Discente: Deivison Venicio Souza

Laboratório Aprendizagem de Máquina

Lab7: Regressão

Considere a base em anexo com os dados de uma usina hidroelétrica, que tem as seguintes características:

f3) Vazão Turbinada

f4) Vazão Vertida

f5) Afluência

f6) Afluência Soma Montante

f7) Volume Prévio

f8) Volume Atual

f9) Volume Prévio Soma Montante

f10) Volume Atual Soma Montante

f11) Volume Prévio Soma Jusante

f12) Volume Atual Soma Jusante

As características [**f5..f12**] devem ser utilizadas para fazer a predição das variáveis **f3** e **f4**. Ou seja, você deve treinar dois regressores, um para cada tarefa. Note que nem todas as características precisam ser utilizadas pelo regressor. Dentre os regressores abaixo, reporte aquele que minimiza o erro quadrado médio (MSE).

• Regressão linear

• SVR

• Redes Neurais (MLP) (+ de uma camada escondida)

• kNN

• Árvore de Decisão

• Random Forest

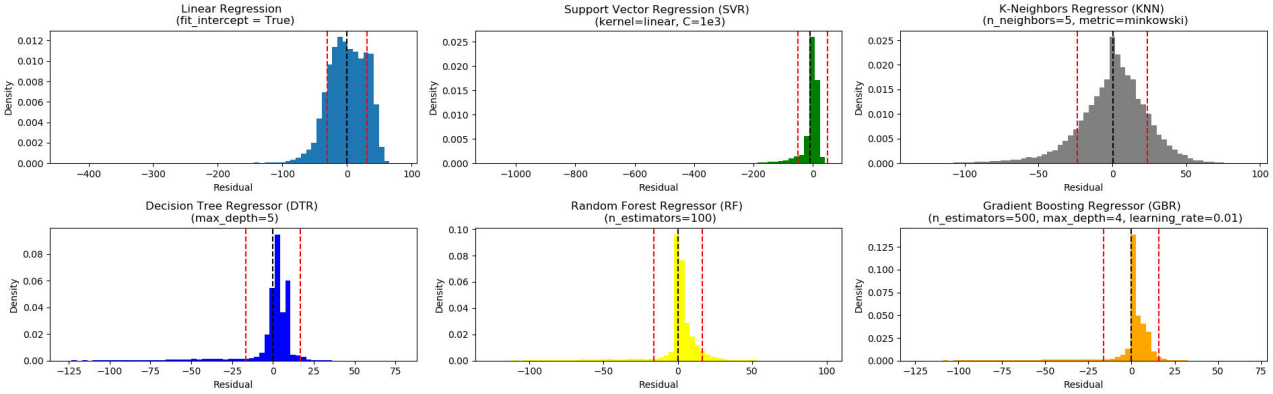
• Gradient Boosting

Divida a base em dois subconjuntos (50-50). Utilize o parâmetro **random_state** na função **train_test_split** para poder comparar os resultados. Para encontrar parâmetros dos regressores, você pode utilizar qualquer método de validação. 50% da base deve ser utilizado exclusivamente para testar o modelo. Faça uma análise de resíduos e verifique se 95% dos resíduos estão no intervalo de um desvio padrão.

---

1) Previsão da Vazão Turbinada (f3)

Foram ajustados vários algoritmos para prever a variável resposta (vazão turbinada = “f3”). Os histogramas das distribuições residuais resultante da aplicação dos modelos ajustados sobre o conjunto de teste estão disponível, bem como os valores de MSE, Erro Médio e Desvio Padrão residual. O modelo ajustado usando o algoritmo “Gradient Boosting Regressor”, com as configurações (n_estimators=500, max_depth=4, learning_rate=0.01) foi o que apresentou melhor desempenho (MSE = 256,7481) para estimar a variável “vazão turbinada”. As linhas pontilhadas sobre os histogramas indicam a média (cor preta) e o desvio padrão (cor vermelha) residual (1 desvio padrão da média). Assim, para o modelo “Gradient Boosting Regressor” percebe-se que a maior parte dos resíduos está dentro do intervalo de 1 desvio padrão da média residual. Por outro lado, os modelos SVR e Linear Regression tiveram pior desempenho (MSE = 2598, 6244 e MSE = 970, 8887, respectivamente), onde uma grande quantidade de scores residuais estiveram acima de 1 desvio padrão da média.

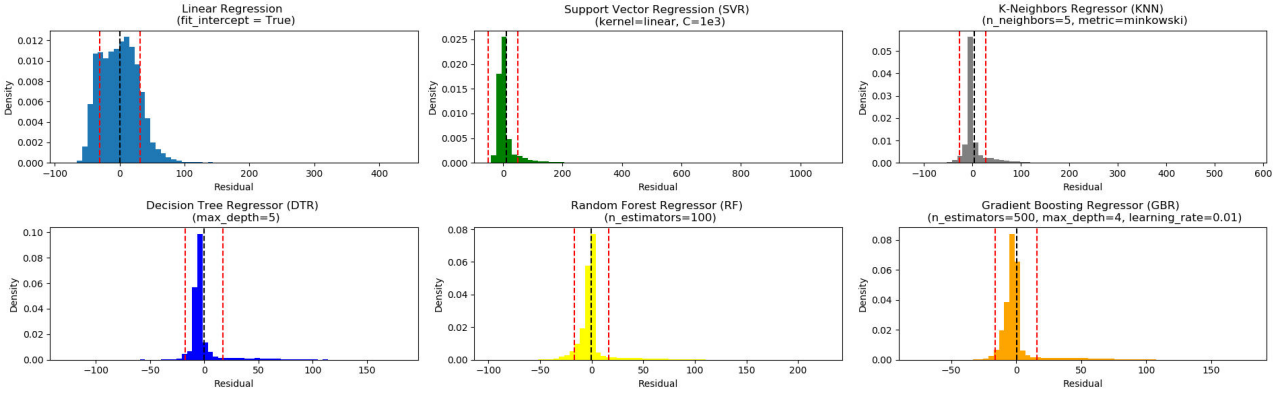


Modelos	MSE	Erro Médio	Desvio
Linear Regression	970.8887	-0.17946	31.1585
Support Vector Regression	2598.6244	-10.1936	49.9471
K-Neighbors Regressor	565.6713	0.60215	23.7762
Decision Tree Regressor	277.93600	-0.21795	16.6699
Random Forest Regressor	264.9658	0.28927	16.2752
Gradient Boosting Regressor	256.7481	-0.2359	16.0216

---

2) Previsão da Vazão Vertida (f4)

Foram ajustados vários algoritmos para prever a variável resposta (vazão vertida = “f4”). Os histogramas das distribuições residuais resultante da aplicação dos modelos ajustados sobre o conjunto de teste estão disponível, bem como os valores de MSE, Erro Médio e Desvio Padrão residual. De modo similar ao verificado para a variável “vazão turbinada”, o modelo ajustado usando o algoritmo “Gradient Boosting Regressor”, com as configurações (n_estimators=500, max_depth=4, learning_rate=0.01) foi o que apresentou melhor desempenho (MSE = 262,1211) para estimar a variável “vazão vertida”, seguido do “Random Forest”. O “Gradient Boosting Regressor” teve maior parte dos resíduos dentro do intervalo de 1 desvio padrão da média residual. Novamente, os modelos de pior desempenho foram o SVR e Linear Regression.



Modelos	MSE	Erro Médio	Desvio
Linear Regression	970.8889	0.17946	31.1585
Support Vector Regression	2611.6392	10.2395	50.0678
K-Neighbors Regressor	767.1215	4.3131	27.3591
Decision Tree Regressor	298.1125	0.1966	17.2648
Random Forest Regressor	272.7032	-0.3275	16.5105
Gradient Boosting Regressor	262.1211	0.2248	16.1885