Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

Artigo de Ciência de Dados:
Prevendo preços de jogos baseado em configurações informadas no site Steam

SCC0275 - Introdução à Ciência de Dados
Profa. Roseli Aparecida Francelin Romero
PAE: Guilherme Vicentim Nardari

Fernando Akio Tutume de Salles Pucci nº 895719
Ewerton Patrick Silva do Amaral nº 10346975
Gabriel Citroni Uliana nº 9779367

São Carlos, 15 de Dezembro de 2020

I. Introdução

Para desenvolvimento de jogos, as empresas realizam investimentos em inúmeras tecnologias ou recursos que podem agregar valor, como por exemplo em dublagens, legendas ou recursos gráficos mais avançados. Esses investimentos são muito importantes, pois tornam o jogo cada vez mais imersivo e inclusivo, atraindo muitos jogadores e consequentemente impactando na venda dos jogos.

Embora o jogo se torne mais imersivo ou atrativo com tantos investimentos, por vezes esses investimentos acabam encarecendo o jogo pelos inúmeros recursos que agregam valor a ele. Os recursos envolvidos no jogo que pode encarecer-lo podem incluir, por exemplo, recursos computacionais e peças de computadores, dublagens oferecidas, plataformas suportadas para o jogo e entre outras informações. Fatores como estes influenciam na decisão de compra do jogo e consequentemente impactam no número de vendas, sendo pontos relevantes de estudo.

A plataforma da Steam fornece uma série de dados sobre jogos relacionados ao nosso problema, buscamos neste artigo explorar esse conjunto de informações com a intenção de prever os preços dos jogos e identificar quais recursos impactam diretamente no preço. Nosso estudo é direcionado para os jogadores e esperamos auxiliá-los a preverem o preço médio que pagarão por jogo a partir de informações, configurações ou recursos desejados. Com isso o jogador poderia decidir, por exemplo, se vale a pena comprar uma nova peça para o computador ou mesmo um novo computador, se vale a pena adquirir o jogo de uma empresa ou mesmo dar ao jogador o conhecimento de quais fatores/componentes que influenciam na redução ou aumento do preço de um jogo.

Nosso trabalho envolve a extração de dados a partir das informações dos jogos contidos na plataforma Steam, a realização do pré-processamento da base de dados, análise sobre as features que impactam o preço do jogo, a classificação e uma análise dos modelos testados.

II. Trabalhos Relacionados

O artigo de estudo sobre review de jogos onlines [1], analisa 79,437 reviews de usuários da Steam para 11 tipos de jogos diferentes, com o intuito de analisar quais fatores impactam na utilidade da classificação e recomendação de jogos, através de análises feitas pela comunidade da Steam. E em função desses dados, sugerir medidas que aumentem a usabilidade e lealdade dos usuários à própria comunidade. Os autores do estudo aplicaram *k-folds cross validation* com *10-folds*, para o treinamento do modelo e em seguida utilizaram *Classification and Regression Tree* e Redes Neurais artificiais para analisar seus dados.

O artigo de aplicações que aplica ciência de dados em jogos [2], realiza uma extensa análise de diversos estudos relacionados ao uso de Jogos Sérios, com o objetivo de entender o processo de aprendizado e melhorá-lo através de análises dos dados coletados pela interação dos estudantes com o ambiente de aprendizado dos jogos. Os autores observam que dentre os modelos de Ciências de Dados mais utilizados, se destacam os modelos lineares para modelos supervisionados, e correlação e clusterização para modelos não supervisionados.

Já o artigo aprendizado não supervisionado para comportamentos de jogos na plataforma Steam [3], utilizou como base de dados um *dataset* retirado da plataforma Steam com mais de 700 mil jogadores considerados *hardcore,* cobrindo mais de 3300 jogos. Aplicando a clusterização de dados com o *K-means,* o estudo identificou os perfis desses jogadores e como os jogos são percebidos por esses usuários.

O artigo de sistema de recomendação híbrida para Steam [4], aborda o sistema de recomendação de jogos para na plataforma de vendas da Steam. O estudo propõe uma recomendação híbrida, que produz sugestões a partir de dados como tempo de jogo, preço e data de lançamento, em adição à já utilizada recomendação por preferências de gêneros e baseada em jogos comprados por seus amigos.

O artigo que determina preço baseado em Análise de Componente Principal [5], aplica *Principal Components Analysis*(PCA) com às Redes Neurais Artificiais, foi utilizado para determinar o preço de imóveis. Nosso trabalho tomou esse artigo como referência de aplicação do PCA para encontrar relações de diversos dados com o preço sendo a informação alvo.

Nosso trabalho se difere no *dataset* escolhido, apesar de artigos anteriores já utilizarem dados obtidos da plataforma da Steam, eles focam na interação dos usuários com os jogos, ou com o sistema de sugestão de produtos. O nosso *dataset* prioriza as especificações do jogo em si, como preço, número de linguagens, hardware recomendado, data de lançamento, entre outros. As únicas informações de interação do jogador com o jogo que colhemos são o número de reviews e a média desses reviews.

Além disso, nosso artigo explora diversos modelos para identificar correlações existentes em um grande conjunto de variáveis, o que potencialmente pode ser benéfico para diversas outras aplicações similares, tanto para área de jogos quanto para outras áreas que possuem grandes variáveis envolvidas.

III. Materiais e Métodos

Dataset

A Steam é uma plataforma que busca disponibilizar jogos para PCs, alguns jogos distribuídos gratuitamente e outros são vendidos. Nela é incluído uma série de informações relevantes sobre o jogo como, por exemplo, configurações recomendadas, dublagens disponibilizadas, *reviews* e entre outras coisas. Decidimos explorar a Steam pela popularidade, estabilidade da plataforma e pela padronização das tags da página, o que auxiliou na obtenção de alguns recursos.

De maneira geral, o nosso dataset apresenta os seguintes componentes para análise do preço: nome do jogo, desenvolvedora, distribuidora, Indie (são jogos desenvolvidos por uma pessoa ou uma pequena equipe independente, com ou sem investimento), quantidade de review, review médio, plataforma singleplayer, multiplayer, data de lançamento, número de linguagens Interface, número de linguagens de dublagem, número de linguagens de legenda, memória ram, espaço em disco, placa de vídeo, processador e preço.

Extração dos dados

Para extração dos dados na Steam, foi realizado *web scraping* utilizando a biblioteca *Beautiful Soup* do Python. A plataforma Steam tem algumas restrições de acesso, como o número de requisições por minutos e verificação de idade para acessar alguns jogos, o que demandou mais pesquisas e estudos para contornar tais problemas. Como solução utilizamos *cookies* na requisição da página para contornar a verificação de idade, já para não infringir o número de acessos por minutos da Steam, incluímos um delay de requisição que varia por cada requisição feita. No final, conseguimos extrair 23.184 jogos da plataforma da Steam.

Na Steam, existe uma certa variabilidade na forma de escrita de alguns dados (como por exemplo a parte de configurações recomendadas), isso gerou dificuldades na obtenção dos dados e tivemos que explorar outros recursos para contornar tais problemas. Exploramos a biblioteca de expressão regular [6] para identificar padrões mais pertinentes para nosso problema, sendo definido um escopo na expressão regular para extrair as informações.

Assim, da plataforma foram extraídas apenas as placas de vídeo da NVIDIA dos modelos GTX e RTX, já com relação aos processadores foram extraídos apenas os da Intel. Embora essa configuração não explore todos os componentes disponíveis no site, ela permitirá identificar claramente qual a influência que cada componente pode ter sobre o preço do jogo. Além disso, delimitar os componentes que serão explorados no dataset permitirá que os jogadores busquem os componentes com desempenho similar ao desejado ou estabelecido na base.

Com relação aos suportes para linguagem, disponibilizados pelos jogos da Steam, foi escolhido obter o número total de linguagens suportadas por cada jogo, em 3 características distintas, a Interface, a Dublagem e as Legendas.

Os preços foram extraídos do valor inteiro do jogo em Reais, ou seja, sem desconto algum. Jogos que não cobram nada, foram considerados Free To Play. Por fim, jogos que não foram lançados ainda ou que não possuem uma forma de comprar, foram considerados Indisponíveis.

Durante a extração dos preços tivemos outro problema relacionados aos *cookies* do site da Steam, a plataforma reconhece o país de onde provêm as requisições, através de *cookies*, isso gerou problemas com a utilização do notebook do Google Colaboratory, pois seu servidor mudava de região com o tempo, podendo produzir valores em diferentes moedas. Para contornar esse problema optamos por realizar as requisições em uma máquina local para montar o *Dataset*.

Exploração e Pré-processamento

Determinadas empresas se destacam pela qualidade, estilo de jogo ou por outros quesitos, o que impacta na popularidade do jogo. Queríamos verificar se existe uma correlação dessa informação com o preço, mas como existe uma alta quantidade de empresas que desenvolvem jogos, a análise desse dado se tornaria difícil. Assim, decidimos fazer um pré-processamento com os rankings de empresas que mais aparecem no dataset, podendo assim analisar melhor uma possível influência desse valor categórico sobre o preço.

No dataset, decidimos pré-processar os dados de data de lançamento, podendo assim transformar o valor categórico em numérico. O valor numérico pré-processado

representa a quantidade de anos em que o jogo foi lançado, podendo assim identificar uma correlação com o preço.

Escolhemos eliminar jogos com números de reviews muito baixos, pois identificamos que no geral eles possuem poucos jogadores, ou seja, são de baixa qualidade e pouco relevantes para nosso problema. Além disso, foram excluídos do *dataset* as instâncias que não apresentavam nenhuma informação do processador ou da placa de vídeo dentro das configurações recomendadas, pois posteriormente analisou-se que a ausência desses valores impactavam negativamente no treinamento.

Para trabalharmos com os variados preços, optamos por retirar jogos gratuitos e pré-processar os demais jogos em 3 categorias de preços diferentes, sendo divididos em preços: baixos, médios e altos. Os preços baixos incluem valores abaixo de R$18, o preço alto envolve jogos acima de R$45 e o preço médio se encontra entre essas duas faixas de valores. Essas categorias se encontram desbalanceadas, como demonstra a Figura 1.

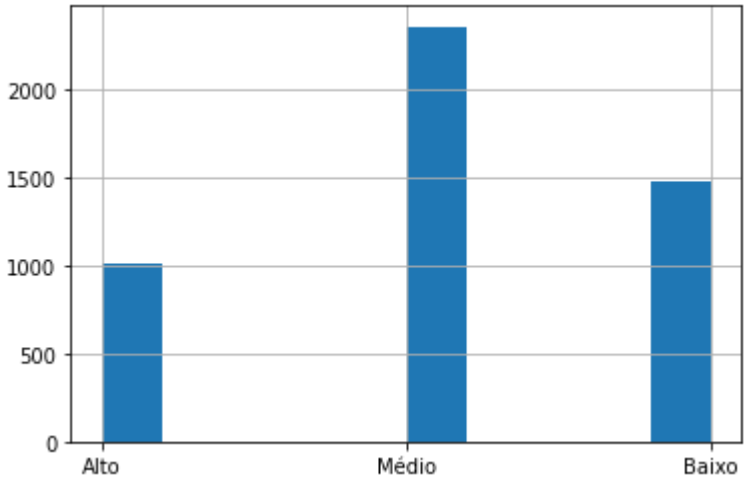

Figura 1: Distribuição desbalanceada dos preços dos jogos da Steam após aplicar o pré-processamento.

Para fazer o balanceamento dos dados aplicamos a combinação de duas técnicas de subamostragem, a *undersampling* e a *oversampling*. Aplicando a combinação das duas técnicas, evitamos perder uma quantidade significativa de jogos que ocorreria se aplicassemos apenas a técnica de subamostragem *undersampling*, e evitamos também repetir muitos valores da base ao utilizar apenas a técnicas de subamostragem *oversampling.* A aplicação da combinação das duas técnicas é apresentada na Figura 2.

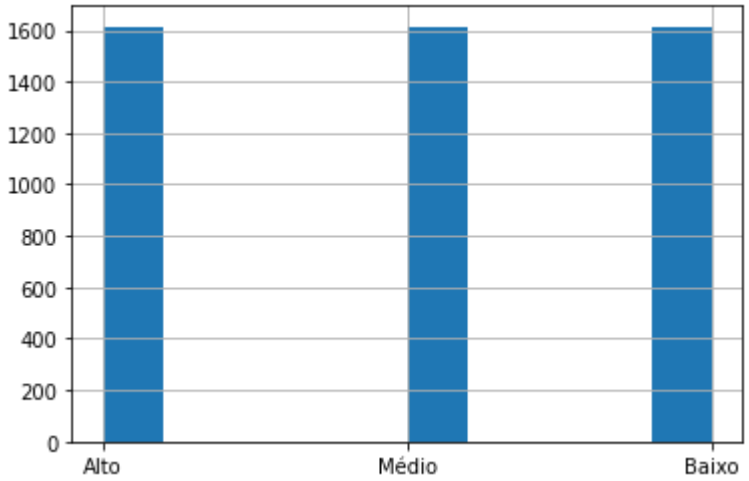

Figura 2: Distribuição da base de dados.

Feature Selection

Primeiramente, nós escolhemos previamente, durante o *scraping* dos dados, os atributos que acreditávamos possuir alguma relevância para montar o *dataset*. Para processarmos a base de dados, retiramos o atributo *Nomes* dos jogos e elencamos os os atributos mais pertinentes para a inferência dos preços dos jogos através da matriz de correlação.

Modelos de Classificação ou Regressão

Utilizamos os modelos da biblioteca sklearn [7] da linguagem python para classificação dos preços, sendo os modelos:

- Perceptron: Biblioteca sklearn.linear_model.Perceptron
- Multi-Layer Perceptron: Biblioteca sklearn.neural_network.MLPClassifier
- SVM Polinomial: Biblioteca sklearn.svm.SVC
- Árvore de Decisão: Biblioteca sklearn.tree.DecisionTreeClassifier
- KNN: Biblioteca sklearn.neighbors.KNeighborsClassifier

Tendo definido as biblioteca, os parâmetros incluídos no código para os modelos foram:

```
classificadores = {
 "Perceptron":{"modelo":Perceptron(),"scores":[]},
 "Multi-Layer Perceptron(15,)":
    {"modelo": MLPClassifier(random_state=1, hidden_layer_sizes=(15,), max_iter=2000),
"scores": []},
 "SVM Polinomial Grau 3": {"modelo":SVC(kernel='poly',degree=3,gamma=1), "scores": []},
 "Árvore Decisão Critério Gini": {"modelo":DecisionTreeClassifier(criterion='gini'),"scores": []},
 "KNN k=5":{"modelo": KNeighborsClassifier(n_neighbors=5), "scores": []}
}
```

IV. Experimento

A partir do *dataset* balanceado, foi gerado a sua matriz de correlação. Um desafio enfrentado foi produzir essa análise utilizando todos os atributos presentes no banco, uma vez que os atributos categóricos *Desenvolvedor, Distribuidora*, *Processador* e *Placa de Vídeo* possuíam múltiplos valores, tornando inviável a análise da correlação com o *heatmap* gerado. Visto isso, obteve-se essa métrica a partir dos dados numéricos e também dos dados categóricos de poucos valores. Analisando a matriz de correlação gerada, é possível notar uma relação significativa de *Preço Alto* com os atributos *Espaço em Disco, Quantidade de Review, Memória Ram e Indie*.

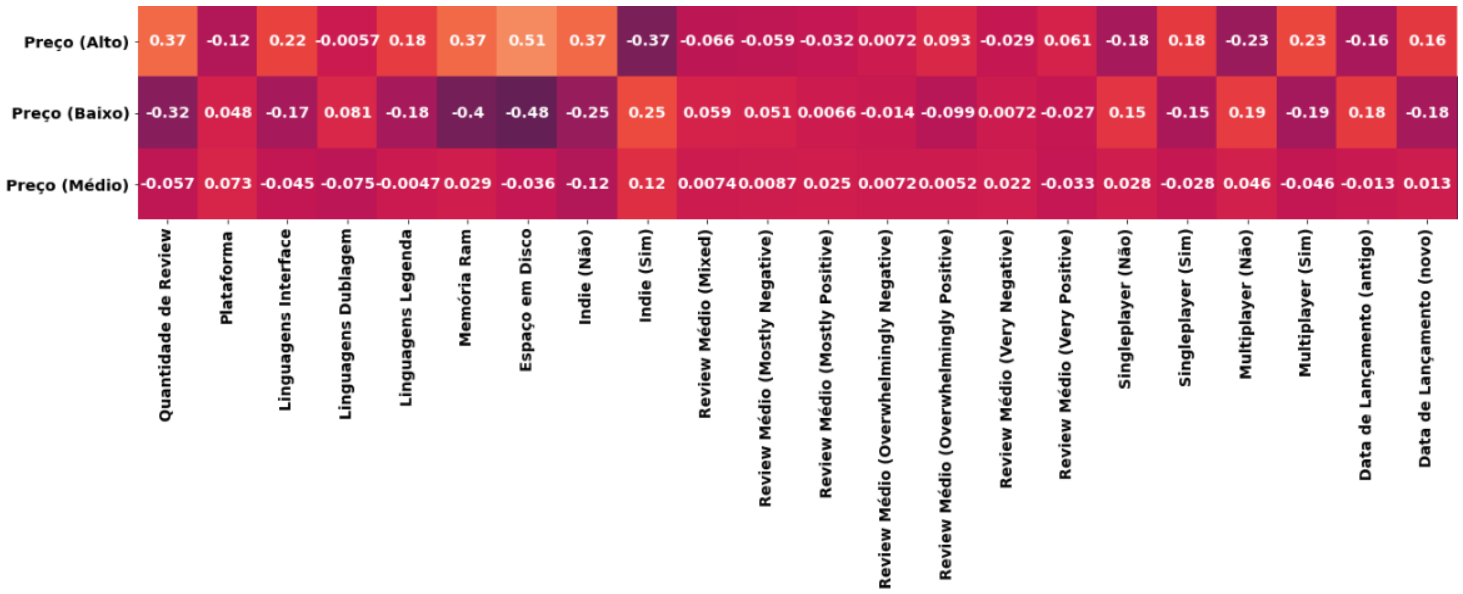

Figura 3: Matriz de Correlações.

Dessa forma foram elencados os doze atributos mais relevantes em relação ao atributo *Preço Alto*. O banco de dados para o treinamento e teste então foi definido como a coleção desses dados extraídos da matriz de correlação, concatenados com dados categóricos anteriormente excluídos.

Atributos	Correlação
Espaço em Disco	0.514446
Quantidade de Review	0.372341
Memória Ram	0.369799
Indie (Sim)	0.368666
Indie (Não)	0.368666
Multiplayer (Sim)	0.231524
Multiplayer (Não)	0.231524
Linguagens Interface	0.217450
Linguagens Legenda	0.182896
Singleplayer (Sim)	0.176644
Singleplayer (Não)	0.176644
Data de Lançamento (novo)	0.164457
Data de Lançamento (antigo)	0.164457

Tabela 1: Atributos com maior índice de correlação para Preço Alto.

Durante o processo de treinamento, foram realizados diversos experimentos com diferentes configurações. Um dos testes realizados foi a aplicação da Análise de Componentes Principais (PCA) com diferentes números de componentes e analisou-se que a pontuação utilizando o *dataset* de dimensionalidade reduzida se mostrou menor do que a do *dataset* padrão. O grupo acredita que esse resultado se deve à grande quantidade de

atributos gerados durante a transformação dos dados categóricos em numéricos; por conta disso, ao reduzir a dimensionalidade do *dataset* há uma perda de informações potencialmente relevantes ao modelo. Outras configurações testadas incluem: a presença de instância com atributos *Processador* ou *Placa de Vídeo* nulos, número de componentes do PCA, utilização de instâncias com o atributo numérico *Review* baixo, *dataset* sem os atributos categóricos *Distribuidora, Desenvolvedor, Processador* e *Placa de Vídeo* e *dataset* sem utilização do PCA. A utilização do *dataset* tratado descrito, sem a utilização do processo de redução de dimensionalidade demonstrou melhor desempenho nos classificadores em geral.

Experimentos	Classificador (%)				
	Perceptron	MLP	SVM Poly Grau 3	Árvore de Decisão	KNN k=5
Sem PCA	55.71	72.95	72.73	77.68	67.37
PCA n_c = 90%	48.44	63.47	68.83	72.94	62.83
PCA n_c = 98%	56.26	71.52	73.16	75.84	64.29
Sem PCA Poucos reviews	49.49	71.50	69.75	71.09	67.28
Sem PCA Sem Attr Desenvolvedor Sem Attr Distribuidora Sem Attr Placa de Vídeo Sem Attr Processado	50.10	57.49	57.71	72.74	62.64
Sem PCA Com valor null em Processador Com valor null em Placa de vídeo	51.37	62.10	62.79	66.74	62.25

Tabela 2: Score dos classificadores com diversos experimentos.

O melhor modelo para esse problema foi a árvore de decisão. O conjunto de dados obtidos para a análise do jogo na Steam apresenta uma grande quantidade de atributos e a não linearidade do problema se refletiu na acurácia do modelo de árvore de decisão, pois a árvore de decisão mapeia muito bem relações não-lineares.

V. Conclusões

Os investimentos nos jogos costumam torná-lo mais imersivo e atrativo, mas por vezes esses investimentos acabam encarecendo o jogo pelos inúmeros recursos que agregam valor a ele. Nosso artigo envolve a análise de atributos de jogos da plataforma Steam para identificar quais informações impactam diretamente no preço do jogo, buscando auxiliar os jogadores na decisão de compra de jogos ou quais peças comportam os jogos desejados.

Buscamos explorar atributos voltados tanto para configuração recomendada, recursos disponibilizados e informações que potencialmente seriam relevantes no preço do

jogo, como por exemplo data de lançamento. Surgiram inúmeras dificuldades na extração dos dados da plataforma da Steam, mas que foram contornadas ao longo do projeto.

Seguindo os métodos estabelecidos para o experimento, os três atributos que mais demonstraram estar relacionados ao preço do jogo foram: Espaço em disco, a quantidade de Review, quantidade de Memória Ram e se o jogo é Indie.

Devido a grande quantidade de atributos analisados na plataforma, o problema se tornou não linear e se o resultado se refletiu na acurácia do modelo de árvore de decisão, sendo que a árvore de decisão mapeia muito bem relações não-lineares. Esse resultado sugere mais estudos sobre outros modelos não lineares que potencialmente descreveriam melhor esse problema, como por exemplo modelos *random forest* e *gradient boosting*.

VI. Referências

[1] Ha-Na Kang, Hye-Ryeon Yong, and Hyun-Seok Hwang, " A Study of Analyzing on Online Game Reviews using a Data Mining Approach: STEAM Community Data," International Journal of Innovation, Management and Technology vol. 8, no. 2, pp. 90-94, 2017.

[2] Alonso-Fernandez, Cristina & Calvo-Morata, Antonio & Freire, Manuel & Martinez-Ortiz, Ivan & Fernández-Manjón, Baltasar. (2019). Applications of data science to game learning analytics data: A systematic literature review. Computers & Education. 141. 10.1016/j.compedu.2019.103612.

[3] Baumann, Florian & Emmert, Dominik & Baumgartl, Hermann & Buettner, Ricardo. (2018). Hardcore Gamer Profiling: Results from an unsupervised learning approach to playing behavior on the Steam platform. Procedia Computer Science. 126. 1289-1297. 10.1016/j.procs.2018.08.078.

[4] Gong, Jin & Ye, Yizhou & Stefanidis, Kostas. (2020). A Hybrid Recommender System for Steam Games. 10.1007/978-3-030-44900-1_9.

[5] Shi, Huawang. (2009). Determination of Real Estate Price Based on Principal Component Analysis and Artificial Neural Networks. Intelligent Computation Technology and Automation, International Conference on. 1. 314-317. 10.1109/ICICTA.2009.83.

[6] PYTHON. Regular expression operations. 2020? Disponível em: <https://docs.python.org/3/library/re.html>. Acesso em: 14 dez. 2020.

[7] SCIKIT LEARN. Scikit-learn Machine Learning in Python. 2020? Disponível em: <https://scikit-learn.org/stable/index.html>. Acesso em: 12 dez. 2020.