UNIVERSIDADE FEDERAL DE MINAS GERAIS - UFMG
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
CURSO DE SISTEMAS DE INFORMAÇÃO

GILLIARD GABRIEL RODRIGUES

APRENDIZADO DE MÁQUINA APLICADO À PREVISÃO DE PREÇOS DE VENDAS DE CASAS

Belo Horizonte
2022

GILLIARD GABRIEL RODRIGUES

APRENDIZADO DE MÁQUINA APLICADO À PREVISÃO DE PREÇOS DE VENDAS DE CASAS

Trabalho apresentado para a Disciplina Mineração de Dados, pelo Curso de Sistemas de Informação da Universidade Federal de Minas Gerais, ministrada pelo Prof. Wagner Meira Júnior.

Belo Horizonte

2022

SUMÁRIO

1 INTRODUÇÃO

O *Ames Housing dataset* é uma base de dados conhecida que contém dezenas de informações sobre residências em Ames, Iowa.

Uma questão interessante que existe no contexto desses dados é a possibilidade de aplicar técnicas de aprendizado de máquina para prever o valor de venda das casas baseado nas diversas informações disponíveis.

Nesse contexto, o presente trabalho irá apresentar uma análise exploratória em cima dos dados das residências de Ames, Iowa, assim como a aplicação de vários algoritmos de máquina supervisionado (regressão) objetivando a previsão do preço de cada casa baseado nas demais informações disponíveis sobre ela.

2 MOTIVAÇÃO

A motivação por trás deste trabalho está pautada na curiosidade e vontade de descobrir se é possível prever o preço das casas com um R^2 elevado.

3 OBJETIVO

3.1 Geral

O objetivo aqui é fazer uma análise exploratória sobre os dados das residências de Ames, Iowa, explorando as dezenas de características disponíveis sobre cada casa e construir um modelo que preveja de forma satisfatória o preço de venda de uma casa a partir de suas demais características.

3.2 Específicos

As tarefas específicas podem ser divididas em:

- Extração dos dados, que virão de uma base em formato *.csv*, retirada do *Kaggle*;
- Limpeza dos dados, pois a base pode conter informações faltantes ou dados desnecessários para o nosso objetivo;
- Análise exploratória dos dados, a fim de obter *insights* sobre as diversas informações disponíveis sobre as residências e decidir quais utilizar no modelo.
- Preparação dos dados, de forma a deixa-los da melhor forma para o modelo de regressão.
- Aplicação dos algoritmos de regressão linear sem e com regularização (*Lasso*, *Ridge* e *Elastic Net*) e árvores de decisão.
- Análise dos resultados, partindo das métricas de avaliação disponíveis.

4 METODOLOGIA

A metodologia foi inspirada no CRISP-DM, ou seja, dividida em: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação.

Uma parte da aplicação dessa metodologia pode ser vista através do link para o *Google Colab*, que apresenta o entendimento dos dados, a preparação dos dados, a modelagem e a avaliação, com os *scripts* já rodados e os resultados já disponíveis.

5 DESENVOLVIMENTO

5.1 Entendimento dos dados

Como recurso disponível, temos uma base de dados, retirada do Kaggle, que pode ser encontrada no seguinte link e está em formato *csv*. Ela contém as seguintes informações sobre cada pokémons:

- **SalePrice** - o preço de venda da propriedade em dólares. Esta é a variável target.
- **MSSubClass**: A classe de construção
- **MSZoning**: A classificação geral de zoneamento
- **LotFrontage**: pés lineares de rua conectados à propriedade
- **LotArea**: Tamanho do lote em pés quadrados
- **Street**: Tipo de acesso rodoviário
- **Alley**: Tipo de acesso ao beco
- **LotShape**: Forma geral da propriedade
- **LandContour**: Planicidade da propriedade
- **Utilities**: Tipo de utilitários disponíveis
- **LotConfig**: configuração do lote
- **LandSlope**: Declive da propriedade
- **Neighborhood**: locais físicos dentro dos limites da cidade de Ames
- **Condition1**: Proximidade da estrada principal ou ferrovia
- **Condition2**: Proximidade da estrada principal ou ferrovia (se houver uma segunda)
- **BldgType**: Tipo de habitação
- **HouseStyle**: estilo de habitação
- **OverallQual**: Material geral e qualidade de acabamento
- **OverallCond**: avaliação geral da condição
- **YearBuilt**: data de construção original
- **YearRemodAdd**: Data da remodelação
- **RoofStyle**: Tipo de telhado
- **RoofMatl**: Material do telhado
- **Exterior1st**: Revestimento exterior da casa
- **Exterior2nd**: Revestimento externo da casa (se houver mais de um material)
- **MasVnrType**: tipo folheado de alvenaria
- **MasVnrArea**: Área de folheado de alvenaria em pés quadrados
- **ExterQual**: qualidade do material externo
- **ExterCond**: Estado atual do material no exterior
- **Foundation**: Tipo de fundação
- **BsmtQual**: Altura do porão
- **BsmtCond**: Estado geral da cave
- **BsmtExposure**: Walkout ou paredes do porão no nível do jardim
- **BsmtFinType1**: Qualidade da área finalizada do porão
- **BsmtFinSF1**: Tipo 1 pés quadrados acabados
- **BsmtFinType2**: Qualidade da segunda área acabada (se presente)
- **BsmtFinSF2**: Tipo 2 pés quadrados acabados
- **BsmtUnfSF**: pés quadrados inacabados da área do porão
- **TotalBsmtSF**: pés quadrados totais da área do porão
- **Heating**: Tipo de aquecimento

- **HeatingQC**: qualidade e condição do aquecimento
- **CentralAir**: ar condicionado central
- Electrical: Sistema elétrico
- **1stFlrSF**: pés quadrados do primeiro andar
- **2ndFlrSF**: pés quadrados do segundo andar
- **LowQualFinSF**: pés quadrados acabados de baixa qualidade (todos os andares)
- **GrLivArea**: Área habitável acima do solo (pés quadrados)
- **BsmtFullBath**: Banheiros completos no porão
- **BsmtHalfBath**: lavabos no porão
- **FullBath**: Banheiros completos acima do nível do solo
- **HalfBath**: Meios banhos acima do grau
- **Bedroom**: Número de quartos acima do subsolo
- **Kitchen**: Número de cozinhas
- **KitchenQual**: qualidade da cozinha
- **TotRmsAbvGrd**: Total de quartos acima do nível (não inclui banheiros)
- **Functional**: classificação de funcionalidade doméstica
- **Fireplaces**: Número de lareiras
- **FireplaceQu**: Qualidade de lareira
- **GarageType**: localização da garagem
- **GarageYrBlt**: ano em que a garagem foi construída
- **GarageFinish**: Acabamento interior da garagem
- **GarageCars**: Tamanho da garagem na capacidade do carro
- **GarageArea**: Tamanho da garagem em pés quadrados
- **GarageQual**: qualidade de garagem
- **GarageCond**: condição de garagem
- **PavedDrive**: Estrada pavimentada
- **WoodDeckSF**: Área de deck de madeira em pés quadrados
- **OpenPorchSF**: Área de varanda aberta em pés quadrados
- **EnclosedPorch**: Área da varanda fechada em pés quadrados
- **3SsnPorch**: Área de varanda de três estações em pés quadrados
- **ScreenPorch**: Área da varanda de tela em pés quadrados
- **PoolArea**: Área da piscina em metros quadrados
- **PoolQC**: qualidade da piscina
- **Fence**: qualidade da cerca
- **MiscFeature**: Diversos recursos não cobertos em outras categorias
- **MiscVal**: $Valor do recurso diverso
- **MoSold**: Mês Vendido
- **YrSold**: Ano Vendido
- **SaleType**: Tipo de venda
- **SaleCondition**: Condição de venda

A fim de analisar as diversas características foram plotados vários histogramas e *box-plots* das *features* e observou-se uma tendência linear positiva entre algumas variáveis independentes (*'OverallQual'*, *'OverallCond'*, *'YearBuilt'*, *'YearRemodAdd'*, *'TotalBsmtSF'*, *'stFlrSF'*, *'GrLivArea'*, *'GarageYrBlt'* e 'GarageArea'e a variável dependente (*SalePrice*), mas nada muito bem definido. Em algumas variáveis foi possível melhorar a distribuição através de transformação logarítmica.

Observou-se também que as variáveis categóricas possuíam categorias muito dispersas, exibindo predominância para alguns valores nas casas dessa região.

Através de algumas visualizações surgiu o indício de que *'BsmtQual'* (Altura do porão), *'BsmtCond'* (Condição do porão), *'BsmtExposure'* (Paredes do porão no nível do jardim ou paralisação), *'Condition1'* (Proximidade de estrada principal ou ferrovia), *'Condition2'* (Proximidade da estrada principal ou ferrovia, se houver uma segunda), *'ExternQual'* (Qualidade do exterior), *'KitchenQual'* (Qualidade da cozinha), *'RoofMatl'* (Material do telhado), *'SaleType'* (Tipo da venda) são *features* que aparentam impactar em preços mais elevados. *'Neighborhood'* (Vizinhança) influenciavam bastante o valor da venda.

5.2 Preparação dos dados

A fim de preparar os dados para aplicação do modelo, as *features* que possuíam mais de 50% de seus dados ausentes foram removidas, os *outliers* corrigidos, algumas variáveis sofreram transformação logarítmica, as variáveis categóricas ordinais foram convertidas para tipos numéricos através de *Label Encoding* e as não ordinais, através de *One Hot Encoding*. Por fim, os dados foram reescalados através do *Standard Scaler*.

5.3 Modelagem

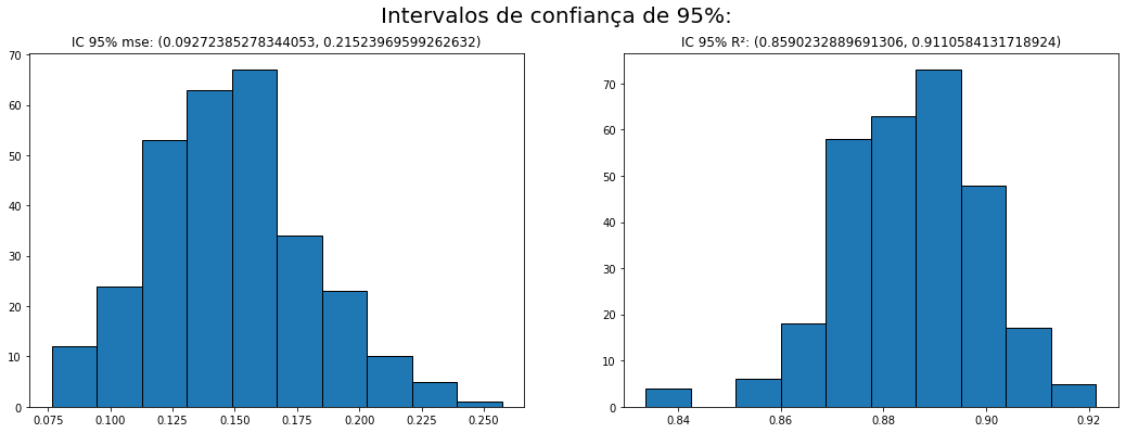
Os dados foram divididos em 80% para treino e 20% para teste. A fim de escolher a técnica a ser aplicada nos dados, foram testados os seguintes modelos: regressão linear, regressão *Lasso*, regressão *Ridge*, regressão *Elastic Net* e árvores de decisão.

A fim de decidir qual algoritmo utilizar, foram gerados intervalos de confiança de 95% via *bootstrap* para o erro quadrado médio e o coeficiente de determinação (R^2). Como os histogramas apresentaram sobreposição, constatou-se que os resultados eram estatisticamente equivalentes (possuindo R^2 aproximadamente entre 88 e 93%). Como a regularização *Ridge* poderia ser útil contra a multicolinearidade dos dados e foi a segunda com maior R^2 no treino (91,82%), ela foi escolhida e o resultado será apresentado na próxima seção.

6 RESULTADOS/CONCLUSÕES

A técnica escolhida para aplicar aos dados de teste foi a regressão linear múltipla com regularização *Ridge* e obteve um coeficiente de determinação de 88,46% nos dados de teste e seu erro quadrado médio ficou entre 0,0927 e 0,2152 (IC de 95%), conforme é possível visualizar na figura 1.

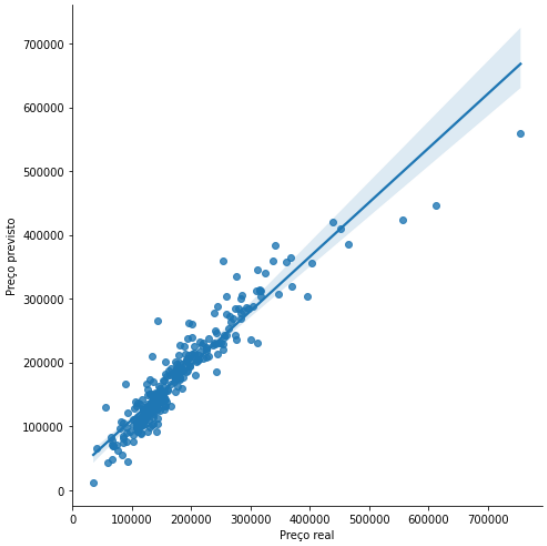
Figura 1 – Intervalo de confiança para R^2 e erro quadrado médio no teste



Fonte: Google Colab.

A figura 2 apresenta um comparativo do **valor real** de y e o **valor previsto** para os preços de casas.

Figura 2 – Valor Real x Valor Previsto



Fonte: Google Colab.