MO432 - Trabalho 4

Jônatas Trabuco Belotti
RA: 230260
jonatas.t.belotti@hotmail.com

I. Introdução

Esse trabalho tem como objetivo realizar a previsão de séries temporais utilizando técnicas de Aprendizado de Máquina. A série temporal em questão é uma série diária com o histórico da Taxa de Cambio entre o Dólar e o Euro de 04/01/1999 até 24/07/2020. Um total de 5581 registros compostos pela data e o valor da taxa no dia. O arquivo completo que foi utilizado como entrada pode acessado através do link `https://github.com/jonatastbelotti/mo432-trab4/blob/master/dados4.csv`.

Foram considerados 2 problemas distintos com a série de Taxa de Cambio. O primeiro é um problema de previsão, onde a partir dos valores anteriores da série se deseja prever qual será o próximo valor da Taxa de Cambio. Já o segundo é um problema de classificação, onde, dados os valores anteriores da série é necessário classificar se no próximo dia o valor da Taxa de Cambio irá subir ou não (note que aqui não vale ser igual, o valor deve ser estritamente maior).

Todos os códigos deste trabalho foram escritos utilizando a linguagem de programação **Python 3.6.9** com as seguintes bibliotecas: **Numpy 1.18.2**, **Pandas 1.1.0**, **Scikit-Learn 0.22.2** e **Matplotlib 3.2.2**. Sendo estas obrigatórias para execução dos códigos. Todo o código desenvolvido juntamente com os arquivos utilizados podem ser acessados no repositório do GitHub `https://github.com/jonatastbelotti/mo432-trab4`.

II. Metodologia

Como já foi mencionado a série temporal possui um total de 5581 registros, sendo cada registro composto por uma data e o valor da taxa. A Figura 1 apresenta o gráfico com a série completa.

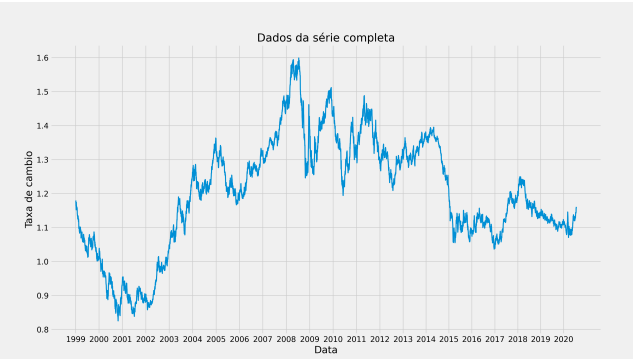

Fig. 1: Série completa da Taxa de Cambio.

Como métricas de desempenho foram utilizadas a Acurácia para o problema de classificação e o RMSE para o problema de previsão. Sendo a Acurácia quanto maior melhor e o RMSE quanto menor melhor.

Inicialmente os registros foram ordenadas do mais antigo para o mais novo, logo em seguida a coluna da data foi removida. Posteriormente os registros foram separados em 2 conjuntos, o primeiro, contendo 90% dos registros é o conjunto de treinamento, que foi utilizado para treinar todos os modelos e para escolher todos os hiperparâmetros. O segundo, é o conjunto de medida, com 10% dos registros, tem o objetivo de realizar as medições finais das métricas de desempenho. Essa separação em treinamento e medida é importante para garantir que as medidas de desempenho não sejam coletadas em registros que foram utilizados na etapa de aprendizado.

A primeira variável a ser testada foi o número de entradas dos modelos, ou seja, quantos valores do passado serão utilizados para prever o próximo valor. Aqui foram testados todos os valores de 1 até 15, no final sendo escolhido o valor com o modelo com a melhor medida de desempenho.

Para cada número de entradas foram aplicadas 2 etapas de Pré-processamento, primeiramente todos os atributos de entrada foram padronizados através das técnicas de *Centering* e *Standard Scaling*. Posteriormente foi aplicado o *PCA* para reduzir a dimensionalidade dos dados até atingir uma variância de 90%.

A escolha dos hiperparâmetros de cada modelo foi realizada através de uma validação cruzada com 5 repetições de um split de treino e teste. Sendo escolhidos os hiperparâmetros com melhor média de desempenho no conjunto de teste.

III. Resultados

Nessa Seção são apresentados os resultados obtidos por cada modelo testado tanto para o problema de previsão (Seção III-B), quanto para o problema de classificação (Seção III-A). Também são apresentados os melhores valores para as métricas de desempenho para o número de entradas escolhido.

A. Classificação

A Tabela I contém a acurácia média obtida para o conjunto de teste e a acurácia para o conjunto de medida para o melhor modelo de classificação em cada número de entradas testado.

Como pode ser observado pelos dados da Tabela I o melhor número de entradas para classificação é 13.

TABELA I: Resultados obtidos para cada número de entradas na classificação.

Entradas	Acurácia (Teste)	Acurácia (Medida)
1	$50,17\%$	$52,32\%$
2	$51,43\%$	$52,15\%$
3	$50,09\%$	$52,32\%$
4	$50,07\%$	$52,15\%$
5	$49,76\%$	$52,32\%$
6	$50,62\%$	$52,15\%$
7	$50,19\%$	$52,50\%$
8	$51,16\%$	$52,15\%$
9	$50,20\%$	$52,15\%$
10	$50,05\%$	$52,32\%$
11	$50,39\%$	$52,50\%$
12	$50,08\%$	$52,68\%$
13	$50,27\%$	$54,83\%$
14	$50,11\%$	$54,65\%$
15	$49,88\%$	$53,22\%$

Os hiperparâmetros de cada modelo utilizado no problema de classificação foram definidos através de uma validação cruzada em cada um dos números de entradas testados. Os modelos finais com os respectivos valores de hiperparâmetros utilizados para a classificação com 13 entradas foram:

- **RL -** Regressão Logística (sem regularização);
- **RL-L2 -** Regressão Logística com Regularização L2, $C = 0.05762721573666136$;
- **LDA -** Análise discriminante;
- **QDA -** Análise Quadrática;
- **SVM-L -** SVM Linear, $C = 2273.302682434829$;
- **SVM-RBF -** SVM com kernel RBF, $C = 6325.235610964958$, $\text{gamma} = 7.939293018902375$;
- **NB -** Naive Bayes;
- **KNN -** K-ésimo Vizinho mais Próximo, $\text{n\_neighbors} = 244$;
- **MLP -** Perceptron multicamadas, $\text{hidden\_layer\_sizes} = 20$;
- **AV -** Arvore de decisão, $\text{ccp\_alpha} = 0.002427644773261082$;
- **GBM -** Gradient boosting, $\text{learning\_rate} = 0.2715551307766882$, $\text{max\_depth} = 3$, $\text{n\_estimators} = 55$.

A acurácia final obtida para cada modelo classificar utilizando 13 entradas pode ser observada na Tabela II.

Observando os dados da Tabela II nota-se que o melhor classificador foi a SVM com kernel RBF, tendo obtido uma acurácia de $54,83\%$ para o conjunto de medida.

B. Previsão

Na Tabela III estão os melhores resultados obtidos no conjunto de medida para cada número de entradas testado e a média do RMSE para o conjunto de teste.

Os hiperparâmetros de cada modelo utilizado no problema de previsão foram definidos através de uma validação cruzada em cada um dos números de entradas testados. Os modelos finais com os respectivos valores de hiperparâmetros utilizados para a previsão com 1 valor de entrada foram:

TABELA II: Resultados finais de classificação com 13 entradas.

Modelo	Acurácia (Teste)	Acurácia (Medida)
RL	$49,24\%$	$45,34\%$
RL-L2	$49,29\%$	$45,34\%$
LDA	$49,24\%$	$45,34\%$
QDA	$49,65\%$	$47,84\%$
SVM-L	$50,99\%$	$52,15\%$
SVM-RBF	$50,27\%$	$54,83\%$
NB	$49,67\%$	$47,84\%$
KNN	$50,13\%$	$52,68\%$
MLP	$50,01\%$	$51,79\%$
AV	$48,98\%$	$47,84\%$
GBM	$49,34\%$	$50,35\%$

TABELA III: Resultados obtidos para cada número de entradas na previsão.

Entradas	RMSE (Teste)	RMSE (Medida)
1	$0,007529$	$0,004772$
2	$0,008433$	$0,005317$
3	$0,009420$	$0,005958$
4	$0,010362$	$0,006593$
5	$0,011263$	$0,007159$
6	$0,012118$	$0,007675$
7	$0,012919$	$0,008123$
8	$0,013717$	$0,008522$
9	$0,014487$	$0,008858$
10	$0,015229$	$0,009131$
11	$0,015939$	$0,009352$
12	$0,016626$	$0,009538$
13	$0,017296$	$0,009708$
14	$0,017951$	$0,009868$
15	$0,018581$	$0,010014$

- **L-L2 -** Linear com regularização L2, $\text{alpha} = 0.0010288884757614191$;
- **L-L1 -** Linear com regularização L1, $\text{alpha} = 0.003037682615559747$;
- **SVM-L -** SVM Linear, $C = 25.430278358788314$, $\text{epsilon} = 0.1$;
- **SVM-RBF -** SVM com kernel RBF, $C = 2379.404519551488$, $\text{epsilon} = 0.1$, $\text{gamma} = 0.0037958557437025825$;
- **KNN -** K-ésimo Vizinho mais Próximo, $\text{n\_neighbors} = 150$;
- **MLP -** Perceptron multicamadas, $\text{hidden\_layer\_sizes} = 195$, $\text{max\_iter} = 400$;
- **AD -** Árvore de decisão, $\text{ccp\_alpha} = 0.0017212507559494306$;
- **GBM -** Gradient boosting, $\text{learning\_rate} = 0.2326361299006572$, $\text{max\_depth} = 4$, $\text{n\_estimators} = 96$.

O RMSE final obtido por cada modelo de previsão com 1 entrada pode ser visto na Tabela IV.

Como pode ser visto nos dados da Tabela IV o melhor modelo preditivo foi a Regressão Linear com regularização L2. A seguir, a Figura 2 contém as previsões realizadas por esse modelo para o conjunto de medida.

TABELA IV: Resultados finais de previsão com 1 entrada.

Modelo	RMSE (Teste)	RMSE (Medida)
L-L2	$0,007529$	$0,004772$
L-L1	$0,011010$	$0,004995$
SVM-L	$0,072532$	$0,053193$
SVM-RBF	$0,057130$	$0,023531$
KNN	$0,052739$	$0,005134$
MLP	$0,016881$	$0,006303$
AD	$0,114759$	$0,041295$
GBM	$0,035093$	$0,005231$

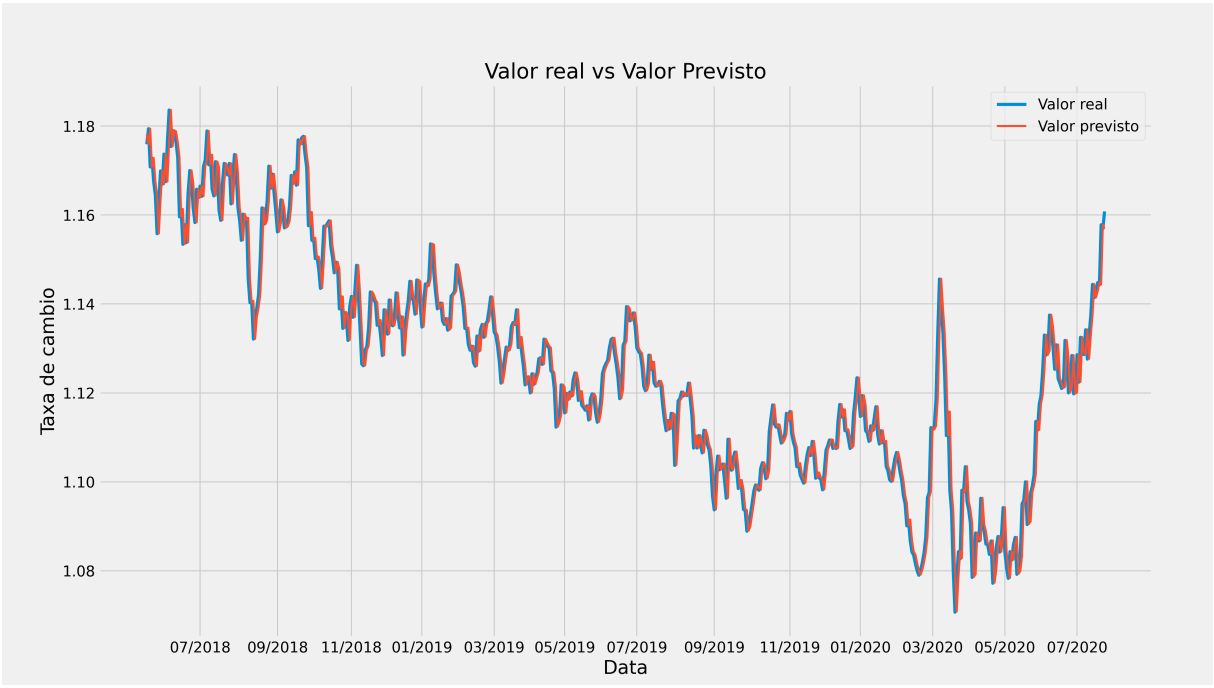

Fig. 2: Valor Real vs Valor Previsto para o conjunto de medida.