279.447613
10	RandomForestRegressor	Validation	0.991866	14.580894	860.304890
11	BayesianRidge	Validation	0.320621	186.288606	71858.594126
12	LinearRegression	Test	0.312564	189.855222	74930.489941
13	LogisticRegression	Test	-0.094791	204.497545	119332.151213
14	DecisionTreeRegressor	Test	0.874726	63.518099	13654.849082
15	AdaBoostRegressor	Test	0.930690	50.334499	7554.829703
16	RandomForestRegressor	Test	0.939110	39.605608	6637.032840
17	BayesianRidge	Test	0.312393	189.854716	74949.132372

CONSIDERAÇÕES FINAIS





THANKS!