Competição Kaggle – House Prices

Por Alvaro Wang

Sobre o Kaggle

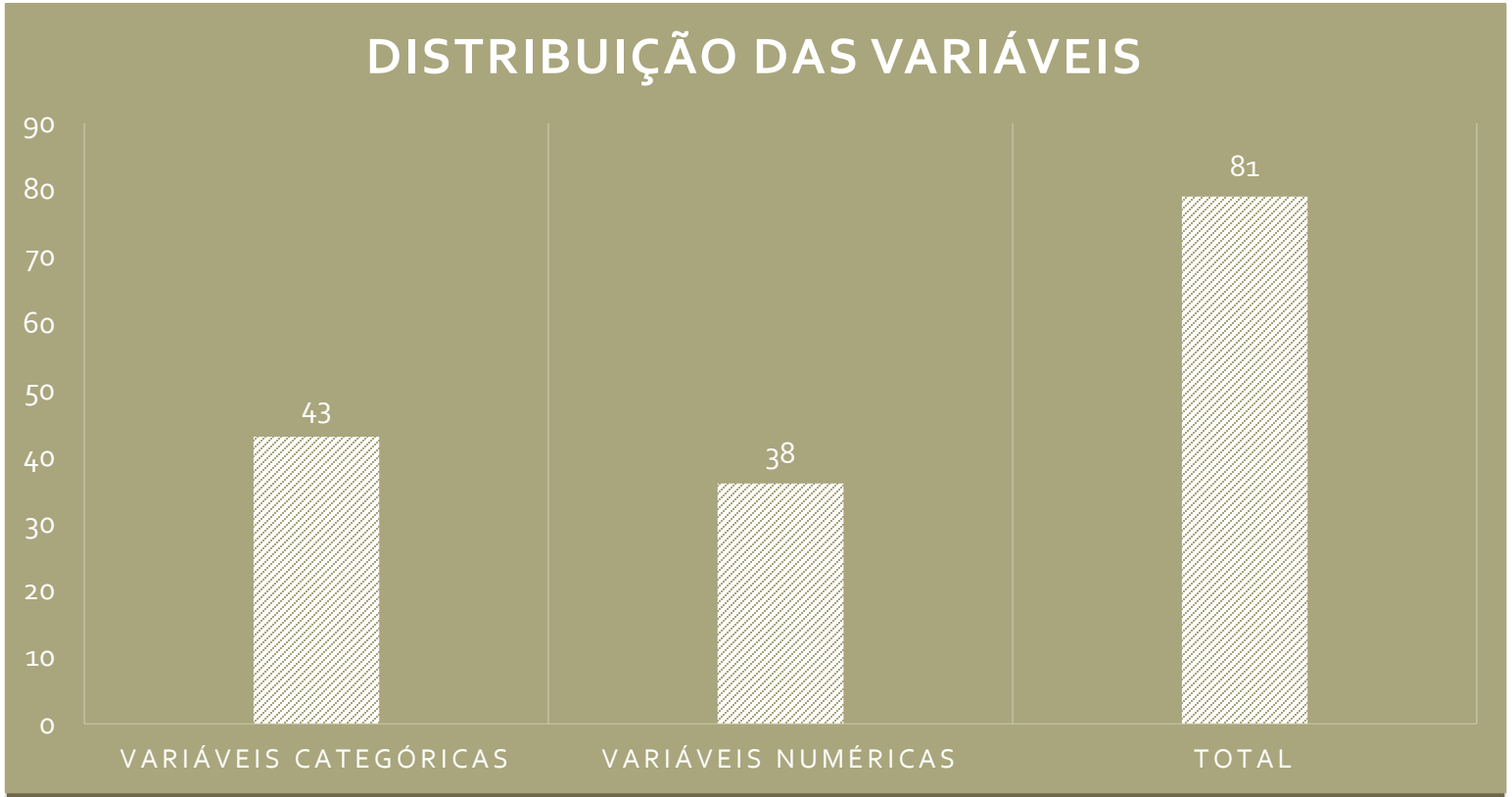
- Kaggle é um website onde frequentemente lançam competições para que praticantes de machine learning possam testar suas habilidades preditivas usando algoritmos de aprendizado de máquina, muitas vezes com prêmios em dinheiro substanciosos.

O contexto

Onde: Awes, Iowa

Campo: Imobiliário

Ação: Usar aprendizado de máquina para precificar imóveis da região baseados em suas características



Exemplos de variáveis categóricas: Tipo de teto, qualidade da cozinha. (descrito com palavras)
Exemplos de variáveis numéricas: Número de banheiros, metro quadrado. (descrito com números)

Os dados em dados

Observações sobre os dados:

- Há 1460 empreendimentos imobiliários
- Há apenas 7 itens com piscina
- 1379 empreendimentos tem garagem
- Os dados faltantes não são aleatórios

0	Id	1460 non-null	int64
1	MSSubClass	1460 non-null	int64
2	MSZoning	1460 non-null	object
3	LotFrontage	1201 non-null	float64
4	LotArea	1460 non-null	int64
5	Street	1460 non-null	object
6	Alley	91 non-null	object
7	LotShape	1460 non-null	object
8	LandContour	1460 non-null	object
9	Utilities	1460 non-null	object
10	LotConfig	1460 non-null	object
11	LandSlope	1460 non-null	object
12	Neighborhood	1460 non-null	object
13	Condition1	1460 non-null	object
14	Condition2	1460 non-null	object
15	BldgType	1460 non-null	object
16	HouseStyle	1460 non-null	object
17	OverallQual	1460 non-null	int64
18	OverallCond	1460 non-null	int64
19	YearBuilt	1460 non-null	int64
20	YearRemodAdd	1460 non-null	int64
21	RoofStyle	1460 non-null	object
22	RoofMatl	1460 non-null	object
23	Exterior1st	1460 non-null	object
24	Exterior2nd	1460 non-null	object
25	MasVnrType	1452 non-null	object
26	MasVnrArea	1452 non-null	float64
27	ExterQual	1460 non-null	object
28	ExterCond	1460 non-null	object
29	Foundation	1460 non-null	object
30	BsmtQual	1423 non-null	object
31	BsmtCond	1423 non-null	object
32	BsmtExposure	1422 non-null	object
33	BsmtFinType1	1423 non-null	object
34	BsmtFinSF1	1460 non-null	int64
35	BsmtFinType2	1422 non-null	object
36	BsmtFinSF2	1460 non-null	int64
37	BsmtUnfSF	1460 non-null	int64
38	TotalBsmtSF	1460 non-null	int64
39	Heating	1460 non-null	object
40	HeatingQC	1460 non-null	object
41	CentralAir	1460 non-null	object
42	Electrical	1459 non-null	object
43	1stFlrSF	1460 non-null	int64
44	2ndFlrSF	1460 non-null	int64
45	LowQualFinSF	1460 non-null	int64
46	GrLivArea	1460 non-null	int64
47	BsmtFullBath	1460 non-null	int64
48	BsmtHalfBath	1460 non-null	int64
49	FullBath	1460 non-null	int64
50	HalfBath	1460 non-null	int64
51	BedroomAbvGr	1460 non-null	int64
52	KitchenAbvGr	1460 non-null	int64
53	KitchenQual	1460 non-null	object
54	TotRmsAbvGrd	1460 non-null	int64
55	Functional	1460 non-null	object
56	Fireplaces	1460 non-null	int64
57	FireplaceQu	770 non-null	object
58	GarageType	1379 non-null	object
59	GarageYrBlt	1379 non-null	float64
60	GarageFinish	1379 non-null	object
61	GarageCars	1460 non-null	int64
62	GarageArea	1460 non-null	int64
63	GarageQual	1379 non-null	object
64	GarageCond	1379 non-null	object
65	PavedDrive	1460 non-null	object
66	WoodDeckSF	1460 non-null	int64
67	OpenPorchSF	1460 non-null	int64
68	EnclosedPorch	1460 non-null	int64
69	3SsnPorch	1460 non-null	int64
70	ScreenPorch	1460 non-null	int64
71	PoolArea	1460 non-null	int64
72	PoolQC	7 non-null	object
73	Fence	281 non-null	object
74	MiscFeature	54 non-null	object
75	MiscVal	1460 non-null	int64
76	MoSold	1460 non-null	int64
77	YrSold	1460 non-null	int64
78	SaleType	1460 non-null	object
79	SaleCondition	1460 non-null	object
80	SalePrice	1460 non-null	int64

Uma explicação sobre dados faltantes

Imputações

Dados faltantes

Dados faltantes podem ser um problema, pois podem indicar desde erros na coleta, como podem indicar vieses na coleta, e, em última análise, podem inviabilizar estudos de Aprendizado de Máquina.

Existem três categorias em que dados faltantes podem cair: MAR, MCAR e MNAR:

- MAR (Missing at random ou faltando aleatoreamente)
- MCAR (Missing completely at random ou faltando de maneira completamente aleatória)
- MNAR (Missing not at random ou faltando de maneira não aleatória)

Definições importantes: Viés - Manifesta-se como uma inclinação irracional a atribuir um julgamento mais favorável ou desfavorável a alguma coisa, pessoa ou grupo. (wikipedia)

MAR (missing at random)

O nome ‘aleatoreamente’ se refere ao fato de que os valores faltantes não se relacionam com outra variável, mas podem estar relacionadas com a variável em si.

Por exemplo, digamos que dentro de um questionário tenha uma pergunta sobre renda. Caso se pergunte isto para menores de idade, há grande probabilidade de que este estrato da população deixe em branco, ou seja, a falta de dados de renda não é causada pela variável renda, mas sim pela variável idade.

Para tratar estes dados faltantes, uma estratégia pode ser tentar entender a razão da lacuna e preencher de acordo com outras variáveis.

Caso não tratemos, introduzimos viés ao nosso modelo.

MCAR (missing completely at random)

Neste caso, os dados realmente são aleatórios, como por exemplo, caso um bug aleatório corrompesse algumas submissões de um questionário, ou se em papel, vento soprasse nas folhas, sendo que nem todas as folhas foram recuperadas.

Este seria um caso simples onde seria seguro tanto imputar, seja com média ou mediana, ou simplesmente eliminar os dados faltantes que não causaria a introdução de viés ao nosso modelo.

Infelizmente, este também é o caso mais raro.

Definições importantes: Mediana - Valor que separa a metade maior e a metade menor de uma amostra, uma população ou uma distribuição de probabilidade.(wikipedia)

MNAR (missing not at random)

- Neste caso, a falta de dado é causada pela variável em si. Por exemplo, pessoas que saem sem aviso de um estudo sobre um remédio devido aos efeitos colaterais do próprio remédio.
- O viés nunca será eliminado neste caso, pois nenhuma informação pode ser obtida sobre a variável perdida.
- Uma imputação cuidadosa pode ajudar a diminuir o viés, mas este nunca será eliminado, e é o pior caso que se pode encontrar.

Nossos casos

- Existem valores vazios como qualidade e tamanho da piscina, tipo e tamanho de porão, garagem, lareira que se relacionam às variáveis piscina, porão e etc, o que siginifica que são valores vazios do tipo MAR
- Para tratar estes valores iremos colocar 'None' para variáveis categóricas e 'zero' para variáveis numéricas

Seleção de variáveis

- Em nosso data set temos 80 variáveis usadas para prever a variável SalePrice.
- Para determinar se devemos usar todas as variáveis, iremos lançar mão de duas técnicas estatísticas diferentes
  - Qui Quadrado
  - Bootstrapping

Teste Chi Quadrado

Seleção de variáveis categóricas

• É o valor que o teste mede	• É um valor calculado a partir do valor de Chi-Quadrado	• É o valor que dá sentido ao Chi – quadrado
Chi – quadrado	p - valor	Valor crítico

Testa a independência entre duas variáveis categóricas

Se p – valor for menor que 0.05, quer dizer que existe 95% de chance de as categorias serem independentes uma da outra

Se Valor Crítico for menor que Chi-Quadrado, reforça-se a ideia de que as categorias são independentes

Qui-Quadrado

Com os testes em mãos definimos que váriaveis como:

Fence,
PoolQC,
PavedDrive,

não se relacionam de maneira estatisticamente significativa com o preço

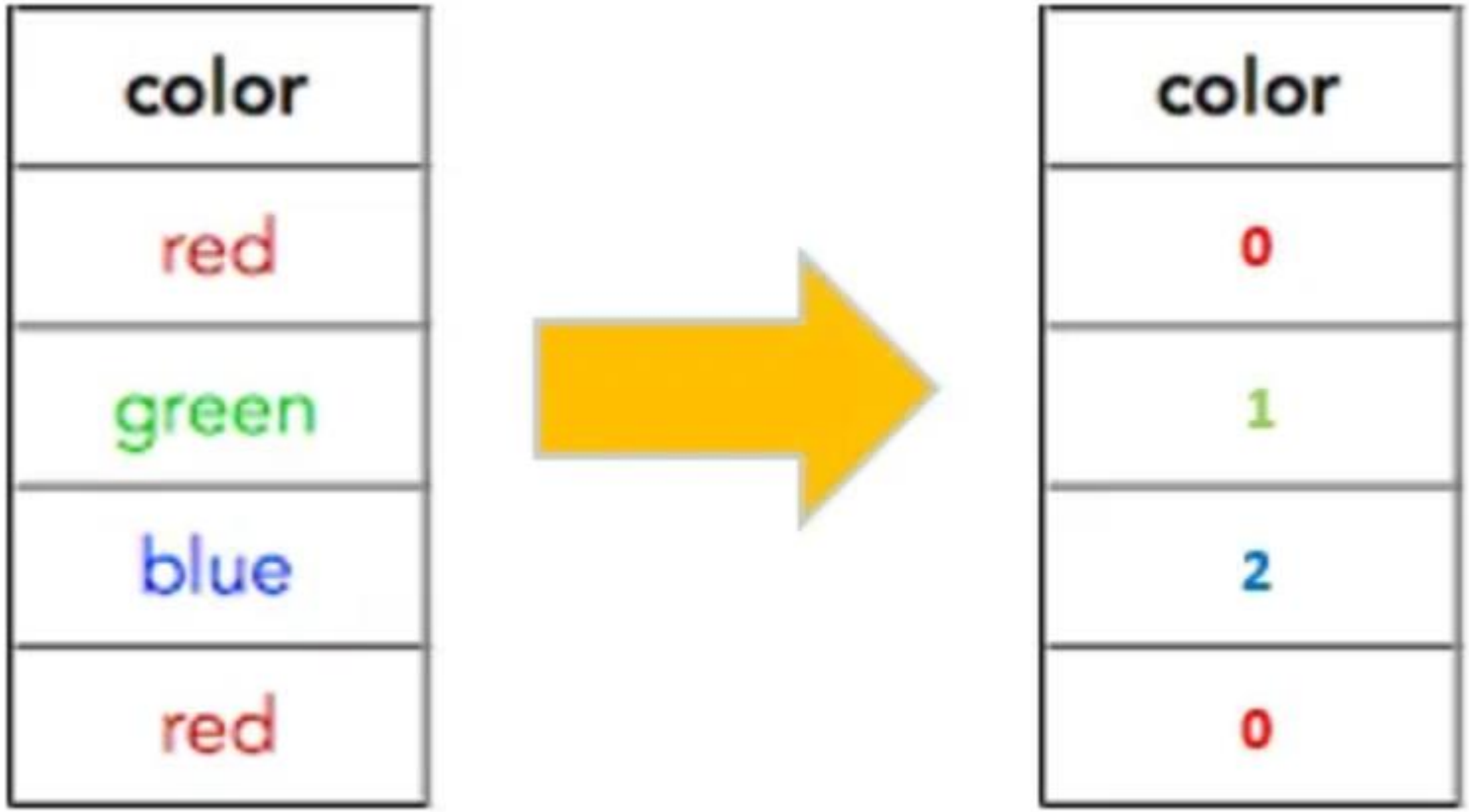
```
Variable: MSZoning
  Chi2: 3147.8911158183737
  p-value: 4.3483250606822396e-11
  Degrees of Freedom: 2648
  Critical Value: 2768.8281535489073
  MSZoning is correlated with SalePrice


Variable: Street
  Chi2: 888.3129945096931
  p-value: 8.338870380464053e-09
  Degrees of Freedom: 662
  Critical Value: 722.9662553606306
  Street is correlated with SalePrice


Variable: Alley
  Chi2: 1233.0758732766717
  p-value: 0.9637567934663973
  Degrees of Freedom: 1324
  Critical Value: 1409.7637498147142
  Alley is NOT correlated with SalePrice


Variable: LotShape
  Chi2: 2446.2353573800365
  p-value: 4.7247291559804O2e-12
  Degrees of Freedom: 1986
  Critical Value: 2090.7894487834265
  LotShape is correlated with SalePrice
```

```
Variable: Utilities
  Chi2: 242.49942883253368
  p-value: 1.0
  Degrees of Freedom: 662
  Critical Value: 722.9662553606306
  Utilities is NOT correlated with SalePrice


Variable: LotConfig
  Chi2: 2771.9854545078742
  p-value: 0.045806211958033756
  Degrees of Freedom: 2648
  Critical Value: 2768.8281535489073
  LotConfig is correlated with SalePrice


Variable: LandSlope
  Chi2: 1388.847891459522
  p-value: 0.10508638737793884
  Degrees of Freedom: 1324
  Critical Value: 1409.7637498147142
  LandSlope is NOT correlated with SalePrice


Variable: Neighborhood
  Chi2: 16898.75578956907
  p-value: 1.3649601026882960-08
  Degrees of Freedom: 15888
  Critical Value: 16182.341330330977
  Neighborhood is correlated with SalePrice
```

LabelEncoder

Traduzindo linguagem humana em computacional

Label encoding atribui um número para cada instância de variáveis categóricas. Esta etapa é importante para garantir que o computador consiga interpretar o conteúdo das variáveis categóricas.



Fonte: https://medium.com/@sunnykumar1516/what-is-label-encoding-application-of-label-encoder-in-machine-learning-and-deep-learning-models-c593669483ed

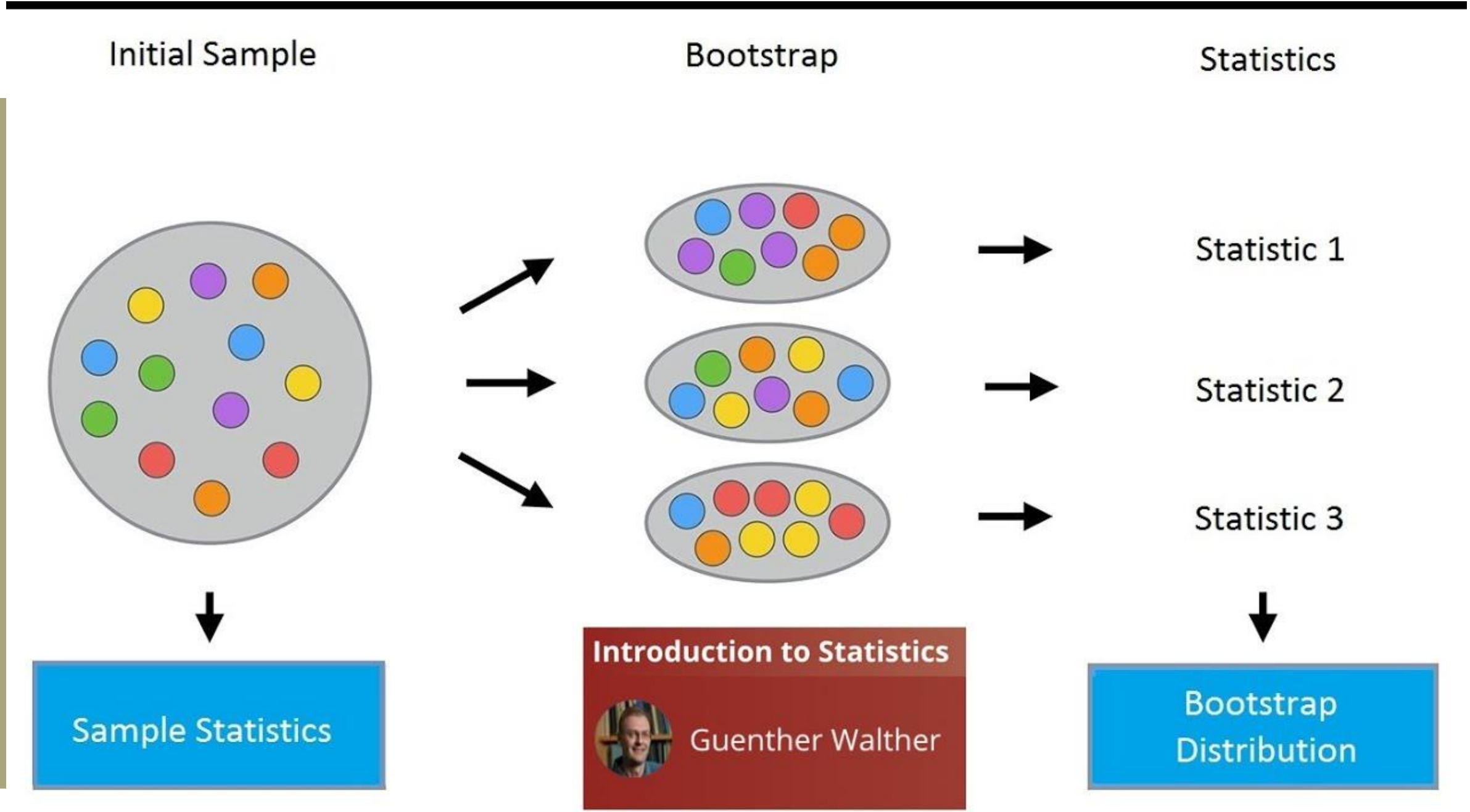
Bootstrapping

Seleção de variáveis numéricas

Para definir as melhores variáveis categóricas, Podemos lançar mão de uma técnica chamada Bootstrapping que se resume em:

- Criar um mini modelo de Aprendizado de Máquina
- Usar diferentes variáveis e testar se eles são bons em prever o preço de um empreendimento imobiliário
- Repete-se esse experimento diversas vezes até que tenhamos alguma certeza de quais variáveis impactam o preço.

```
Selected Best Features: ['MSSubClass', 'OverallCond', 'YearBuilt', 'GarageCars', 'OverallQual', 'KitchenAbvGr', 'Bedr
oomAbvGr', 'BsmtHalfBath', 'TotRmsAbvGrd', 'FullBath', 'Fireplaces', 'BsmtFullBath', 'HalfBath', 'YrSold', 'GrLivAre
a', 'YearRemodAdd', 'MoSold', 'ScreenPorch', 'LotFrontage', 'PoolArea', '1stFlrSF', 'MasVnrArea', '2ndFlrSF', 'WoodDe
ckSF', 'LowQualFinSF']
```



https://www.youtube.com/watch?app=desktop&v=d3mcuJycJfI

Aprendizado de Máquina

Um resumo

Para que serve?

- Aprendizado de máquina tenta a partir de algoritmos estatísticos, produzir previsões sem que o resultado tenha sido explicitamente programado por um humano.

(https://pt.wikipedia.org/wiki/Aprendizado_de_m%C3%A1quina)

Tipos de aprendizado de máquina

Supervisionado

- Sabemos quais os resultados são corretos
- A máquina tenta aprender com uma base de dados
- E então replicar seu aprendizado em outra base de dados distintas
- Comparamos seu desempenho com outros modelos de máquina ou até modelos humanos

Não Supervisionado

- Não sabemos quais os resultados corretos
- A máquina tenta aprender com uma base de dados
- O humano analisa se o que a Máquina aprendeu é satisfatório ou não.
- Exemplo: Máquinas que detectam se a foto tem gato ou cachorro

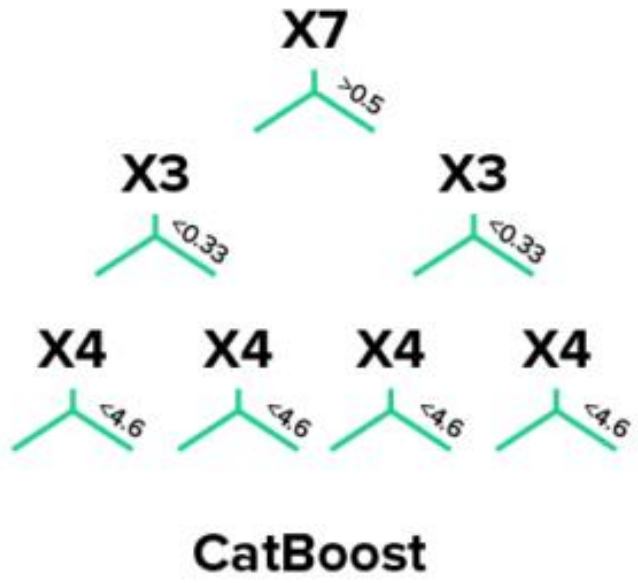
Passo a passo

- Definir o objetivo do projeto
- Limpar os dados para que não tenhamos valores nulos ou que o computador seja incapaz de ler
- Buscar entender quais dados são relevantes para o seu problema, correndo o risco de confundir o algoritmo caso esta etapa seja pulada
- Buscar a melhor maneira de apresentar os dados ao computador a fim de facilitar o algoritmo
- Comparar diversos algoritmos através de métricas relevantes
- Selecionar o modelo e medir seu desempenho final

O algoritmo:

CatBoost

- CatBoost funciona através de várias árvores de decisão que buscam chegar a uma conclusão sobre o seu problema (preço de imóvel). A partir da primeira árvore, árvores subsequentes são criadas tentando melhorar o modelo anterior, criando uma floresta com alto poder de predição.



Julgando o desempenho do computador:

MSE

- Este número é o que se chama MSE, ou mean squared error ou erro médio quadrado. O valor é obtido:
- Subtraindo o preço previsto pela máquina pelo preço real de cada empreendimento
- Tirando se a média destes valores
- Elevando à raiz quadrada

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

Y_i = observed values

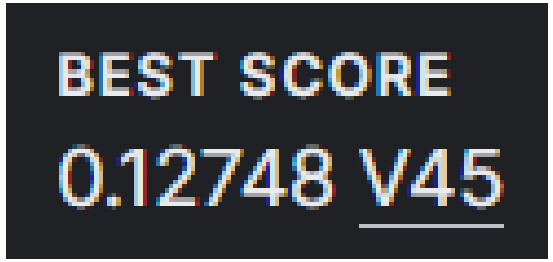
$\hat{Y}_i$ = predicted values

Interpretando MSE

- Para o nosso contexto, um MSE de 0.12748 significa que se tirarmos a raiz quadrada deste número, obtemos, em dólar, a variação média entre o preço real e o preço previsto pela máquina, que dá aproximadamente U$$ 0.357.

- Em miúdos, a variação entre o preço real e o previsto pela máquina é de aproximadamente 0.357 centavos de dolar em média, podendo inclusive ser bem distante de 0.357 centavos.

- O que este resultado não significa: Todo empreendimento tem uma variação de exatos 0.357 centavos de dólar de seu preço real



Obrigado

Contatos:

Email: alvarowang@gmail.com

Linkedin: https://www.linkedin.com/in/alvaro-wang-1b0588155/

Telefone: (11) 984644425