MO432 - Aprendizado Supervisionado
Trabalho 04

Patrick de Carvalho Tavares Rezende Ferreira - 175480

10 de agosto de 2020

1 Introdução

O objetivo deste trabalho é realizar previsões em séries temporais, trabalhando sobre um dataset que descreve a variação da taxa de câmbio entre dólar e euro de 1999 até julho de 2020. As grandezas a serem previstas são a taxa de câmbio no dia seguinte e a previsão de alta ou queda para o próximo dia, ambas levando em consideração uma janela W que resume os fechamentos de dias anteriores.

2 Descrição do algoritmo elaborado

Antes de utilizar os dados para qualquer procedimento, é necessário tratá-los e preprocessá-los adequadamente. Em seguida, os dados são utilizados para treinamento dos melhores regressores e classificadores utilizados nos modelos já vistos até aqui e então avaliados. As seções a seguir descrevem este procedimento.

2.1 Preprocessamento dos dados

Os dados são fornecidos em um arquivo CSV com duas colunas, sendo a primeira pra a data da cotação e a segunda para a taxa de câmbio em si. O primeiro procedimento é eliminar a coluna de data conforme solicitado pelo roteiro. Em seguida, antes de fazer qualquer consideração acercados dados sendo utilizados, é necessário separar o conjunto de teste (ou de medida), que será utilizado ao final do experimento para verificarmos se o treinamento resultou em um bom preditor. O conjunto de teste consiste nos 10% dos dados mais recentes, enquanto que os 90% restantes serão utilizados para o processo de treino e validação cruzada.

Realizamos o procedimento de centering e scaling dos dados de treino, sem levar em conta os dados de teste, no qual realizamos a mesma transformação com os parâmetros obtidos na primeira, a fim de evitar contaminação. Os dados de treino são então divididos em 5 splits de 2 folds, conforme representado na figura 1. Cada split 50% para treino e 50% para validação, sendo concatenado com os dados de treino de cada split anterior a serem usados junto com o treino.

A cross validação é realizada de forma compatível com a API do sklearn[1], facilitando a busca pelos hiperparâmetros (W) e classificadores envolvidos.

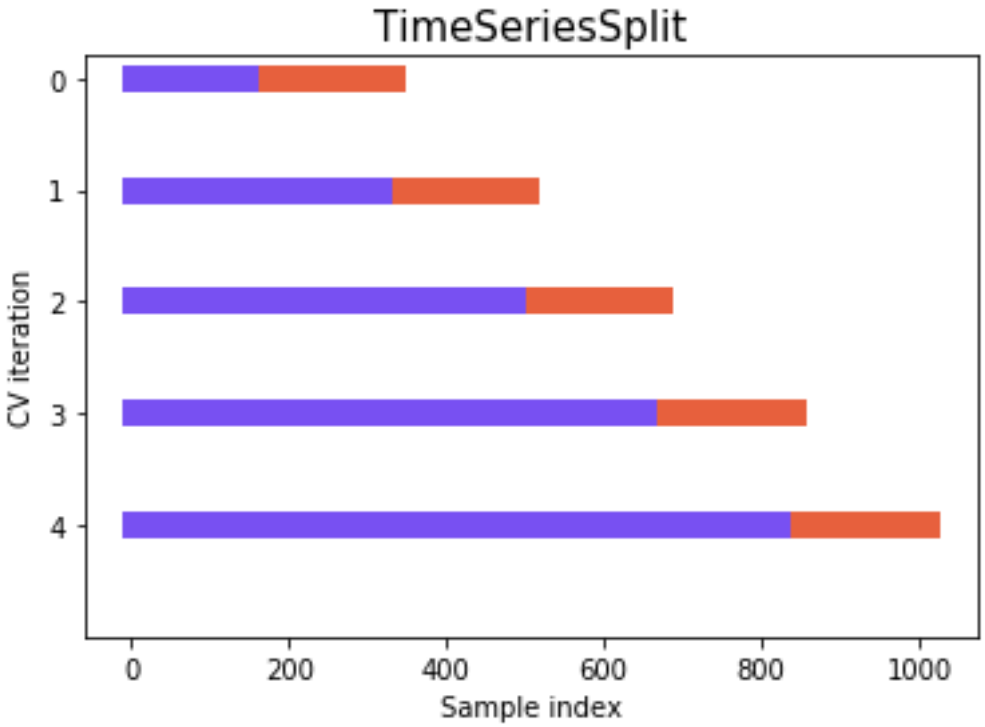

Figura 1: Estrutura dos splits de cross-validation[2].

2.2 Busca por regressores

Com base nos experimentos anteriores, os melhores regressores encontrados foram Random Forest, SVM, Gradient Boosting Machine (GBM) e MLP, todos com parâmetros default para não tornar a busca muito demorada, já que comprovou que a maioria dos default do sklearn possuem bom desempenho.

Foi utilizada então uma busca do parâmetro W (janela de dias anteriores sendo considerados) com estes 4 tipos de regressores. O range de para esta janela era uma busca aleatória sobre uma distribuição uniforme de 8 a 20 dias, com 5 amostras. Os resultados são exibidos na seção 3.

2.3 Busca por classificadores

Assim como na busca por regressores, foram utilizados os classificadores de melhor desempenho nos experimentos até agora, sendo eles o Random Forest, SVM, GBM, MLP e Regressão Logística, todos com parâmetros default para não tornar a busca muito demorada, já que comprovou que a maioria dos default do sklearn possuem bom desempenho.

A busca novamente é pelo parâmetro W, com os classificadores supracitados. O range de para esta janela era uma busca aleatória sobre uma distribuição uniforme de 8 a 20 dias, com 5 amostras. Os resultados são exibidos na seção 3.

3 Resultados

Os resultados abaixo são exibidos apenas para as melhores combinações de regressores e classificadores. A tabela 1 tem os resultados para regressores e a tabela 2 para os classificadores.

Tabela 1: Resultado dos regressores.

Regressor	**Janela W [dias]**	**RMSE**
Random Forest	12	0.003904
Gradient Boosting Machine	14	0.006594
SVM	18	0.044295

Tabela 2: Resultado dos classificadores.

Classificador	**Janela W [dias]**	**Acurácia**
Random Forest	11	0.900806
Gradient Boosting Machine	14	0.708282
MLP	11	0.512096

Os melhores resultados se concentraram em Random Forest, GBM, o que era esperado. O SVM provavelmente obteria um melhor desempenho se fosse realizada uma busca de parâmetros, mas esta se demonstra muito lenta e não se distanciou significativamente da Random Forest nos últimos experimentos, então foram mantidos os parâmetros default. Em ambos os casos, a Random Forest obteve um resultado melhor que os demais, então foi utilizada para o teste nos dados de medida.

No conjunto medida, o RMSE do Random Forest foi avaliado em 0.005449, um aumento de 40% no erro. O valor do RMSE foi mantido baixo, embora tenha tido um previsível aumento dado o fato que escolhemos o conjunto de parâmetros que possui o melhor desempenho nos dados de teste, produzindo um certo overfitting.

Para a acurácia, a perda de desempenho foi mais significativa, caindo de 90% de acerto para 52,4% no conjunto de medida ao se utilizar o Random Forest. É claro que existe um processo de overfitting ao se escolher o classificador com melhor desempenho nos dados da validação, e 90% é realmente um valor muito alto de acurácia, mas há outros fatores que influenciam o desempenho nestes dados de teste em específico. Por exemplo, o conjunto de testes se concentra aproximadamente entre 2018 e 2020, em que ocorreram eventos fora do comum em termos de economia, como o Brexit e a pandemia de coronavirus. Estes foram eventos bastante atípicos e que afetaram os mercados e taxas de câmbio de uma forma incomum e para a qual o classificador não estava familiarizado no treino, o que pode justificar esta perda de desempenho.

Além dos eventos mais recentes, os dados de alta ou baixa na taxa de câmbio são binários, então mesmo que a rede erre a variação real prevista por muito pouco, esta pode ser a diferença entre prever um dia de alta ou baixa em períodos de variação menos intensa e diminuir a acurácia. Isto explicaria o fato de obtermos um erro RMSE relativamente baixo enquanto que a acurácia não supera tão bem os 50% de acerto.

4 Conclusão

Este roteiro permitiu realizarmos experimentos para tentar obter os melhores preditores para a taxa de câmbio dólar/euro com base em um dataset que a registra há mais de 10 anos. É evidente que estas atividades relacionadas ao mercado financeiro são afetadas por incontáveis possíveis fatores e que olhar apenas o histórico das mesmas não permite prever com grande acurácia quais serão seus próximos valores.

A baixa acurácia encontrada indica porque não há ainda uma hegemonia da predição dos mercados financeiros através dos métodos de aprendizado de máquina, embora a predição de valores de tendência tenha demonstrado um RMSE relativamente baixo e que pode ser utilizado para operações a médio prazo.

Algumas técnicas como ensemble (voto majoritário) e mistura de especialistas poderiam ser aplicada para se obter um desempenho ainda mais elevado em ambas as predições (classificação e regressão), porém estas são técnicas que conhecidamente aumentam o desempenho por unir vários classificadores que divergem no erro, não implicando em um ganho de desempenho através da busca de melhores estruturas de estimadores ou hiper-parâmetros em si e, por isso, não foram aplicadas aqui.

Referências

[1] *scikit-learn: machine learning in Python — scikit-learn 0.23.2 documentation*. URL: `https://scikit-learn.org/stable/` (acesso em 11/08/2020).

[2] Packt Editorial Staff. *Cross-Validation strategies for Time Series forecasting [Tutorial]*. en-US. Section: Featured. Mai. de 2019. URL: `https://hub.packtpub.com/cross-validation-strategies-for-time-series-forecasting-tutorial/` (acesso em 11/08/2020).