Nanodegree Engenheiro de Machine Learning - Udacity

Proposta de projeto final - 09 de setembro de 2018

Autor: Marcos Paulo da Silva Falcirolli

Proposta do projeto

A proposta deste projeto é criar um modelo estatístico de capaz de estimar o valor de venda de um imóvel, este é um problema conhecido da área de Engenharia de Avaliação de Bens.

Em 1977 foi criada a NB-502, que foi a primeira norma brasileira para avaliação de imóveis urbanos, este foi um marco para o mercado imobiliário brasileiro, pois antes disso o valor de um imóvel era obtido de maneira subjetiva, o avaliador estimava o valor baseados em sua experiencia. Com esta norma foi determinado a necessidade de um modelo de regressão linear com níveis de precisão definidos para as avaliações, assim substituindo a subjetividade pela ciência.

Em 1989 surge a NBR 5676 que é uma revisão da NB-502. Em 2001 surge a NBR 14653 que tem sua versão de 2011 vigente até os dias de hoje. Em todas essas normas é imprescindível que os avaliadores de imóveis utilizem inferência estatística a partir de um modelo de regressão linear para avaliação de imóveis urbanos. (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2011)

Nos dias de hoje normalmente um engenheiro de avaliação de bens ao realizar um laudo coleta os dados quase dos imóveis quase que manualmente, porém com a informatização do setor esperamos que em breve os engenheiros tenham grandes bases de *bigdata* com esses dados assim causando uma possibilidade de mudança no setor. Com grandes volumes de dados nos permitirá aplicar técnicas estatísticas mais avançadas trazendo possivelmente resultados mais precisos e de maneira mais prática. Portanto a exploração de técnicas de regressão mais avançadas pode ser de grande contribuição para esta mudança.

O laudo de avalição de um imóvel que determina o seu valor tem inúmeras aplicações, como por exemplo: negociação de venda, financiamento imobiliário, hipotecas, separação de bens e outras.

Motivações pessoais para este trabalho:

Sou engenheiro civil e já tive a oportunidade de realizar diversos laudos de avaliação de imóveis urbanos, com este estudo pretendo analisar o potencial de técnicas mais avançadas para esta aplicação. Também pretendo conseguir uma boa pontuação no Kaggle, assim incluindo este projeto em meu portfólio.

Descrição do problema

O problema a ser resolvido é obter o preço de venda de um imóvel.

A solução proposta neste trabalho é criar um modelo utilizando técnicas de regressão, que terá com input diversas características do imóvel, assim poderemos estimar/inferir os preços de venda. A performance do modelo poderá ser medida através do *Root-Mean-Squared-Error* entre os valores de venda apresentados no conjunto de dados e os valores de venda previstos ou do coeficiente de determinação (R^2).

Conjuntos de dados e entradas

O conjunto de dados neste trabalho será o conjunto da competição Kaggle "House Prices: Advanced Regression Techniques" que deverá nos permitir uma boa aplicação de diversas técnicas de regressão e assim um bom estudo e aplicação da biblioteca do scikit-learn (regressão).

Este conjunto é composto por setenta e nove variáveis dependentes que são características relacionados ao imóvel e uma variável dependente que é o valor de venda do imóvel (SalePrice). As *features* podem ser vistas na tabela abaixo ou no acessadas no site https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Tabela 1 – Descrição das *features* (variáveis):

Feature	Descrição	Feature	Descrição
SalePrice	The property's sale price in dollars.	Heating	Type of heating
MSSubClass	The building class	HeatingQC	Heating quality and condition
MSZoning	The general zoning classification	CentralAir	Central air conditioning
LotFrontage	Linear feet of street connected to property	Electrical	Electrical system
LotArea	Lot size in square feet	1stFlrSF	First Floor square feet
Street	Type of road access	2ndFlrSF	Second floor square feet
Alley	Type of alley access	LowQualFinSF	Low quality finished square feet (all floors)
LotShape	General shape of property	GrLivArea	Above grade (ground) living area square feet
LandContour	Flatness of the property	BsmtFullBath	Basement full bathrooms
Utilities	Type of utilities available	BsmtHalfBath	Basement half bathrooms
LotConfig	Lot configuration	FullBath	Full bathrooms above grade
LandSlope	Slope of property	HalfBath	Half baths above grade
Neighborhood	Physical locations within Ames city limits	Bedroom	Number of bedrooms above basement level
Condition1	Proximity to main road or railroad	Kitchen	Number of kitchens
Condition2	Proximity to main road or railroad (if a second is present)	KitchenQual	Kitchen quality
BldgType	Type of dwelling	TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)
HouseStyle	Style of dwelling	Functional	Home functionality rating
OverallQual	Overall material and finish quality	Fireplaces	Number of fireplaces
OverallCond	Overall condition rating	FireplaceQu	Fireplace quality
YearBuilt	Original construction date	GarageType	Garage location
YearRemodAdd	Remodel date	GarageYrBlt	Year garage was built
RoofStyle	Type of roof	GarageFinish	Interior finish of the garage
RoofMatl	Roof material	GarageCars	Size of garage in car capacity
Exterior1st	Exterior covering on house	GarageArea	Size of garage in square feet
Exterior2nd	Exterior covering on house (if more than one material)	GarageQual	Garage quality
MasVnrType	Masonry veneer type	GarageCond	Garage condition
MasVnrArea	Masonry veneer area in square feet	PavedDrive	Paved driveway
ExterQual	Exterior material quality	WoodDeckSF	Wood deck area in square feet
ExterCond	Present condition of the material on the exterior	OpenPorchSF	Open porch area in square feet
Foundation	Type of foundation	EnclosedPorch	Enclosed porch area in square feet
BsmtQual	Height of the basement	3SsnPorch	Three season porch area in square feet
BsmtCond	General condition of the basement	ScreenPorch	Screen porch area in square feet
BsmtExposure	Walkout or garden level basement walls	PoolArea	Pool area in square feet
BsmtFinType1	Quality of basement finished area	PoolQC	Pool quality
BsmtFinSF1	Type 1 finished square feet	Fence	Fence quality
BsmtFinType2	Quality of second finished area (if present)	MiscFeature	Miscellaneous feature not covered in other categories
BsmtFinSF2	Type 2 finished square feet	MiscVal	$Value of miscellaneous feature
BsmtUnfSF	Unfinished square feet of basement area	MoSold	Month Sold
TotalBsmtSF	Total square feet of basement area	YrSold	Year Sold
Heating	Type of heating	SaleType	Type of sale
HeatingQC	Heating quality and condition	SaleCondition	Condition of sale

Observação importante: o conjunto de dados fornecido pelo Kaggle é dividido em duas partes a primeira chamada de “train”, esta traz informações de 1460 imóveis. O segundo chamado de “test” com 1459 imóveis, no entanto o conjunto teste não apresenta a variável independente. Portanto utilizaremos apenas o “train” para treinar nosso modelo, mas ainda assim queremos criar um modelo que seja compatível com o “test” para posteriormente comparar a solução desenvolvida neste trabalho com o a de outros participantes da mesma competição do Kaggle.

Descrição da solução

A solução consiste na criação de um modelo preditivo dos valores de venda.

Para a criação deste modelo será realizado o estudo das variáveis, assim quando necessário a aplicação de transformação, combinação e até o drop de variáveis.

Em seguida uma breve etapa pré-processamento do conjunto de dados para que o conjunto de dados possa ser interpretado pelos algoritmos de regressão.

Continuaremos com a aplicação de técnicas de regressão do scikit-learn como:

LinearRegression, LinearRegression com PolynomialFeatures, SVR (Support Vector Regression), SGDRegressor, DecisionTreeRegressor, GradientBoostingRegressor.

A avalição individual de cada modelo levando em consideração, *Root-Mean-Squared-Error* entre os valores de venda e os valores de venda previstos em um subconjunto do train que será nosso conjunto de validação cruzada, o coeficiente de determinação (R^2) e o tempo de execução de cada algoritmo

Tentaremos a combinação dos resultados dos algoritmos mais promissores assim criando alguns modelos Ensemble, e comparando sua performance com os demais. Por fim fazermos a previsão do conjunto test e este será submetido no Kaggle para comparação com os demais competidores.

Modelo de referência (benchmark)

Utilizaremos dois modelos de referência, o primeiro visa validar a proposta do trabalho, o segundo verificar se o melhor modelo encontrado apresenta uma performance satisfatória.

Modelo 1 – Regressão Linear Simples:

Para validar a proposta do projeto iremos comparar o modelo gerado a partir da técnica de regressão linear simples levando em consideração as métricas de avaliação *Root-Mean-Squared-Error* e R^2 com os outros modelos gerados a partir das outras técnicas de regressão utilizadas neste trabalho.

Modelo 2 - Kaggle leaderboard:

#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲2302	DSXL			0.06628	3	5d
2	▼1	Igor S			0.06946	8	8d
3	▼1	Zheng Pan			0.08021	3	1mo
4	▼1	Javale			0.08397	1	2mo
5	▼1	mohammed khamis			0.10567	1	12d

Figura 1 - Kaggle Leaderboard. Fonte: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/leaderboard

O melhor modelo apresentado neste trabalho será submetido na competição, assim o modelo poderá ser comparado com o de outros competidores através do score fornecido pelo Kaggle.

Como objetivo secundário deste trabalho o score deverá estar entre os 10% melhores competidores.

Design do projeto

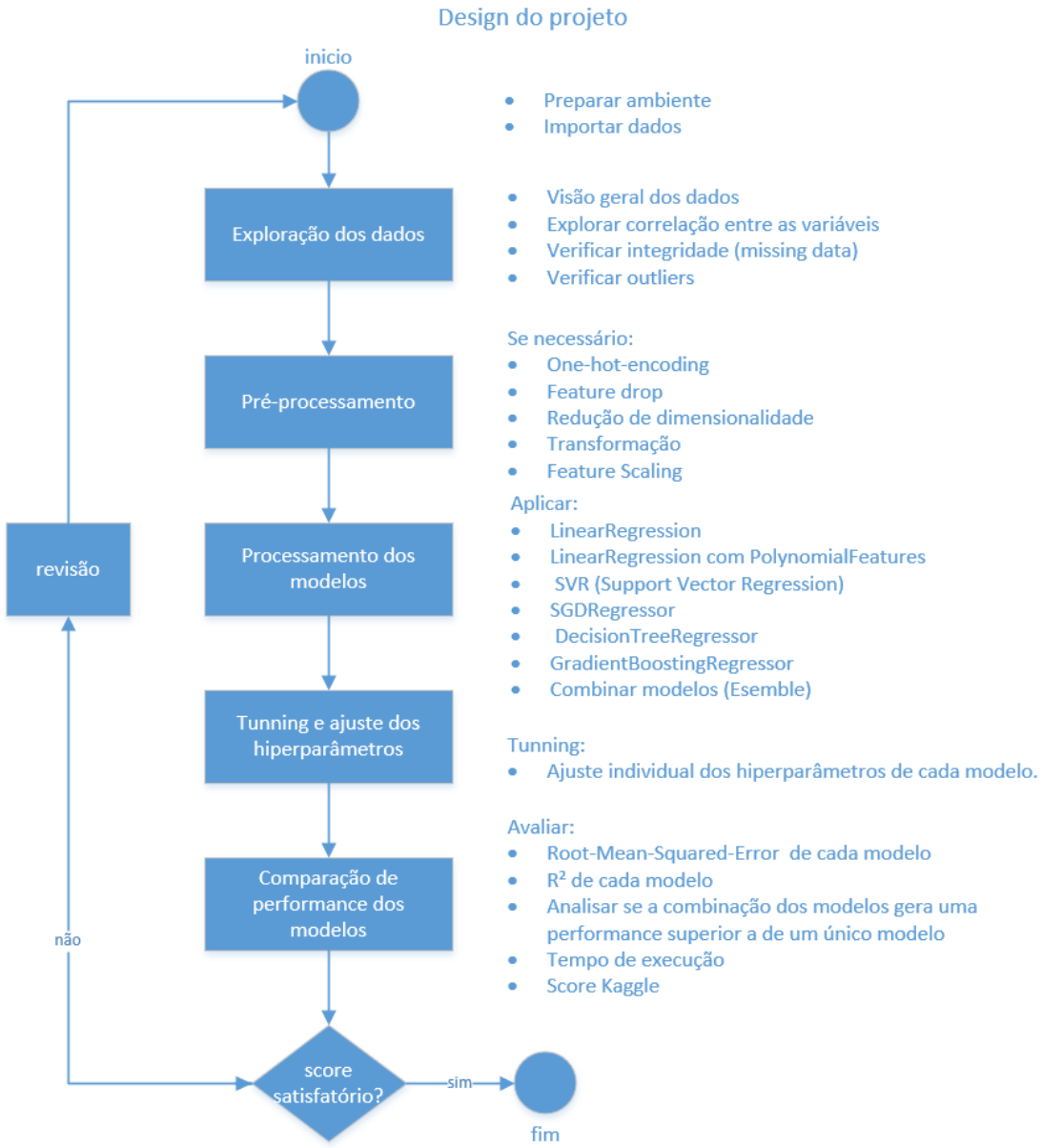

Figura 2 – Design do Projeto.

Referências bibliográficas

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2011, **NBR 14653-1: Avaliação de bens**. São Paulo. 2001