Fundação Vanzolini

Dominando Big Data com o uso de Plataformas Gratuitas (nível intermediário)

Aula 5

LexisNexis® RISK SOLUTIONS

RELX Group

PRO

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

USP UNIVERSIDADE DE SÃO PAULO

Bem-vindo! – Agenda da aula 5

- ✓Desafio Lending Club
- ✓Boosted Forest
- ✓Intervalo
- ✓Disponibilização de consultas

Exercício prático:

Crie o data frame do dataset do Lending Club

- Considere a aplicação de aprendizagem supervisionada
- Se baseie nos resultados do perfilamento de dados

Fundação Vanzolini

Métodos Ensemble

Bibliotecas de machine learning

Não supervisionado

Clusterização
DBSCAN
K-Means

PLN
Text Vectors

Supervisionado

Classificação
SVM
Árvores de decisão
Regression logística
Classification Forest

Regressão
Regressão linear
GLM
Regression Forest

Redes neurais & Deep Learning

Autoencoders

Redes neurais convolucionais

Redes neurais recorrentes

Perceptrons

Métodos ensemble

Random Forest

Gradient Boosted Forest

Gradient Boosted Trees

Modelos de árvore de decisão

Algoritmo de aprendizagem supervisionada que utiliza um conjunto hierárquico de condições de separação de dados para determinar a variável dependente.

XOR Truth Table

Feature 1	Feature 2	Result
0	0	0
0	1	1
1	0	1
1	1	0

(https://hpccsystems.com/blog/learning-trees-guide-to-decision-tree-based-machine-learning)

Exemplo prático de ML

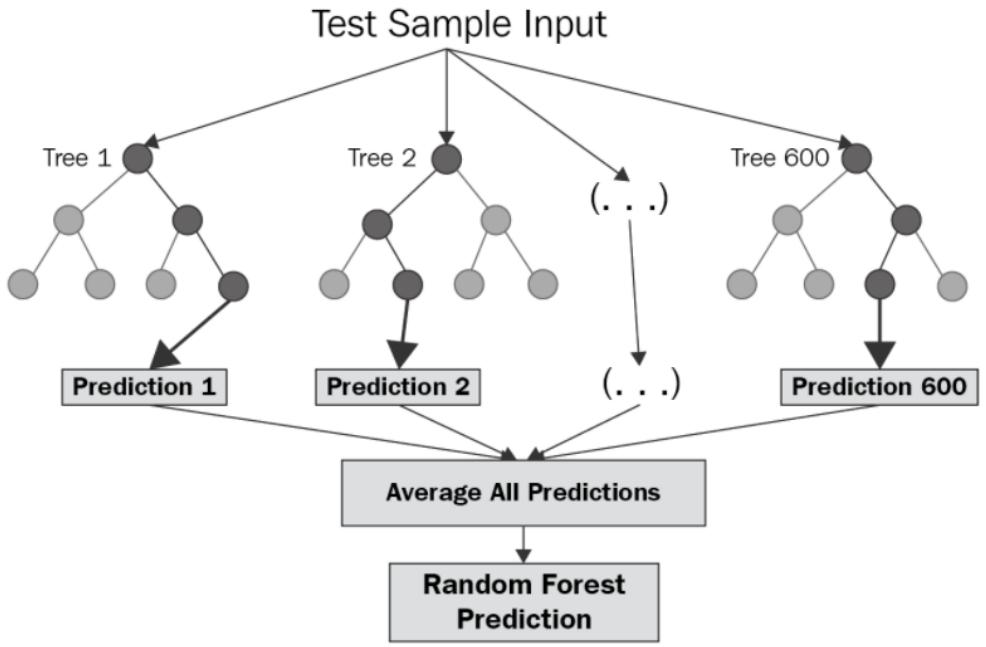
- Dado o conjunto de dados sobre árvores em uma floresta:

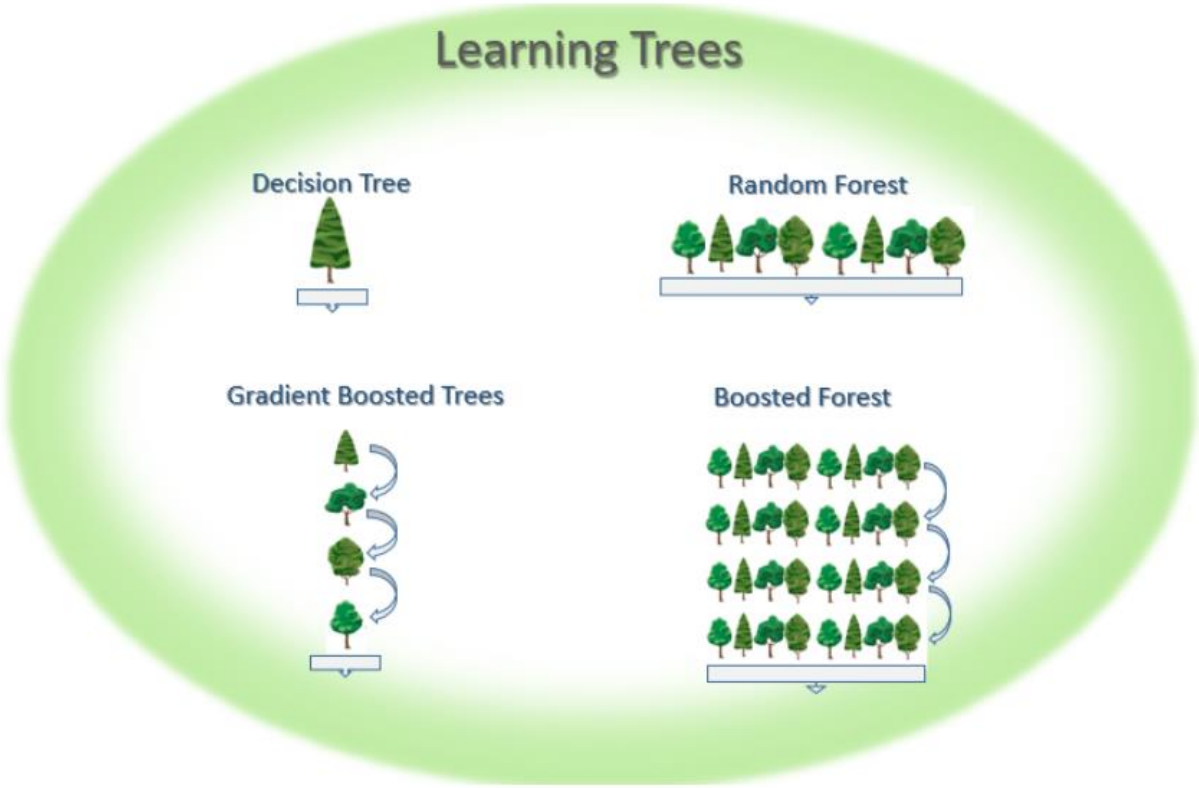
Altura	Diâmetro	Altitude	Pluviosidade	Idade
50	8	5000	12	54
56	9	4400	10	75
72	12	6500	18	60
47	10	5200	14	53

- Modelo de árvore de decisão:
  - Se altura for <= 50 e diametro <=10 e altitude >=5000 e pluviosidade >=12, então idade é 52

Métodos Ensemble

Métodos que utilizam múltiplos algoritmos de aprendizagem para obter um desempenho melhor do que seria obtido com o uso isolado do algoritmo.





(https://hpccsystems.com/blog/learning-trees-guide-to-decision-tree-based-machine-learning)

Fundação Vanzolini

Tutorial de Boosted Forest

Fluxo de aprendizagem de máquina

1. Definição do problema
2. Extração dos dados
3. Preparação dos dados
4. Conversão e segregação das variáveis
5. Treinamento e avaliação do modelo
6. Implantação do modelo

1. Definição do problema

“Dado um conjunto de atributos de uma propriedade (localização, metragem, ano de construção), como predizer o seu valor?"

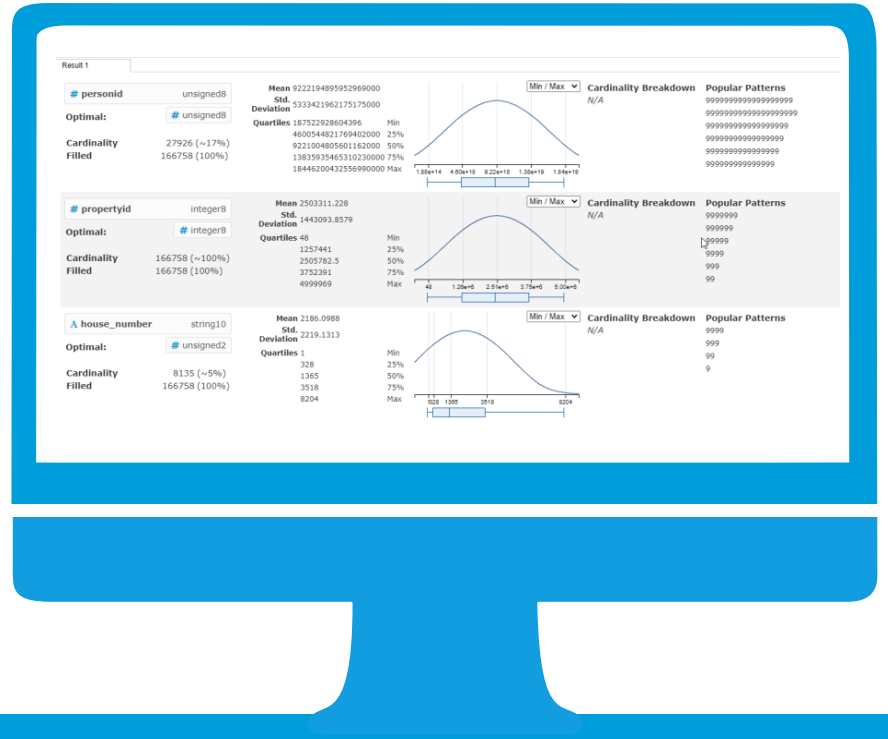
propertyid	house_numbe	house_nu	predir	street	streett	postdir	apt	city	state	zip	total_value	assessed_value	year_acquired	land_square_foot	living_square_fee	bedrooms	full_baths
828195	144			MCKIERNAN	DR			WALNUT CREEK	CA	94597	62614	62614	2006	20418	2485	3	2
1144455	281			CENTER	ST			BALTIMORE	MD	21136	105500	10550	2007	4807	1368	0	0
1494347	483			NEWTON	RD			FLAGSTAFF	AZ	86011	2220	2220	0	5654	1011	3	1
1910847	802			HATCHERY	CT			WOODLAND	WA	98674	356000	356000	0	6094	0	2	1
4267562	5007		E	ROY ROGERS	RD			TROY	MI	48085	327253	327253	2007	3484	0	3	0
4888602	7607			PEBBLESTONE	DR		000009	KERNVILLE	CA	93238	732179	732179	2010	19597	6132	6	6
48725	4			LONG	AVE			SUNRISE	FL	33323	271000	271000	2008	6880	2392	4	2
83528	6			TRILLUM	LN			WAYLAND	MA	02193	79889	79889	2007	7657	1657	4	1
94604	7			PARMENTER	AVE			PLYMOUTH	MN	55441	23800	23800	2005	19994	1754	3	2
220326	17			TIMBER	RD			LOS ANGELES	CA	90063	89000	89000	2008	7840	954	3	1
994609	212			FREYER	DR	NE		PHILOMONT	VA	20131	59800	59800	2009	11199	1241	3	0
1836173	724			EASTER	ST			ALLENTOWN	PA	18102	191600	191600	0	9100	2534	4	2
2910797	1903			SADDLE BROOK	DR			CLIO	CA	96106	61610	61610	2007	0	0	0	0
3083959	2158			RIVERSIDE	DR			UPPER MORELA…	PA	19006	90300	0	0	0	1235	3	2
3952189	4040			GRAND VIEW	BLVD		000054	RIO LINDA	CA	95673	0	0	0	2700720	0	0	0
4186238	4726			LAS PALMAS	CT			WAELDER	TX	78959	18816	18816	2009	2159	1320	0	0
4597143	6213			WILSON	RD			ZOLFO SPRINGS	FL	33890	72600	0	0	8496	0	3	1
4624905	6321			STONEWALL	LN			PATERSON	NJ	07514	139880	139880	2008	10454	1391	4	2
92326	7			KNOLLCREST	DR			NARANJA	FL	33032	76214	76214	2008	4800	930	2	0
1792852	704			ERIN	DR			TRABUCO	CA	92678	28010	28010	2007	5200	0	3	1
1843977	728		S	ARLINGTON HE…	RD			BLOOMING GRO…	TX	76626	130400	130400	2007	36154	1629	3	1
4214872	4821			MYRTLE OAK	DR		000025	SAN BERNARDI	CA	92376	33250	0	2007	93654	0	0	0

Desafio: Lending Club

Exercício prático:

Treine e avalie um modelo de análise de risco de pedido de empréstimo

- Considere a aplicação de aprendizagem supervisionada



Fundação Vanzolini

Até a próxima aula!!!