Big Data Analytics com R e Microsoft Azure Machine Learning

Classificação com Linguagem R e Azure Machine Learning

Seja Bem-Vindo(a)!

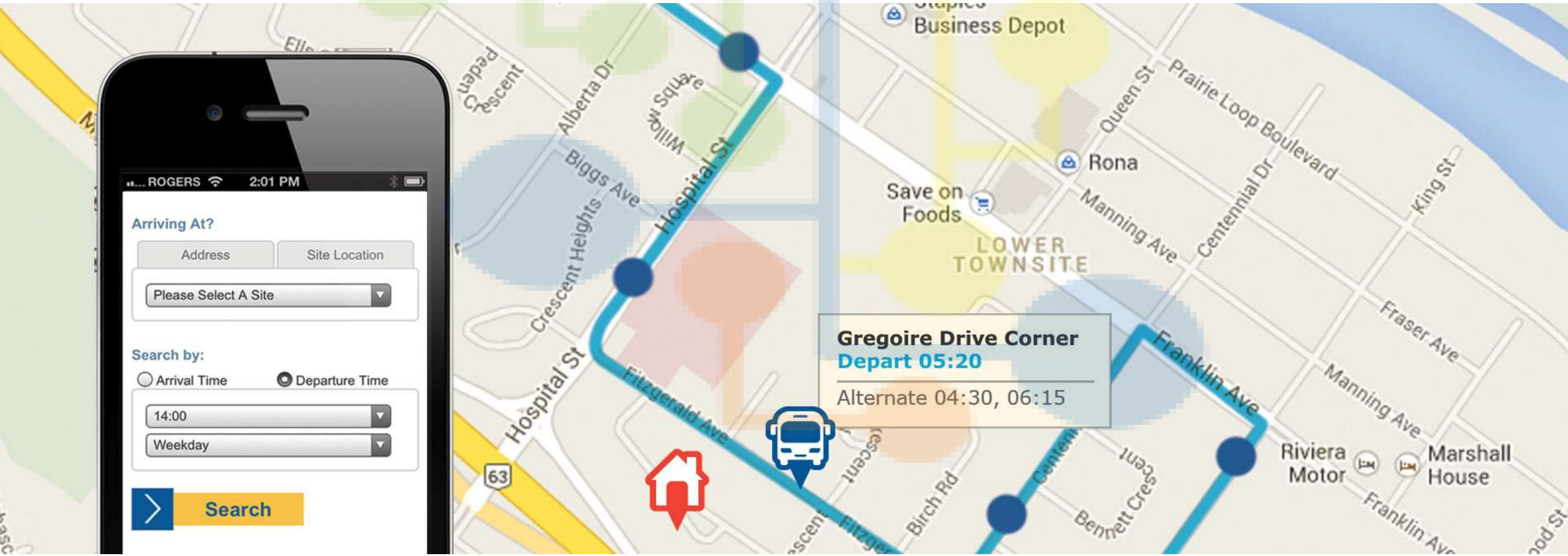
Classificação com Linguagem R e Azure Machine Learning

Hora de consolidar tudo que estudamos e colocar o conhecimento em prática!

O Que é Classificação?

Seja Bem-Vindo(a)!

O Que é Classificação?



O Que é Classificação?

O Que é Classificação?

Classificação

- Aprendizagem Supervisionada
- Classe de modelos para categorizar valores
- Métodos Two-class e Multi-class
- Erros são medidos pelas taxas de classificações incorretas
- Alguns erros podem ser mais críticos que outros e trade-offs terão que ser feitos

O Que é Classificação?

Performance dos Modelos de Classificação

Confusion Matrix

	Previstos	
Atuais	Sim	Não
Sim	True Positive (TP)	False Negative (FN)
Não	False Positive (FP)	True Negative (TN)

O Que é Classificação?

Medidas de Performance

Medida de Performance	Definição
Accuracy	Total de resultados corretos / Total de casos analisados
Recall	Total de resultados positivos / Total de resultados corretos
Precision	Proporção de ”true” / Total de resultados corretos
F-Score	F = 2 * TP / (2 * TP + FP + FN) (É o Balanceamento entre Precision e Recall)
AUC	AUC = Area Under the Curve. Plot de TP no eixo y e FP no eixo x

O Que é Classificação?

Medidas de Performance

	Previstos	
Atuais	Sim	Não
Sim	True Positive (TP)	False Negative (FN)
Não	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F-Score} = 2 * TP / (2 * TP + FP + FN)$$

Recomendações Sobre Otimização

Seja Bem-Vindo(a)!

Recomendações Sobre Otimização

Dicas Gerais:

- Diferentes conjuntos de variáveis
- Utilizar outros algoritmos
- Aplicar quantization a variáveis numéricas (transformá-las em variáveis categóricas)
- Otimizar os parâmetros dos algoritmos

Recomendações Sobre Otimização

Decisões de Negócio:

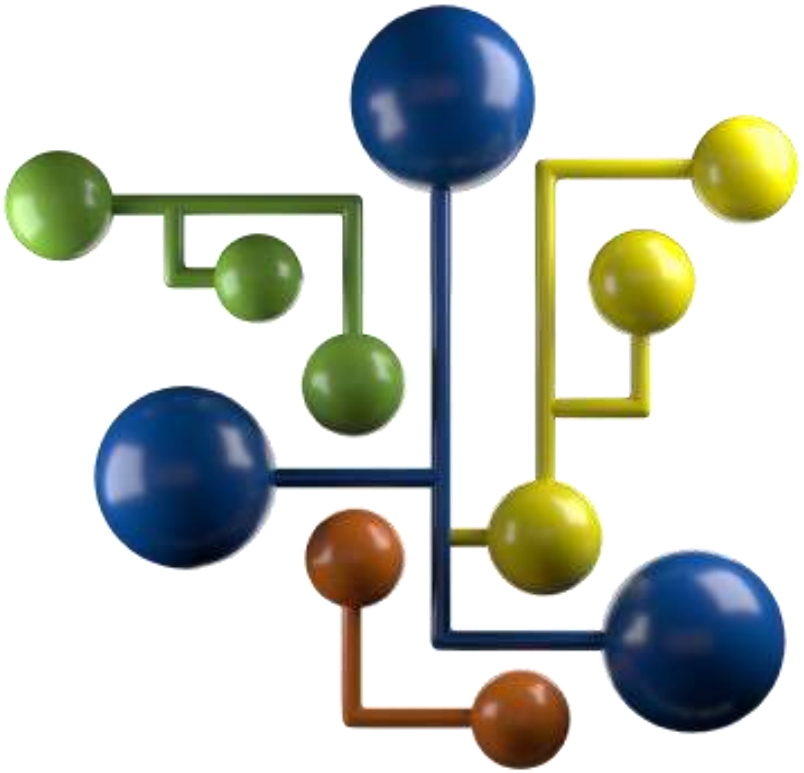
- Quais features (variáveis) são mais relevantes?
- Trade-off entre falsos positivos e falsos negativos
- O problema pode ser resolvido com esses dados?

Recomendações Sobre Otimização

Cada etapa do processo pode ser otimizada:

- Limpeza e Preparação de Dados
- Exploração dos Dados
- Feature Selection
- Testar e Avaliar o Modelo
- Otimizar o Modelo

Muito Obrigado por Participar!



Tenha uma Excelente Jornada de Aprendizagem.

Equipe Data Science Academy