Extreme Gradient Boosting para Classificação da Popularidade de Artigos Online – Plataforma Mashable Inc.

Aline Pelegrino Shikasho
Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas
Campinas, Brasil

Ana Alice Scalet
Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas
Campinas, Brasil

Diogo Henrique Dias
Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas
Campinas, Brasil

Fernando Henrique Guedes de Oliveira
Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas
Campinas, Brasil

***Resumo*— Mashable Inc. é um website de mídia social com cerca de 7.3 milhões de seguidores no Facebook e quase 9.9 milhões no Twitter. Plataformas online como a Mashable ou por exemplo, BuzzFeed, publicam centenas de artigos todos os dias. Esses artigos podem se enquadrar em diversas categorias como tecnologia, esportes, entretenimento, e serem publicados em diferentes dias da semana. Foram coletadas informações de aproximadamente 39000 artigos publicados por esta plataforma entre os anos de 2013 e 2015. Com base nisso, algoritmos de predição baseados no conceito de árvore foram implementados com o propósito de classificar um artigo como sendo popular ou não. A ideia principal foi proporcionar aos autores dos artigos um maior conhecimento sobre a popularidade destes antes de publicá-los. Os algoritmos utilizados foram o *Extreme Gradient Boosting*, a Árvore de Decisão, o *Bagging* e a Floresta Aleatória, onde o intuito foi comparar o desempenho, medido através da acurácia, do algoritmo *Extreme Gradient Boosting* com os demais supracitados. Por ordem decrescente de acurácia, tem-se o *Extreme Gradient Boosting* (0.6712), a Floresta Aleatória (0.6594), o *Bagging* (0.6568) e a Árvore de Decisão (0.6195).**

Palavras-chave: Extreme Gradient Boosting, Classificação, Popularidade Online

I. Introdução

Com o avanço da tecnologia no mundo, mais pessoas desfrutam de acesso à internet e, consequentemente, às informações nela contidas. Devido a isso, muitos jornais, revistas e outros meios de comunicação estão migrando e adicionando seus conteúdos para o sistema de plataformas online.

Há, portanto, uma preferência da população por materiais online, e isso demonstra que são necessárias mudanças nas técnicas de vendas e marketing no geral. Novas técnicas incluem propagandas em sites, rastreamento e coleta de informações pessoais, todas voltadas à maior lucratividade e venda dos produtos das empresas. As que procuram colocar propagandas em sites, por exemplo, necessitam torná-las sempre visíveis para o maior número de pessoas possível.

Um método útil para aumentar a abrangência das propagandas é analisar as visualizações das postagens dos sites. Um número elevado de visualizações gera um bom retorno para as empresas das propagandas.

O caso de interesse deste estudo, portanto, é a predição da popularidade dos artigos postados na plataforma online Mashable Inc dentro do período especificado. Ao predizer se um artigo será popular ou não, de acordo com suas características, prediz-se algo relevante e interessante para a lucratividade da Mashable Inc.

II. Banco de Dados

O banco de dados utilizado é formado por 39644 artigos da Mashable Inc., publicados entre os anos de 2013 e 2015. Cada artigo foi caracterizado a partir de 61 variáveis

(categóricas e numéricas), sendo algumas destas o assunto do artigo, a quantidade de palavras presentes em seu título, o dia da semana em que foi publicado, dentre outras.

A medida de interesse é o número de compartilhamentos do artigo apresentada no banco de dados com o nome Shares. Essa medida é quantitativa, isto é, um número. Além disso, é não negativo e inteiro. Como o interesse é predizer se um novo artigo será popular ou não, o número de compartilhamentos foi categorizado tomando como referência o seu valor mediano.

Com isso, a nova medida foi chamada de Popularidade e possui duas categorias, sendo 1 para popular e 0 para não-popular.

III. Metodologia

Foram utilizados métodos de predição, através do pacote estatístico R, para classificar a popularidade dos artigos. Quatro algoritmos baseados no conceito de árvore em sua formação foram escolhidos com essa finalidade. Algoritmos com essa característica são amplamente utilizados em diversas áreas, para fins de predição de valores quantitativos (regressão) ou de categorias (classificação), sendo este último o caso deste projeto.

O primeiro algoritmo e também o mais simples, é a Árvore de decisão. O segundo é o Bagging, que pode ser visto como um conjunto de Árvores de decisão e, além disso, ser tratado como um método ensemble (grupo). O penúltimo é o algoritmo Floresta aleatória, que também é considerado um método ensemble e é uma melhoria do algoritmo Bagging. O último algoritmo é o Extreme Gradient Boosting, que é o mais complexo dos quatro e vem sendo amplamente utilizado por ser bem eficaz, geralmente retornando boas predições.

Uma maneira de determinar o quão eficiente é o algoritmo implementado é verificar a sua acurácia, isto é, calcular a proporção de artigos não-populares classificados como não-populares pelo algoritmo (VN - Verdadeiro Negativo), juntamente com os artigos populares classificados como populares pelo mesmo (VP - Verdadeiro Positivo).

Ambos os casos citados acima são casos em que o algoritmo implementado executa um acerto. Por outro lado, também existem os casos em que este executa um erro, isto é, classificar artigos que são populares como não-populares (FN - Falso Negativo) e classificar os que não são populares como populares (FP - Falso Positivo).

A tabela a seguir ilustra como a acurácia de um algoritmo com a finalidade de predizer categorias é calculada.

		Predito	
		0 - Não-popular	1 - Popular
Real	0 - Não-popular	VN	FP
	1 - Popular	FN	VP

$$\text{Acurácia} = \frac{VN + VP}{VN + FP + FN + VP}$$

Figura 1: Ilustração do cálculo da acurácia.

De acordo com a Figura 1, pode-se notar que a acurácia é representada por um número que varia de 0 a 1. Quanto maior o valor da acurácia, melhor é o algoritmo. Por exemplo, um valor de acurácia igual a 1 indica que este classificou corretamente todos os artigos populares como populares e também classificou corretamente todos os artigos não-populares como não-populares.

Na prática, obter uma acurácia igual a 1 em observações novas é praticamente impossível por qualquer que seja o algoritmo.

A. Árvore de Decisão

A Árvore de decisão é um método de treinamento supervisionado baseado na segmentação do espaço preditor em áreas mais simples. É indicado para problemas de regressão (resposta quantitativa) e classificação (resposta qualitativa). No caso deste artigo, utiliza-se como classificação.

A partir da divisão dos dados em dados de treinamento e dados de teste, a árvore é calibrada pelos de treinamento, criando então um modelo de predição geral. Este, é posteriormente executado pelas observações de teste. De acordo com as características e regiões às quais se enquadram, as observações de teste serão classificadas em alguma das categorias da variável resposta.

A seguir, é apresentado um exemplo de uma árvore de classificação.

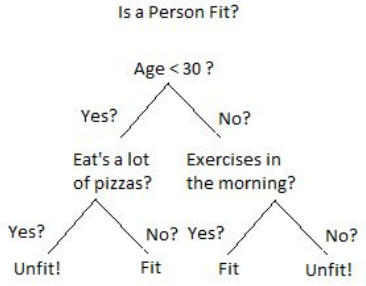

Figura 2: Árvore de Classificação.

A árvore é construída por perguntas Sim ou Não, além das respostas para cada um dos casos em cada ramo. De acordo com a Figura 2, vê-se que se uma pessoa tiver menos

de 30 anos e comer muita pizza, ela irá ser rotulada como "Não Fitness".

Os acertos e erros resultantes do algoritmo são analisados por uma tabela, chamada matriz de confusão (mostrada na Figura 1), além de sua acurácia.

B. *Bagging*

É um algoritmo e é também visto como uma técnica utilizada para reduzir a variância das predições. Nesta, combina-se o resultado de vários classificadores, modelados em diferentes sub-amostras do mesmo conjunto de dados.

Dado isso, tem-se as seguintes etapas do *Bagging*:

- 1- Criar vários conjuntos de dados: A amostragem é feita com a substituição dos dados originais e formação de novos conjuntos de dados (amostras *bootstrap* dos dados originais). Constrói-se *k* novos conjuntos idênticos e replicam-se esses dados de forma aleatória, para construir *k* conjuntos independentes por reamostragem com reposição.
- 2- Criar múltiplos classificadores: Classificadores são construídos em cada conjunto de dados. Geralmente, o mesmo classificador é modelado em cada conjunto de dados, e a partir disso as previsões são feitas.
- 3- Combinar classificadores: As previsões de todos os classificadores são combinadas usando-se a média ou a maioria dos votos, dependendo do problema em si.

Os valores combinados são geralmente mais robustos do que em um único modelo. Um maior número de modelos é geralmente considerado melhor, ou resulta em um desempenho semelhante ao de números mais baixos. Pode-se mostrar que, em teoria, a variância das previsões combinadas é reduzida para $\frac{1}{n}$ (*n* número de classificadores) da variância original, sob algumas premissas.

C. *Floresta Aleatória*

Floresta aleatória é um método de *ensemble* para classificação. Os métodos de *ensemble* utilizam uma combinação de outras técnicas estatísticas para obter uma melhor performance de predição.

Na Floresta aleatória, criam-se diversas Árvores de decisão, com o objetivo de diminuir eventuais casos de *overfitting* (quando o modelo criado ajusta-se perfeitamente aos dados de treino, tendo uma performance ruim de predição para os dados de teste) que podem surgir em uma única Árvore de decisão.

As árvores da Floresta aleatória são construídas a partir de uma seleção aleatória de um número *m* de variáveis do banco de dados. Com essas variáveis selecionadas, criam-se os nós de uma Árvore de decisão. Para cada nó de uma árvore nova criada para essa floresta, são selecionadas outras *m* variáveis aleatórias. O número de árvores presentes no modelo pode variar, porém um maior número delas não significa em uma melhor predição.

Para a predição final de uma observação, a mesma passa por todas as árvores presentes no modelo e a floresta escolhe a classificação final de acordo com a categoria mais escolhida entre as árvores, ou seja, pelo voto da maioria.

A técnica de Floresta aleatória também é utilizada para mensurar a importância das variáveis de um banco de dados.

Para o cálculo desta medida, em cada árvore conta-se o número de observações que foram preditas corretamente no conjunto de teste, e aleatoriamente permuta-se os valores da variável *m* na árvore. As observações são repassadas nesta árvore recém-criada, e conta-se o número de acertos das mesmas. Subtrai-se o número de acertos iniciais pelo número de acertos obtidos da árvore com a variável *m* aleatorizada. O valor médio obtido em todas as árvores resultará no nível de importância da variável *m*. Os níveis de significância de todas variáveis podem ser normalizados utilizando-se o desvio padrão dessas diferenças.

D. *Extreme Gradient Boosting*

É uma variante do algoritmo *Gradient Boosting*, que é uma das técnicas atuais mais poderosas na criação de modelos preditivos. *Gradient Boosting* é um modelo composto que combina diversos classificadores fracos de modo que, ao final, seja possível obter um classificador poderoso e forte. A ideia da aprendizagem por trás do algoritmo *Gradient Boosting* é criar diversos classificadores que possuam uma relação com o seu classificador anterior, isto é, criar classificadores de modo sequencial.

Uma vez que se crie um classificador, este deve produzir predições erradas que precisam ser convertidas em predições corretas. Para tal, cria-se o próximo classificador utilizando-se dessa informação associada aos erros cometidos pelo classificador anterior, de maneira que o atual possa dar um maior peso para os artigos que já foram classificados de modo errado e que precisam ser classificados de modo correto.

Com isso, ao longo da sequência de classificadores, espera-se que a quantidade de erros cometidos pelo algoritmo diminua, de maneira que, ao final, este produza uma boa acurácia.

O algoritmo *Extreme Gradient Boosting* surge como uma melhoria do algoritmo *Gradient Boosting* permitindo alguns controles interessantes e eficazes como, por exemplo, a ideia de regularização e processamento em paralelo. Este último faz com que o algoritmo seja processado em um menor período de tempo, o que na prática pode representar um menor custo.

A regularização permite evitar o já mencionado *overfitting*, que é uma grande preocupação quando se fala em *Machine Learning*.

IV. Desenvolvimento e Resultados

Inicialmente, verificou-se que não há existência de valores faltantes, erros de digitação e valores sem sentido no banco de dados. Dado isso, uma varredura foi realizada nas variáveis para garantir que o seu tipo estivesse correto.

Percebeu-se que haviam oito variáveis *dummy* associadas aos dias da semana, sendo que sete delas representavam propriamente cada dia da semana e uma delas representava se o artigo havia sido postado durante o final de semana ou não.

Em resumo, a variável citada acima estava perfeitamente correlacionada com as variáveis que retratavam a publicação de um artigo ter ocorrido em um Sábado ou em um Domingo, ou seja, a variável relacionada ao final de semana continha a mesma informação das demais. Então, alguma ação deveria ser tomada: ou retirar a variável relacionada ao final de semana e manter as outras sete, ou o contrário.

Para se chegar a uma conclusão, foi construído um gráfico do tipo boxplot do logaritmo do número de compartilhamentos dos artigos ao longo de todos os sete dias da semana. O logaritmo foi utilizado para tornar melhor a visualização da dispersão dos dados, visto que a escala do número de compartilhamentos se altera, ficando mais compactada. Este gráfico é apresentado na figura abaixo.

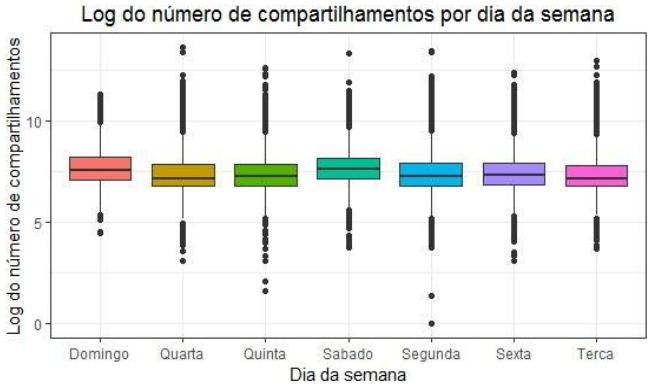

Figura 3: Boxplot do log do número de compartilhamentos por dia da semana.

De acordo com a Figura 3, verifica-se que o número de compartilhamentos dos artigos nos diferentes dias da semana não parece divergir muito em seu valor mediano, porém a variabilidade existente no número de compartilhamentos é razoavelmente diferente ao longo dos diferentes dias da semana. Portanto, decidiu-se manter as variáveis relacionadas aos sete dias da semana e excluir a variável relacionada ao final de semana.

Outro passo importante foi procurar não excluir as variáveis relacionadas às naturezas dos artigos, por entender que tal informação seria interessante para o problema e deveriam estar presentes na análise.

Sobretudo, é apresentado a seguir um gráfico do tipo boxplot do logaritmo do número de compartilhamentos por natureza.

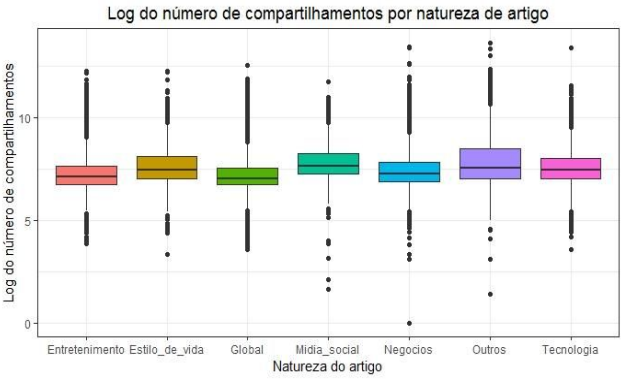

Figura 4: Boxplot do log do número de compartilhamentos por natureza do artigo

De acordo com a Figura 4, tem-se que as naturezas dos artigos não se diferenciam muito em relação a mediana, havendo uma pequena diferença na dispersão dos dados com relação a cada natureza. No entanto, decidiu-se manter as variáveis.

Um passo importante durante a análise foi verificar a existência de possíveis outliers. Alguns potenciais outliers foram encontrados, porém, ao se testar os diferentes algoritmos, os mesmos apresentaram comportamento igual, com ou sem esses pontos. Tal acontecimento leva a acreditar que modelos baseados em árvores são robustos quando se trata de outliers. Com isso, decidiu-se manter todos os artigos do banco de dados.

O último passo antes da implementação dos algoritmos foi a categorização da medida de interesse, denominada como *Shares*. Foi decidida a criação de duas categorias, 0 e 1, sendo que 1 representa que o artigo é popular e 0 representa que o artigo não é popular. Tal passo foi executado tomando como referência a mediana do número de compartilhamentos.

Uma ilustração do antes e depois é mostrada na figura a seguir.

Antes	
Mínimo	1
1° Quartil	946
Mediana	1400
Média	3395
3° Quartil	2800
Máximo	843300

Depois	
0	1
20082	19562

Figura 5: Ilustração da distribuição da medida de interesse, antes e depois da categorização

De acordo com a Figura 5, nota-se que antes a medida quantitativa de interesse, o número de compartilhamentos, tinha mínimo igual a 1, mediana igual a 1400 e máximo igual a 843300. Decidiu-se adotar como divisor entre as

categorias 1 e 0 a mediana, e não a média, já que o número de compartilhamentos não é simétrico em torno de sua média.

Após a categorização baseada na mediana, a nova medida categórica de interesse passou a possuir um total de 20082 artigos não-populares e um total de 19562 artigos populares. Em outras palavras, todos os 20082 artigos não-populares são artigos que possuem o número de compartilhamentos menor ou igual ao valor de 1400. Da mesma maneira, todos os 19562 artigos populares são artigos que possuem o número de compartilhamentos maior que 1400.

Dentro da fase de implementação dos algoritmos, o primeiro a ser implementado foi a Árvore de decisão, por ser o mais simples, de fácil interpretação e também por ser possível visualizar com clareza os critérios de classificação.

A árvore resultante é apresentada a seguir.

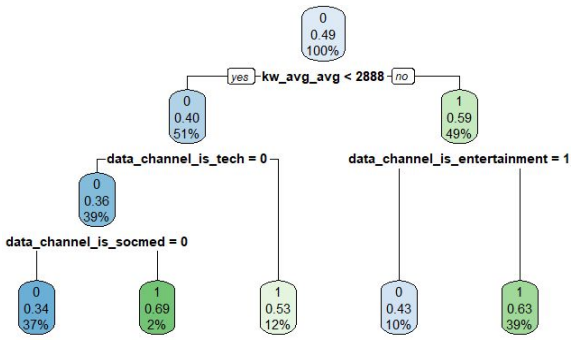

Figura 6: Árvore de decisão do algoritmo

De acordo com a Figura 6, é possível verificar que somente quatro variáveis foram utilizadas na construção da árvore. Todas as outras variáveis não foram consideradas. A acurácia resultante desse algoritmo foi de 0.6195.

Por uma breve interpretação da árvore tem-se que, se o número médio de palavras contidas no artigo for maior ou igual a 2888, o artigo será considerado popular. Porém, se a natureza do artigo for entretenimento, a popularidade muda.

Em contrapartida, se o número médio de palavras for menor que 2888, a natureza não for tecnologia e nem mídias sociais, o artigo é considerado como não popular.

O segundo algoritmo implementado foi o *Bagging*. Esse algoritmo possui um hiperparâmetro a ser escolhido, que é o número de árvores que ele irá considerar. Foram considerados cinco possíveis valores para esse hiperparâmetro, sendo eles 40, 80, 120, 160 e 200 árvores. Como esse algoritmo é um conjunto de diversas árvores, não é possível apresentá-lo visualmente.

Sobretudo, é apresentado abaixo um gráfico do valor da acurácia de acordo com o número de árvores escolhido.

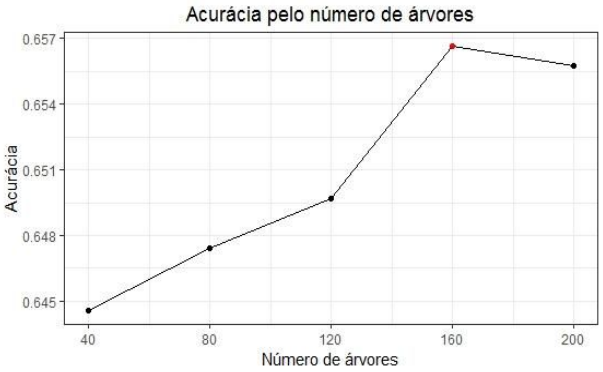

Figura 7: Valor da acurácia de acordo com o número de árvores

É possível verificar pela Figura 7 que a maior acurácia obtida com o algoritmo *Bagging* foi considerando o número de árvores igual a 160. O valor resultante da acurácia foi de 0.6568.

O terceiro algoritmo implementado foi a Floresta aleatória. Este algoritmo possui dois hiperparâmetros, sendo eles o número de árvores considerado e o número de variáveis amostradas para a criação de cada nó. Para o número de árvores, foram considerados valores iguais a 100, 200, 300, 400 e 500. Para o número de variáveis amostradas por nó, foram considerados os valores de 7 e 8, para que a raiz do número de variáveis utilizadas pelos algoritmos, $\sqrt{58} = 7.61$, estivesse dentro desse intervalo e assim seguisse o consenso, que é utilizar $\sqrt{p}$. Com isso, dez diferentes classificadores foram criados.

A seguir é apresentado o gráfico com o valor da acurácia para cada número de árvores e para cada número de variáveis amostradas por nó.

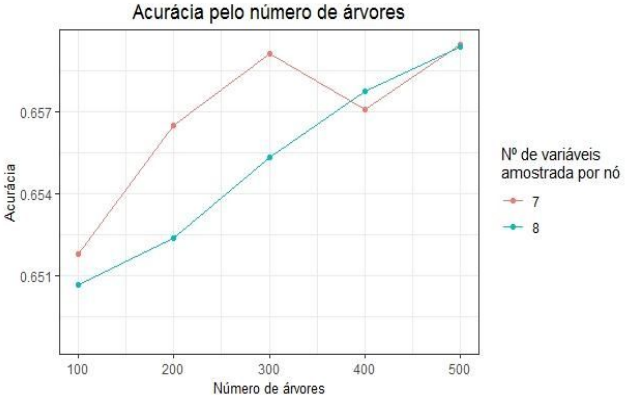

Figura 8: Valores da acurácia pelo número de árvores

De acordo com a Figura 8, é possível verificar que a combinação de hiperparâmetros que resultou na maior acurácia foi com um número de árvores igual a 500 e com número de variáveis amostradas por nó igual a 7. O valor da acurácia obtida foi de 0.6594.

O último algoritmo implementado foi o *Extreme Gradient Boosting*. Este algoritmo possui diversos parâmetros.

Foram controlados seis parâmetros. São eles:

- 1° - A taxa de aprendizado, que auxilia no controle contra *overfitting*.
- 2° - A profundidade máxima de cada árvore.
- 3° - O número máximo de iterações do algoritmo, que representa a quantidade de árvores construídas de modo sequencial.
- 4° - A proporção de artigos, dentre os 39644 artigos possíveis, coletada através de amostragem para a construção de cada uma das árvores.
- 5° - A proporção de variáveis, dentre as 58 variáveis possíveis, coletada através de amostragem para a construção de cada uma das árvores.
- 6° - A redução de perda mínima necessária para a criação de um novo nó.

A melhor combinação foi encontrada variando todos os seis parâmetros dentro de uma grade de valores para cada um.

A estratégia adotada foi variar um parâmetro e considerar todos os anteriores fixos nos seus melhores valores encontrados, isto é, que retornasse a maior acurácia, e considerar os próximos com seus valores padrão.

A tabela contendo a melhor combinação dos valores dos parâmetros é mostrada a seguir.

	Nome	Valor
1°	*eta*	0.21
2°	*max_depth*	3
3°	*nround*	150
4°	*subsample*	0.8
5°	*colsample_bytree*	0.7
6°	*gamma*	0.96

Figura 9: Combinação ótima de valores para os parâmetros

Considerando os valores dos parâmetros mostrados na Figura 9, a acurácia resultante obtida foi 0.6712.

Conclusão

Estabelecendo relações entre os algoritmos de Árvore de decisão, Floresta aleatória e *Bagging*, foi visto que a acurácia resultante das predições pelo *Extreme Gradient Boosting* foi um pouco maior do que as dos outros métodos. A melhora mais expressiva foi de pouco mais de 5%, deste em relação à Árvore de decisão.

A seguir são apresentadas as acurácias encontradas em cada algoritmo, em ordem crescente de acurácia.

Algoritmo	Acurácia
Árvore de decisão	0.6195
Bagging	0.6568
Floresta aleatória	0.6594
Extreme Gradient Boosting	0.6712

Figura 10: Valor da acurácia por algoritmo

De acordo com a Figura 10, é possível verificar que o algoritmo *Extreme Gradient Boosting* apresentou um poder preditivo de aproximadamente 67.12%.

Cabe ressaltar que os desempenhos dos algoritmos *Bagging* e Floresta aleatória foram bem próximos entre si, em torno de 65%, diferindo-se a partir da terceira casa decimal. O algoritmo Árvore de decisão atingiu quase 62% de poder preditivo, sendo o menos poderoso.

Ao ser considerado o tempo de execução e o gasto de recurso de máquina para cada algoritmo, tem-se que o *Bagging* e a Floresta aleatória são os algoritmos que exigem mais recurso de máquina, logo consomem um maior tempo para execução, sendo o *Bagging* o algoritmo que consome o maior tempo. Já a Árvore de decisão, que apresentou um poder preditivo menor em relação aos demais, em termos de execução conjuntamente com o gasto de recurso de máquina, é considerado o mais rápido, sendo seguido pelo *Extreme Gradient Boosting*.

Apesar de não possuir um amplo ganho na acurácia, em comparação aos outros, foi comprovado que o *Extreme Gradient Boosting* retorna predições em um bom período de tempo. Isso é uma ótima vantagem, já que hoje em dia cada vez mais os bancos de dados estão maiores. Outra vantagem vista foi a do retorno das variáveis mais importantes para o modelo.

Assim, tem-se a Figura 11 com as dez variáveis mais importantes para o algoritmo *Extreme Gradient Boosting*, em ordem decrescente de importância.

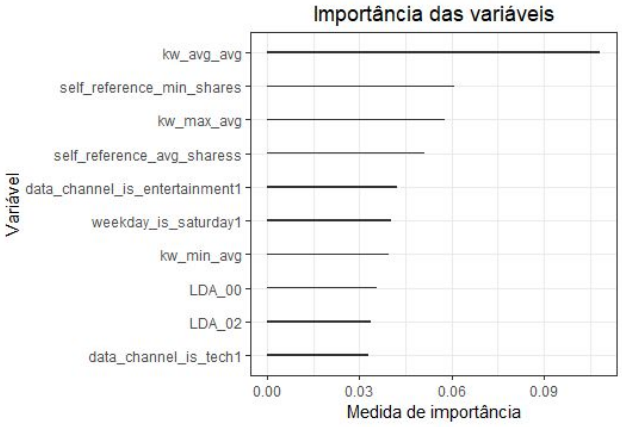

Figura 11: Importância das variáveis para o algoritmo *Extreme Gradient Boosting*

Problemas Encontrados

Inicialmente, houve uma dificuldade na determinação do significado das variáveis presentes no banco de dados. Em boa parte das variáveis não foi possível encontrar e entender o seu significado, o que fez com que não fosse optado por remover qualquer variável não conhecida.

Em um segundo momento, verificou-se a necessidade de testar a implementação de centenas de classificadores, de acordo com a variação de seus parâmetros, a fim de obter o classificar com o maior poder preditivo. O consumo de tempo foi relativamente grande.

Um grande desafio foi a determinação de como apresentar, em formato de gráficos, os passos que foram executadas na tentativa de encontrar o melhor classificador em cada algoritmo. Alguns algoritmos possuíam diversos parâmetros, o que tornou a decisão mais difícil.

No algoritmo *Extreme Gradient Boosting* não foi possível apresentar algo visual que transmitisse facilmente os passos que foram executados para se chegar na melhor combinação de parâmetros do mesmo.

O poder preditivo final alcançado em cada um dos quatro algoritmos apresentados foi razoável, porém poderia ter sido maior. Um grande obstáculo foi entender como, por exemplo, a acurácia poderia ser melhorada. Talvez uma saída poderia ser a melhor manipulação das variáveis disponíveis.

Referências

[1] Dua, D. and Karra Taniskidou, E. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2017

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York: Springer, 2013.

[3] Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc., 2001.