**Predição da classificação de livros do GoodReads**

**Beatriz de Carvalho Pacheco Lourenço**

**Fernando Kenji Sakabe**

**Pedro Henrique Menezes de Oliveira**

**Sophia Kerber Shigueoka**

1. **Ideia do Projeto**
2. **Pergunta e Relevância**
3. **Análise Exploratória**
4. **Limpeza do Código**
5. **Regressão Linear**
6. **Random Forest Classification**
7. **Aperfeiçoamento e Aprofundamento**
8. **Conclusões e proposta**
9. **Técnicas de verificação**
10. **Contribuições**
11. **Ideia do Projeto**

O projeto foi inspirado na ideia de prever a avaliação (*rating*) de livros do GoodReads a partir das variáveis encontradas no dataframe disponível no Kaggle. Um cenário no qual o modelo poderia ser utilizado é o de avaliar livros que ainda estão por sair, dando a possível avaliação dele a partir das variáveis escolhidas. Mais do que isso, poderia ser feito uma personalização do modelo para usuários individuais, vendo quais gêneros e autores o usuário mais lê e indicar o livro para ele.

1. **Pergunta e Relevância**

Pode um livro ser julgado por apenas algumas informações, como gênero, quantidade de páginas, etc, e não pelo conteúdo? Se nossa hipótese nula (que a avaliação do livro não pode ser prevista sem ter que ler o livro) for quebrada, um modelo preditivo é capaz de determinar o rating do livro apenas por dados superficiais (hipótese alternativa).

Nossa pergunta é de que se é possível prever a avaliação de um livro a partir do número de páginas, formato do livro (capa dura, capa de papel, versão Kindle…), inicial do autor, gênero do livro e número de reviews (ou seja, a quantidade de pessoa que comentaram nas suas avaliações).

1. **Análise Exploratória**

Foram feitas diversas análises exploratórias para dar um norte sobre quais variáveis deveriam ser levadas em conta quando utilizadas para fazer o modelo, por exemplo, relacionamos o rating com o gênero do livro, o rating com o número de reviews do livro, o rating com o número de ratings.

Pode-se perceber pelos scatter plots que, inicialmente, não parece haver uma relação entre variáveis quantitativas na análise. Apesar disso, mostra-se necessário justificar os motivos pelos quais o grupo pensa que poderia haver uma correlação entre as variáveis dos dados coletados.

Notamos que, como a avaliação dos livros varia entre 0 e 5 e somente com no máximo 2 casas decimais, temos um número finito de possibilidades. Além disso, como as avaliações estão, de maneira geral, concentradas na faixa de 3.5 a 4.5, temos um problema que se reflete em uma mudança pequena que cada variável exerce na avaliação. Assim, nossos gráficos exploratórios (sobretudo gráficos de dispersão) tem uma região hiper concentrada de dados que dificulta a visualização e análise de dados.

Exemplos disso:

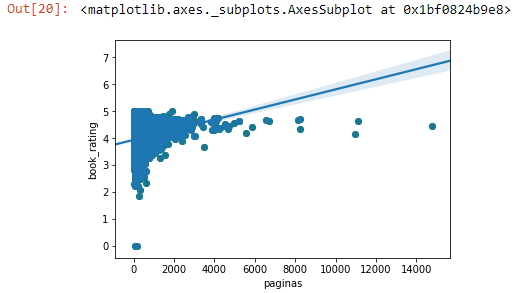


Gráfico 1 - Avaliação do livro em função do número de páginas

Dessa forma, tivemos que analisar mais profundamente nessa região concentrada de dados. Há um exemplo disso a seguir.

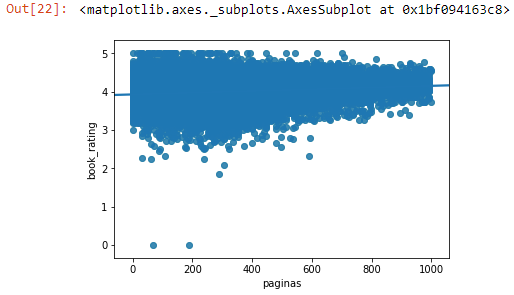


Gráfico 2 - Avaliação do livro em função do número de páginas (com no máximo 1000 páginas)

No gráfico acima é possível perceber que a influência do número de páginas até 1000 é praticamente desprezível, porém para livros com mais que isso é interessante utilizá-la, por isso ela foi deixada no modelo.

Uma hipótese que surge para explicar o comportamento que não aparenta ter nenhuma relação é uma correlação diferente entre as variáveis. É possível, por exemplo, que o modelo siga uma regressão espúria, ou seja, além das duas variáveis estudadas, há uma terceira, que afeta as duas outras, de modo que a relação não seja sempre causa e consequência.

1. **Limpeza do Dataframe**

Primeiramente, retiramos aquelas variáveis que não poderiam ser utilizadas ou não acrescentariam no modelo. Essas variáveis foram: descrição do livro, código de barras, a edição do livro, o número de avaliações (e não de reviews), o título da obra e a imagem da capa. Em seguida, retiramos todos os NaN do novo DataFrame sem as variáveis acima. Além disso, ajustamos as variáveis do nome do autor para pegar somente a letra inicial do primeiro nome; o número de páginas transformamos em uma variável quantitativa; e filtramos os livros nos vários gêneros a que eles pertenciam e só ficou com o gênero principal.

1. **Regressão Linear**

A regressão linear é um método útil para encontrar correlações entre duas variáveis contínuas. No projeto, utilizamos a regressão linear como um modo de previsão do rating de um livro. Encontrando uma correlação entre as variáveis, é possível aproximar os dados de uma linha, de modo que a soma dos quadrados dos erros seja diminuída. A partir dessa linha que é possível a partir de valores da abcissa, prever o valor da ordenada (rating).

1. **Classificação**

Após usarmos a regressão linear para nosso modelo e entender seu baixo rendimento, cogitamos usar o método de classificação para apurar nossos resultados e, por conseguinte, obter uma maior acurácia.

O random forest classification é um método para se fazer classificações levando em conta diferentes variáveis. Esse método é similar a uma distribuição de várias árvores de classificação com inputs não correlacionados entre si. No nosso caso, optamos por fazer a classificação dos livros usando este método.

1. **Aperfeiçoamento e Aprofundamento**

Para melhorar o modelo, poderíamos usar outro classificador além do GoodReads para uma maior exatidão das avaliações. Além disso, poderia-se também usar as palavras que foram utilizadas em cada livro (ou pelo menos, no primeiro capítulo) para enxergar se há alguma correlação entre elas e a avaliação do leitor. Por exemplo, um livro que usa palavras pouco usuais ou mais rebuscadas tem uma maior probabilidade de receber uma nota baixa, dado que seus leitores têm dificuldade de entendê-lo. Com mais detalhes, ainda seria possível usar o quanto de lucro um livro gerou para classificar com maior acurácia sua avaliação.

Para além, pensou-se em utilizar o método de regressão decision tree ou árvore de decisão, para que o modelo tivesse uma acurácia ainda melhor, uma vez que por meio de testes nas diferentes variáveis pode-se chegar em um resultado mais preciso. Parte dele está no Jupyter “decision tree.ipynb” e que apesar dos gráficos lá presentes, não foi analisado com a merecida atenção devido ao prazo para ser entregue o trabalho e por conta disso não foi colocado na composição principal desse modelo e sim na parte de aperfeiçoamento.

1. **Conclusões e Proposta**

Para a regressão, a média de nosso R² foi de 0,32, o que indica que nossas medidas variam em média, 0,32 para cima ou para baixo da medida real, o que pode ser considerado um bom resultado dado que nossas medidas variam de 0 a 5. Logo, caso quiséssemos prever se um livro terá uma determinada classificação numérica, poderíamos afirmar a classificação de nosso modelo com até 0,32 unidades de erro.

Ao final da análise classificatória concluiu-se que os resultados foram corretos em pouco mais que a metade dos testes (53,8%) porém, considerando o número de classificações (cinco no total), a predição mostrou-se razoável pois a chance de acertar a avaliação do livro apenas com um chute aleatório é de 20%. Portanto, nosso modelo preditivo pode ser considerado satisfatório.

Para que a experiência do usuário seja otimizada, propõe-se uma interface que roda o código do jupyter por meio de um programa no flask. Assim, a ideia é pegar as informações do livro e lançá-las no flask, e o programa retornará o rating previsto.

1. **Métricas**

A primeira técnica de verificação da qualidade do resultado utilizada foi a raiz quadrada da média do R2, de modo que a medida fica na mesma escala da variável em que o erro está sendo calculado, ou seja, consegue ter uma palpabilidade maior sobre o valor dessa medida.

A outra métrica foi a acurácia dos testes de classificação, calculando o número de acertos sobre o número de testes, em que é possível saber se é uma boa medida ou não, pois esse resultado é em porcentagem, ou seja, o modelo consegue prever corretamente o resultado em x% das vezes.

1. **Contribuições**

**Pedro** - Ajudou na construção do modelo baseado na técnica de regressão, limpou os dados e fez comentários sobre as partes produzidas nos arquivos Jupyter.

**Beatriz** - Auxiliou na construção do modelo baseado na técnica de classificação, contribuiu na formatação do arquivo Jupyter a ser entregue, produziu conclusões sobre o modelo classificatório e limpou os dados.

**Sophia** - Contribuiu na construção do modelo de regressão, documentou a trajetória do modelo em um documento de texto separado, e auxiliou na formatação do arquivo Jupyter a ser entregue.

**Fernando** - Fez sua contribuição auxiliando o grupo na divisão de tarefas, comentou o código, ajudou na documentação dos resultados e dos processos realizados para obter o modelo atual.