Como avaliar desempenho de algoritmos de Machine Learning para predição de valores

Wellington Gallo, Leanderson André

Bacharelado em Sistemas de Informação   
Universidade da Região de Joinville (Univille) Joinville – SC – Brasil

(wellington.gallo, leanderson.andre)@univille.br

**Abstract.** This article reports the results of performance appraisals of Machine Learning algorithms, using linear regression, to predict values ​​of al purchased in the region of São Paulo. The data dataset is available on Kaggle's website for further queries. With the research completed, it is possible to fulfill that, to use regression algorithms, the correlation between the variables needs to be between -1 and 1, in addition to identifying possible improvements in the data, and to analyze the best available algorithm for the proposed data.

**Resumo.** Este artigo relata os resultados das avaliações de desempenho de algoritmos de Machine Learning, utilizando regressão linear, para prever valores de aluguéis na região de São Paulo. O dataset dos dados está disponível no site da Kaggle para demais consultas. Com a pesquisa finalizada, é possível concluir que, para utilizar algoritmos de regressão, a correlação entre as variáveis precisa ser entre -1 e 1, além de identificar possíveis melhorias nos dados, e analisar o melhor algoritmo disponível para os dados propostos.

**Palavras chaves:** Regressão linear, Machine Learning, aprendizado de máquina, predição de valores

**Introdução**

A inteligência artificial é um ramo de pesquisa da ciência da computação que desenvolve mecanismos e dispositivos tecnológicos para poder simular o raciocínio humano. De acordo com Lemes (2018), o objetivo da IA (Inteligência Artificial) é desenvolver computadores que consigam pensar, bem como, ver, ouvir, andar, falar e sentir.

Hoje a inteligência artificial aborda diferentes áreas do conhecimento, podendo ser útil em vários estudos e descobertas cientificas, pela forma que os algoritmos aprendem e absorvem a informação.

Existem vários exemplos que demonstram como a inteligência artificial pode encontrar soluções e realizar previsões de modo a ajudar diferentes áreas de alguma forma. A inteligência artificial possui diversos mecanismos para identificar padrões, e isso possibilita rápidos aprendizados para que seja possível identificar a melhor forma para solução de algum problema.

Os desafios da escolha de um imóvel para locação são enormes devido à grande diversidade de variáveis nas características deste imóvel, tais quais são a localização, tempo de construção, quantidade de quartos, andar etc. Os pontos a serem considerados em uma negociação são os mais variados possíveis, e variam de imóvel para imóvel.

Geralmente para alugar um imóvel, se esse for o caso, é preciso ter um fiador, para que o locatário tenha uma garantia que o pagamento total do contrato será realizado. Além disso, os valores de taxas precisam ser negociados, pois há muita variação dessas taxas que impactam diretamente no valor total em negociação.

Atualmente, o mercado imobiliário está “aquecido”, pelo fato de muitas pessoas estarem se mudando para locais mais seguros ou mais agradáveis por questões pessoais. No entanto, é preciso identificar tudo o que há disponível num imóvel antes de tomar alguma decisão, e para validar se uma mudança faz sentido, é necessário avaliar alguns pontos, tais qual sua localização, além de identificar se o valor total, incluindo todas as taxas, cabem no bolso.

Um trabalho publicado por Thiago Alves e Rodrigues, com o título, avaliação de residências usando modelos de regressão linear, apresenta verificações estatísticas e normativas para a validação do modelo de regressão, inclusive com a análise de sensibilidade dele. O trabalho apresentou o resultado várias simulações, 4 variáveis independentes explicaram estatisticamente a variável dependente valor total: área construída, área do terreno, número de quartos e padrão construtivo. Obteve-se a equação de regressão com nível de significância de 0,01%, apresentando correlação no valor de 86,96%, o que significa uma forte relação entre a variável dependente e as variáveis independentes utilizadas no modelo. O coeficiente de determinação encontrado foi de 75,63%, o que significa dizer que 24,37% do valor total não foi explicado pelo modelo de regressão por prováveis variáveis não consideradas ou erros ocasionais de medidas.

Este artigo tem como objetivo realizar um experimento com algoritmos de *Machine Learning* para predição dos valores de aluguéis de imóveis.

A partir deste experimento, será possível identificar os algoritmos que tiveram melhor desempenho com os dados coletados, além de identificar como o desempenho pode ser melhorado.

O artigo está estruturado em 5 (cinco) etapas. Após a introdução, na segunda etapa é apresentada a revisão da literatura explicando os conceitos de *Machine Learning*, Inteligência artificial, Aprendizado supervisionado, regressão linear e *webscraping*. Na terceira etapa são expostos os procedimentos metodológicos utilizados para realização da pesquisa. Na quarta etapa é abordada a análise dos dados e discussões dos resultados. Na quinta etapa são feitas as considerações finais, seguidas pelas referências bibliográficas.

**2. Revisão da literatura**

Nesta etapa serão abordados os principais tópicos relacionados ao tema de pesquisa, onde são explicados os métodos utilizados, com suporte de algumas referências bibliográficas e científicas, para que seja possível dar andamento ao trabalho.

**2.1 Aprendizado de Máquina (*Machine Learning – ML*)**

O aprendizado de máquina (Machine Learning – ML) é um subcampo da computação que evoluiu o estudo do reconhecimento de padrões na inteligência artificial.

Em 1995, Arthur Samuel definiu aprendizado de máquina (Machine Learning – ML) como o campo de estudo que dá os computadores a habilidade de aprender sem serem programados. Segundo GÉRON (2019, p.4) Aprendizado de Máquina é a ciência da programação de computadores para que eles possam aprender com dados, ou seja, isso faz com que algoritmos de ML (Machine Learning) tenham maior facilidade em resolver problemas variados com base nas informações que o possui. Quanto mais informações, mais padrões a máquina poderá identificar.

Existem alguns tipos de aprendizagem para ser utilizado, como o aprendizado não supervisionado, que serve para quando é necessário identificar padrões sem possuir massas de dados, segundo Honda (2017) exemplos de aplicações de aprendizado não supervisionados são sistemas de recomendação de filmes ou músicas, detecção de anomalias e visualização de dados. Já aprendizagem semi-supervisionada, consiste em uma metodologia de aprendizado no meio do caminho entre a aprendizagem supervisionada e a não supervisionada. Um grande exemplo de como é aplicado o aprendizado semi-supervisionado é o serviço de hospedagem e fotos do Google, segundo Géron (2019) ao carregar todas as suas fotos de família, o aplicativo reconhecerá automaticamente que a mesma pessoa (A) aparece nas fotos, 1, 5 e 11 enquanto outra pessoa (B) aparece nas fotos 2,5 e 7. Com apenas um rótulo por pessoa ele será capaz de nomear todos.

**2.2 Aprendizado supervisionado**

Para realizar o aprendizado supervisionado, é necessário ensinar à máquina com uma massa de dados. Assim, com base nestes dados, ela lhe trará resultados ou soluções esperadas. Segundo Massaron (2019), a estratégia supervisionada é semelhante à aprendizagem humana sob supervisão de um professor, ou seja, tudo é realizado com base em exemplos e demonstrações, até que o aluno memorize o assunto que está sendo abordado em aula.

O aprendizado supervisionado e as técnicas citadas são as que contêm maior precisão e resultados de aplicações bem-sucedidas, pelo fato de os problemas estarem bem definidos, por exemplo, ao tentarmos prever uma variável dependente a partir de uma lista de variáveis independentes, conforme será exemplificado na tabela 1:

Tabela 1: Exemplo de aprendizado supervisionado com variáveis fixas

|  |  |
| --- | --- |
| **Var. Independentes** | **Var. Dependentes** |
| Anos de Carreira, Formação, Idade | Salário |
| Idade do Carro, Idade do Motorista | Risco de Acidente Automotivo |
| Texto de um livro | Escola Literária |
| Temperatura | Receita de venda de sorvete |
| Imagem da Rodovia | Ângulo da direção de um carro autônomo |
| Histórico escolar | nota no ENEM |

Fonte: Honda; Facure (2019)

**2.3 Regressão linear**

A regressão linear é um dos conceitos mais utilizados dentro do *Machine Learning*, sendoela é denominada por uma linha reta traçada, a partir de um diagrama de dispersão. Segundo Pierson (2019, p. 73), regressão linear é um método de aprendizagem de máquina que você pode usar para descrever e quantificar a relação entre sua variável alvo “y”, com a variável dependente.

Um exemplo simples de algoritmos de *Machine Learning* utilizando a regressão linear é no setor imobiliário, em que há a previsão de valores de locação imóveis; ou até mesmo no setor de vendas, onde é possível prever valores de algum determinado produto. Há uma série de fatores, que determinam uma regressão linear, que segundo Pierson (2019, p. 73) são:

1. a regressão linear funciona apenas com variáveis numéricas, não categóricas;
2. se houver valores ausentes em seu conjunto de dados, isso causará problemas. Resolva a questão dos valores ausentes antes de tentar criar um modelo de regressão linear;
3. se seus dados contiverem valores atípicos, seu modelo produzirá resultados impressos. Verifique antes de continuar;
4. a regressão linear supõe que há uma relação linear entre os recursos do conjunto de dados e a variável-alvo. Confira se este é o caso, e se não for, use uma transformação logarítmica para compensar;
5. o modelo de regressão linear supõe que todos os recursos são independentes entre si;
6. os erros de previsão, ou residuais, devem ser distribuídos normalmente.

Ainda, é possível utilizá-lo em cálculos de expectativa de vida em um país, ou até mesmo calcular pressão sanguínea de algum paciente. Estes são alguns exemplos nos quais podem ser aplicados a regressão linear. Segundo Damaceno (2020), esse tipo de algoritmo é aplicado quando há uma boa correlação linear (positiva ou negativa) entre os dados, isto é, quando o relacionamento ou associação entre os dados pode ser definido com uma reta, conforme mostra a figura 1.

Figura 1: Gráfico representando regressão linear

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Fonte: Datageeks (2019)

**2.4 Avaliação dos algoritmos**

Para avaliação de modelos utilizando algoritmos de regressão, é possível utilizar as métricas de avaliação, neste caso pode ser utilizado as medidas mais conhecidas, que são Erro Quadrático Médio ou MSE (Mean Squared Error), Erro Médio Absoluto ou MAE (Mean Absolute Error) e Raiz quadrada do erro medido absoluto ou RMSE (Root Mean Squared Error), segundo Sampaio, Bernardini, Paes, Andrade e Viterbo (2019) As métricas MSE, MAE e RMSE devem sempre apresentar valores não negativos. Quanto menor o valor dessas métricas, melhor o modelo será.

Além das métricas de erro, é possível avaliar modelos de *Machine Learning* através de coeficiente de determinação (R²), que realiza uma análise de regressão no modelo e verifica o quão ajustado o modelo está em relação a regressão. Segundo Souza (2019 o cálculo do coeficiente de determinação envolve três medidas: Soma total dos quadrados (STQ), Soma dos quadrados dos resíduos (SQU), e soma dos quadrados de regressão SQR, o valor de R² é o próprio SQR dividido pelo SQT, ou seja, é divisão da variação explicada pela variação total dos dados.

# 3. Procedimentos metodológicos

Nesta etapa são apresentados todos os passos para análise e aplicação do conjunto de dados dos imóveis para aluguel na região de São Paulo; desde os procedimentos metodológicos, como o passo a passo para obtenção dos dados, até a aplicação dos algoritmos de Machine Learning.

Para que o processo de aprendizagem de máquina fosse eficiente, foi necessário aplicar os dados de maneira correta. Para isso, foi utilizado o fluxo de KDD (knowledge-discovery in databases), que é composto por cinco etapas: seleção de dados, pré-processamento, transformação, mineração e interpretação e avaliação, este tipo de processo pode ser repetido inúmeras vezes desde que necessário, para que tenha os melhores resultados. A base de dados original está disponível para download no Kaggle, e contém informações de várias cidades do Brasil, contendo 10693 linhas, e 14 colunas, sendo elas. Para este trabalho, foram filtrados os dados utilizando apenas a cidade de São Paulo como referência, e utilizado apenas dados que fizessem sentido com os algoritmos de regressão, que foram dados numéricos, além disso foram coletadas 2853 linhas, cada uma representando características dos imóveis, como, valor total, área, número de quartos, número de banheiros, quantidade de vagas de garagem e andar.

Após o processo de coleta de dados, foi realizada uma pesquisa para identificar as possíveis formas de identificar os melhores algoritmos de regressão linear para base de dados proposta, e chegou-se à conclusão, que com a biblioteca PyCaret era possível gerar uma relação dos algoritmos, identificando a melhor performance através de três medidas de erros, que são: MAE, MSE e RMSE. Cada uma destas medidas, fazem cálculos de erro no modelo para identificar possíveis falhas nos dados coletados. A partir destes resultados, passamos a analisar os melhores algoritmos para este modelo, conforme sua medida de erro, e para exemplificação, foram utilizados o pior e o melhor cenário proposto pelo PyCaret para discussão dos resultados.

A partir de toda análise dos, foi desenvolvido algoritmos de *Machine Learning* com a aprendizagem supervisionada de regressão, e separado o modelo em dados de treino e teste. Os dados de treino costumam representar 70% da totalidade dos dados, enquanto os dados de teste 30%, o modelo defini a aleatoriedade dos dados utilizando funções prontas do python, para que isso se torne mais rápido e faça com que o modelo realmente aprenda com os dados.

# 4. Análise dos dados e discussão dos resultados

Os resultados deste estudo se iniciam pela análise das variáveis do modelo proposto. Neste caso, foi utilizado o coeficiente de correlação de Pearson, que mede o grau de correlação entre duas variáveis. Este coeficiente é apresentado por valores entre 1 e -1, esses valores são representados pelas descrições da tabela abaixo:

Tabela 1 – Valores do coeficiente de correlação

|  |  |
| --- | --- |
| Correlação | Interpretação |
| 0,00 a 0,19 ou 0,00 a -0,19 | Correlação bem fraca |
| 0,20 a 0,39 ou -0,20 a -0,39 | Correlação fraca |
| 0,40 a 0,69 ou -0,40 a -0,69 | Correlação moderada |
| 0,70 a 0,89 a -0,70 a -0,89 | Correlação forte |
| 0,90 a 1,00 ou -0,90 a -1,00 | Correlação muito forte |

Fonte: (O autor, 2021)

Como mostrado na tabela acima, para que uma base de dados, possa ser considerada relevante para utilizar algoritmos de regressão, é necessário que a medida da correlação de Pearson fique entre 0,40 ou -0,40 a 1,00 ou -1,00. Esta medida representa que as variáveis dependentes possuem uma boa correlação. Esses valores, positivos e negativos, significam a seguinte situação:

1. Correlação negativa - enquanto uma variável aumenta a outra diminui
2. Próximo de 1 significa que, quando uma variável aumenta, a outra aumenta também.

Na base de dados utilizada para o estudo, a correlação entre as variáveis, teve um resultado positivo, isto avaliando a variável de predição que é o valor total, com as características que são, área, quantidade de quartos, quantidade de banheiros, quantidade e vagas de garagem, conforme ilustra a figura 2, concluindo-se que é possível utilizar algoritmos de regressão.

Figura 2 – Matriz de correlação

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Fonte: (O autor, 2021)

Para exemplificação dos resultados, será mostrado os gráficos do melhor e do pior desempenho, como mostra a figura 3. Nos algoritmos testados, foram utilizados 30% dos dados para teste e os demais para treinamento, de uma base com 2853 linhas.

Conforme mostrado na tabela de correlação, foi visível que a base de dados tem características adequadas para utilização de algoritmos de regressão. Na figura 3, é apresentado diversos algoritmos com seus desempenhos a partir de métricas de erros, neste caso, será avaliado o algoritmo com a melhor e pior performance segundo a biblioteca PyCaret. Os algoritmos apresentados serão, *Least Angle Regression* (LARS) que realiza seus cálculos de predição a partir da característica que mais se correlaciona com o dado previsor, e que de acordo com as métricas de erro, teve seu pior desempenho em relação aos outros, e o *Linear Regression* que tenta encontrar a reta que mais se adequa com seus pontos no gráfico.

Conforme citado acima, os algoritmos que serão comparados, vão ser apenas o de melhor e pior desempenho, os demais apresentados na figura 3, não devem ser considerados.

Figura 3 – Relação de performance de algoritmos de regressão

Texto

Descrição gerada automaticamente

Fonte: (O autor, 2021)

Ambos os algoritmos, mostram um resultado de regressão muito bom, sendo difícil identificar o modelo que teve a melhor performance, mesmo que com a biblioteca PyCaret podemos ter as métricas de erro MAE, MSE e RMSE, estes mesmos erros são muito próximos um dos outros, e não garantem que algum algoritmo teve um desempenho melhor do que o outro. A partir disto, foi utilizado o coeficiente de determinação (R2), que é uma medida de ajuste do modelo, o R2 valida a diferença dos modelos em relação a média e normaliza o valor para uma escala que varia entre 0 e 1, quanto mais próximo do 1, mais performático o modelo está em relação ao problema proposto.

Figura 4 – Algoritmo de treino e teste com *Least Angle Regression*

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Fonte: (O autor, 2021)

Figura 5 – Algoritmo de treino e teste com *Linear Regression*

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Fonte: (O autor, 2021)

Olhando os gráficos, a diferença entre os resultados é mínima por isto, vamos identificar o melhor desempenho através do coeficiente de determinação, como citado acima. Para este caso, é possível definir que o algoritmo de Regressão linear teve o melhor desempenho, pois o seu coeficiente ficou em 0.48 mais próximo de 1 do que o algoritmo de LARS, que teve seu coeficiente definido como 0.25. Estes cálculos foram realizados com a biblioteca *sklearn* do python, utilizando o método de *metrics.r2score* que recebe como parâmetro o valor real, e o valor previsto.

# 5. Conclusão

Para realizar predição de valores, são utilizados algoritmos de regressão linear, porém, é necessário tomar muito cuidado com a base de dados e os parâmetros de cada algoritmo, isso porque, nem toda base terá dados suficientes para que seja possível uma correlação linear adequada, caso isso ocorra, é necessário que os dados sejam trabalhados utilizando alguma normalização, ou até mesmo buscando mais informações sobre o assunto.

No caso desta pesquisa, podemos identificar que a correlação dos dados era extremamente boa, sendo possível realizar a predição dos valores. Conforme a relação da performance dos algoritmos na figura 3, é possível visualizar que até mesmo o algoritmo que foi considerado o pior em performance, superou as expectativas, e conseguiu trazer um resultado adequado, conforme mostra a figura 7.

Durante o desenvolvimento dos algoritmos, foi trabalho os parâmetros dos dados, para que o resultado fosse satisfatório, e como identificado em parágrafos acima, conclui-se que o melhor desempenho para esta base de dados, é utilizando o algoritmo *Linear Regression*. Reforçando que, apenas para este modelo apresentado, foi possível extrair os resultados citados acima, cada modelo age de maneira diferente, podendo gerar outros resultados ou até mesmo, nem ser possível utilizar algoritmos de regressão.

# Referências

LEMES, Giovanni Bugni. **INTRODUÇÃO A INTELIGENCIA ARTIFICIAL**. Joinville: Clube de Autores, 2018.

SAMUEL, AL (1959). **ALGUNS ESTUDOS EM APRENDIZADO DE MÁQUINA USANDO O JOGO DE DAMAS.** IBM Journal of research and development

GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**: conceitos, ferramentas e técnicas para a construção de sistemas inteligentes. Rio de Janeiro: Alta Books, 2019.

MUELLER, John; MASSARON, Luca. **Aprendizado de Máquina Para Leigos**. Rio de Janeiro: Editora Alta Book, 2019.

PIERSON, Lillian. **Data Science Para Leigos**. 2. ed. Rio de Janeiro: Editora Alta Books, 2019.

DAMACENO, Laura. **Regressão Linear?** 2020. Disponível em: https://medium.com @lauradamaceno/regressão-linear-6a7f247c3e29. Acesso em: 26 ago. 2020.

RODRIGUES, Gutelvam. **Métricas para avaliação de Modelos de Regressão**: variáveis contínuas/numéricas. Variáveis contínuas/numéricas. 2020. Disponível em: https://gutto-rdj.medium.com/ métricas-para-avaliação-de-modelos-de-regressão-variáveis-contínuas-numéricas-41d240375035. Acesso em: 13 maio 2021

HONDA, Hugo; FACURE, Matheus; YAOHAO, Peng. **Os Três Tipos de Aprendizado de Máquina**. 2017. Disponível em: https://lamfo-unb.github.io/2017/07/27/tres-tipos-am/. Acesso em: 07 jun. 2020.

GOMES, Pedro César Tebaldi. **Regressão Linear**: entenda como utilizar. entenda como utilizar. 2019. Disponível em: https://www.datageeks.com.br/regressao-linear/. Acesso em: 29 jun. 2020.

SQL MAGAZINE. **Descoberta de conhecimento utilizando o processo KDD**. Disponível em: https://www.devmedia.com.br/descoberta-de-conhecimento-utilizando-o-processo-kdd/38709. Acesso em: 13 maio 2021.

SAMPAIO, Igor Garcia; BERNARDINI, Flavia; PAES, Aline; ANDRADE, Eduardo de Oliveira; VITERBO, José. **Avaliação de Modelos de Predição e Previsão Construídos por Algoritmos de Aprendizado de Máquina em Problemas de Cidades Inteligentes**. Porto Alegre: Minicursos Sbsi, 2019. Disponível em: https://sol.sbc.org.br/livros/index.php/sbc/catalog/download/33/133/319-1?inline=1. Acesso em: 25 maio 2021

SOUZA, Emanuel G de. **Implementando Regressão Linear Simples em Python**. 2019. Disponível em: https://medium.com/data-hackers/implementando-regress%C3%A3o-linear-simples-em-python-91df53b920a8#:~:text=O%20R%C2%B2%2C%20tamb%C3%A9m%20chamado%20de,sua%20m%C3%A9dia%20elevado%20ao%20quadrado. Acesso em: 25 maio 2021.