IND2604 - Proposta de Projeto Final

Matheus Nogueira (2412716)
PUC-Rio
.
matnogueira@gmail.com

RESUMO

Documento contendo a proposta de projeto final para a disciplina IND1604 Probabilidade e Estatística.

1. Descrição do problema

É de conhecimento geral que notícias podem impactar o preço de ativos nas bolsas de valores de diferentes países e que, somado a isso, a percepção do público sobre um determinado ativo ou cenário macroeconômico é uma variável importante para entender, ou até prever, o desempenho do mercado de capitais.

O presente projeto propõe-se a investigar a existência de uma correlação entre os dados de pesquisa do *Google Trends*, utilizados como uma medida quantitativa da percepção e interesse da população brasileira sobre determinados temas, e os preços de fechamento de um ativo específico da bolsa de valores. O *Google Trends* é uma ferramenta que permite analisar o volume de pesquisas realizadas pelos usuários no mecanismo de busca do Google ao longo do tempo, fornecendo *insights* sobre o interesse público em determinados tópicos.

A hipótese subjacente a este estudo é que variações no interesse público por temas correlatos à economia e finanças, refletido nas pesquisas no Google, podem estar correlacionadas com mudanças nos preços de fechamento de ativos na bolsa de valores. Tal correlação potencial pode ser explorada para desenvolver estratégias de análise e previsão de mercado mais robustas.

O principal trabalho que inspirou esse projeto é Preis et al. [2013], no qual foi implementada uma estratégia de compra e venda de um ativo financeiro da bolsa americana baseado na variação do *trends* de diferentes termos relacionados à finanças e economia. Embora o resultado financeiro obtido neste trabalho tenha sido significativo, a referência não apresentou nenhum estudo formal que buscasse quantificar ou medir o grau de correlação, ou de explicabilidade, dos *google trends* em relação ao preço de fechamento do ativo. É justamente esta lacuna que este projeto final pretende preencher.

2. Objetivos

O objetivo deste projeto é estudar se notícias ou interesse popular em assuntos relacionados à economia e finanças podem impactar o mercado financeiro brasileiro.

Com esse objetivo em mente, foram escolhidos os *google trends* como medidas quantitativas de interesse popular, uma vez que eles representam o volume de buscas de determinados termos na plataforma de pesquisa *Google*. Além disso, como indicador do andamento do mercado brasileiro, foi escolhido o preço de fechamento diário do índice Ibovespa, principal índice do mercado de capitais nacional.

Para avaliar se existe essa relação entre *google trends* e o preço de fechamento do Ibovespa, serão utilizadas diferentes técnicas estatísticas, a serem detalhadas na seção 4.

3. Dados

Os dados a serem utilizados neste projeto final consistem em dois tipos:

- Séries financeiras históricas
- Séries históricas do Google Trends

As séries financeiras consistem, primeiramente, da série histórica de preço de fechamento ajustado diário das ações (ou índices) a serem estudados. Como comentado na seção de objetivo, primeiramente planejo investigar os valores de fechamento diário do índice IBOVESPA, reconhecido pelo *ticker* "^BVSP". Essa série será retirada do site Yahoo Finance.

Além dessa série, também serão utilizadas séries de variáveis macroeconômicas, a saber, SELIC e CDI diários, obtidos por meio de um pacote em *Python* python-bcb que disponibiliza APIs para o site do Banco Central do Brasil.

Os dados principais deste projeto serão os *Google Trends* históricos diários de 10 termos relacionados com economia e investimentos. Os 10 termos são: 'Ibovespa', 'Bolsa de Valores', 'Ações', 'Dividendos', 'Renda Fixa', "Inflação", "CDI", "Dolar", "Bitcoin", "Renda Variável". Todos os *trends* foram obtidos a partir de um pacote *Python* pytrends que implementa APIs de acesso ao site Google Trends.

É importante ressaltar alguns fatos sobre a natureza do *Google Trends*. Primeiramente, os valores de *trends* não são absolutos, mas sim relativos ao período de busca selecionado e sempre contidos no intervalo de 0 a 100. Isso quer dizer que, para um mesmo termo em uma mesma data, o valor do *trend* pode ser diferente dependendo do período de busca selecionado. Além disso, não é possível especificar, a priori, a granularidade dos *trends*, isto é, a plataforma não permite que o usuário escolha por *trends* diários, semanais ou mensais. Essa granularidade é definida a partir do período de busca, por exemplo: buscas de um intervalo de 1 mês retornam *trends* diários e buscas de um intervalo de 1 ano geram *trends* semanais. Essas duas características dos *google trends* podem ser um desafio tanto para a obtenção dos dados quanto da interpretação de seus resultados. Por fim, a plataforma do *Google Trends* permite especificar a região de busca dos termos. Isso foi utilizado para garantir que as buscas dos 10 termos listados acima fosse restrita ao Brasil, visto que este projeto se limita a estudar o mercado de capitais brasileiro.

Todos os dados históricos foram obtidos para o intervalo de 01-01-2010 até 01-01-2024, ignorando feriados e finais de semana, isto é, contabilizando apenas os dias úteis.

4. Metodologia

Podemos separar a metodologia deste projeto em duas partes: (1) a preparação dos dados para as análises a serem realizadas e (2) a definição de quais análises serão realizadas, o que inclui especificar quais métodos serão implementados.

Em relação ao ponto (1), o objetivo é montar um *dataset* com a série temporal de preço de fechamento a ser explicada e todas as variáveis explicativas. Vale, portanto, definir algumas notações:

- y_t é o valor da série de preço de fechamento do ativo a ser explicado no tempo t;
- $P^{(y)}$ o conjunto de *lags* a ser utilizado para a variável dependente;
- y_{t-p} é o valor da série de preço de fechamento defasada em $p \in P^{(y)}$ dias;
- $g_t^{(i)}$ é o valor do *google trend* do termo i no instante t;
- $\Delta g_t^{(i)} = g_t^{(i)} - g_{t-1}^{(i)}$ é a variação do *google trend* do termo i entre os dias $t-1$ e t;
- $P^{(g)}$ é o conjunto de *lags* a ser utilizado para os *google trends*;
- $g_{t-p}^{(i)}$ é o valor do *google trend* do termo i no instante $t-p$ com $p \in P^{(g)}$;
- $\Delta g_{t-p}^{(i)} = g_{t-p}^{(i)} - g_{t-p-1}^{(i)}$ é a variação do *google trend* do termo i entre os dias $t-p-1$ e $t-p$;
- $x_t^{(j)}$ é o valor da variável explicativa j no instante t com $p \in P^{(g)}$.

Uma vez definidos esses dados, o núcleo da parte (2) é definir uma função f como:

$$y_t = f(y_{t-p}, g_{t-p}^{(i)}, \Delta g_{t-p}^{(i)}, x_t^{(j)}) \forall i, p, j, t$$

De tal forma que essa função me permita, de algum modo, estudar a correlação, ou explicabilidade, ou dependência da variável dependente em relação a cada variável independente.

Ainda não foram definidas as ferramentas estatísticas para estudar essa relação, mas algumas estão sendo pesquisadas para entender qual pode melhor se encaixar nesse projeto. São elas:

- Regressão Linear e testes de hipótese (t e F) para significância das variáveis explicativas;
- LASSO para seleção de variáveis explicativas;
- Marginal Effects para estudo do efeito marginal das explicativas em relação à dependente.
- Boruta para obtenção de importância de variáveis a partir de um modelo de *Random Forest*.

Todas essas técnicas já estão implementadas em *Python*, de tal forma que não excluo a possibilidade de utilizar mais de uma para estudar a relação dos *google trends* e o valor diário de fechamento do índice IBOVESPA.

Por fim, também seria interessante montar uma estratégia de compra e venda do índice IBOVESPA para avaliar se, apesar da existência ou não de um grau de explicabilidade dos *google trends*, eles podem ser utilizados para definir uma estratégia de investimentos.

Referências

Preis, T., Moat, H. S., e Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Scientific reports*, 3(1):1–6.