Programa Nanodegree Engenheiro de Machine Learning

Proposta de projeto de conclusão de machine learning - Capstone

Robson Azevedo Rung
27/07/2019

Background do domínio

Imóveis são bens de grande relevância para qualquer sociedade, e têm seus preços formados não de forma única, mas a partir das características que possuem. [A]

Assim sendo, os consumidores potenciais têm a percepção de diferenciar as diversas possibilidades de características em função do que é tido como prioritário. Desta maneira, um determinado consumidor pode escolher seu "pacote" de características disponíveis para cada bem ou serviço em função da percepção de utilidade.

Muitos estudos buscam determinar os atributos intrínsecos e extrínsecos pertencentes a cada residência, a fim de verificar quais deles apresentam maior representatividade para a composição dos instrumentos de demanda e oferta, utilizando-se modelos de preços hedônicos, por meio dos quais é possível analisar a importância relativa de cada atributo em função dos diferentes perfis sociodemográficos. [B]

Enunciação do problema

Uma das atividades de empresas e profissionais do ramo imobiliário é a analisar o valor de imóveis, com objetivo de vendê-los em tempo razoável e maximizar seus lucros.

Entender quais características são as mais relevantes e como impactam os preços é tarefa complexa por si só. Além disso, tais fatores podem mudar ao longo do tempo e variam entre regiões geográficas. Podem, ainda, ser afetadas por diferenças culturais e climáticas.

Por conseguinte, o uso de técnicas de machine learning pode ser de grande valia para esse importante ramo da atividade econômica.

Conjunto de Dados e Inputs

O conjunto de dados deste trabalho será aquele disponibilizado na competição intitulada "*House Prices: Advanced Regression Techniques*" da plataforma Kaggle. [C]

Constam dos dados diversos atributos, por exemplo:

- Tipo da propriedade;
- Tamanho do lote;
- Recursos públicos disponíveis (eletricidade, gás, água e saneamento);
- Condições de acesso;
- Acabamento (materiais);
- Estado de conservação;
- Tipo da fundação;
- Tipo de aquecimento;
- Sistema elétrico;
- Características da garagem;
- Idade do imóvel;
- etc.

Os dados são fornecidos em dois grupos, um de treinamento e outro de teste.

Explanação da solução

Para resolver o problema proposto serão aplicadas técnicas de machine learning, mais especificamente, dois algoritmos de aprendizagem supervisionada para realização de regressões: regressão linear e árvore de decisão.

O objetivo de uma regressão é prever uma variável alvo a partir de outras variáveis de entrada. É necessária a existência de um conjunto de dados com vários exemplos de variáveis de entrada e seus respectivos valores para a variável alvo, de modo que os algoritmos possam ser treinados para encontrar a correlação entre os dados.

Após o treinamento, será possível prever a variável alvo para novos dados de entrada que não constam da base de dados inicial.

Além disso, a mesma tarefa será realizada usando técnicas de deep learning. Será criada e treinada uma rede neural com os dados fornecidos.

Será possível, portanto, comparar a performance das duas técnicas.

Os códigos serão escritos em Python, usando as bibliotecas NumPy, Pandas, Scikit-learn, Seaborn, Matplotlib e Keras.

Avaliação Métrica

Para avaliar a performance dos modelos, será calculado o coeficiente de determinação (R2), um modelo bastante usado para analisar regressões.

O valor de R2 indica o percentual de correlação quadrática entre os valores previstos e os valores reais. Quando o resultado é igual a 0, o modelo de regressão se equivale a um modelo que sempre tem como resultado a média amostral (dados de treinamento). Por outro lado, quando R2 é igual a 1, significa que o modelo foi capaz de prever com precisão os valores da variável alvo.

É possível que R2 tenha como resultado um valor negativo, o que significa que o modelo de regressão é pior do que um modelo que sempre prevê a média.

Design do projeto

O primeiro passo será explorar os dados de modo a tentar entender um pouco a relação entre as características dos imóveis e seus valores de venda, bem como possíveis correlações entre as características. Algumas visualizações serão usadas para esse fim.

Serão calculadas, também, as estatísticas da base de treinamento, como média, mediana, desvio padrão, valores máximos e mínimos.

Na sequência, uma busca por *outliers* e dados faltantes será realizada, de modo a evitar distorções nos modelos durante o treinamento.

Um passo importante virá a seguir, que é a transformação dos atributos (*feature transformation*), no qual serão analisados os seguintes aspectos:

- Existência de dados ordinais textuais que precisam ser convertidos em numéricos;
- Normalização dos dados, com o objetivo de não permitir que algumas características tenham mais peso no treinamento simplesmente por suas faixas de valores serem maiores (isso também será muito útil para o treinamento da rede neural);
- Existência e ajuste de assimetrias (*skew*);
- One-hot encode de variáveis categóricas.

Dar-se-á, então, a execução de alguns passos complementares:

- Regularização (*regularization*) e seleção de atributos (*feature selection*);
- Redução de dimensionalidade usando PCA (*Principal Component Analysis*).

Agora será realizado o treinamento dos modelos. Para o caso da aprendizagem supervisionada, será conduzida uma otimização dos hiperparâmetros usando a técnica de *grid search*.

For fim, o coeficiente de determinação será calculado para avaliar a performance dos modelos.

Modelo de benchmark

Considerando tratar-se de uma competição do Kaggle, a ideia é usar o ranking da competição para avaliar se os modelos desenvolvidos neste projeto tiveram bons resultados.

Para isso, o melhor resultado obtido dentre os modelos propostos será enviado para avaliação por meio da página da competição.

Referências

[A] Almeida, Pedro Henrique Ramos. Fatores determinantes para a formação de preço no mercado imobiliário de Brasília. Brasília. 2001. Universidade de Brasília. Disponível em http://bdm.unb.br/bitstream/10483/2122/1/2011_PedroHenriqueRamosdeAlmeida.pdf

[B] Belfiore, Patrícia Prado. Fávero, Luiz Paulo Lopes. Lima, Gerlando A. S. Franco. Modelos de precificação hedônica de imóveis residenciais na região metropolitana de São Paulo: uma abordagem sob as perspectivas da demanda e da oferta. São Paulo Jan./Mar. 2008. Estud. Econ. vol.38. Disponível em http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-41612008000100004

[C] Kaggle. House Prices: Advanced Regression Techniques. Disponível em https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview