UNIASSELVI

GRADUAÇÃO E PÓS

TURMA FLC15450-BDI

Tutor Wellington Costa

UNIASSELVI

#ADistanciaImporta

A **distância nunca foi tão importante** em um momento como esse. A **UNIASSELVI**, referência na metodologia EAD semipresencial, ajustou o seu modelo de ensino para que **todas as aulas e provas sejam totalmente virtuais.** Assim você não precisa sair de casa para estudar. ***Com atitudes como essa, você colabora para a integridade da saúde de todos.***

Ah, e fique atento: quando a situação se resolver, voltaremos para o nosso modelo semipresencial – a metodologia mais indicada para uma educação de qualidade. ***#ADistânciaImporta***

APRESENTAÇÃO

- Acadêmico:
  - Adilio de Sousa Farias
  - Matrícula: 3275870
- Disciplina: Projeto I - Aplicação de Métodos de Aprendizagem de Máquina (19370).
- Tema: Aplicação de Métodos de Aprendizagem de Máquina utilizando uma base com dados de clientes de um banco comercial.

UNIASSELVI
GRADUAÇÃO E PÓS

DESCRIÇÃO DO TEMA

- A Base de dados utilizada neste projeto foi uma amostra do banco de dados de clientes de um banco em formato CSV.
- Nesta aplicação o método escolhido foi de aprendizado supervisionado, a categoria aplicada será de classificação, onde será avaliada a performance de três algoritmos. São eles: **Decision Tree, RandomForest Classifier e Logistic Regression**.
- Assim determinando qual deles terá melhor performance e acuracidade de previsão sobre o objetivo proposto utilizando métodos de avaliação **F1 Score e Curva AUC**.

UNIASSELVI
GRADUAÇÃO E PÓS

OBJETIVOS

- Classificar o cliente do banco em faixas de risco de crédito para ajudar na tomada de decisão de emprestar dinheiro ou aumentar o limite de crédito de cartão do cliente.

UNIASSELVI
GRADUAÇÃO E PÓS

ESPECIFICAÇÃO TÉCNICA

- A base de dados está no formato CSV, sendo sua divisão em 11 colunas e 249.678 linhas.
- Features (colunas): ID, sexo, UF, n_dependentes, casa_propria, max_dias_atraso_6m, programa_reneg, escolaridade, status_civil, tempo_empregado, modelo_renda_v2, classe_renda_inf.
- Divisão da base de dados: 70% Train e 30% Test
- Tarefa de Aprendizado: Classificação.
- Modos de aprendizado: Supervisionado.
- Algoritmos Avaliados: Decision Tree, Random Forest e Logistic Regression.
- Métricas utilizadas: F1-Score Curva ROC.

UNIASSELVI
GRADUAÇÃO E PÓS

PREPARAÇÃO DOS DADOS

- Métodos de Pré-processamento: check de dados faltantes, remoção de informações duplicadas e/ou nulas, verificação de variância dos dados, e transformação de dados categóricos em numéricos para aplicar alguns métodos de Machine Learning.
- Mapeamos valores inteiros para colunas categóricas para não haver nenhum erro na aplicação dos algoritmos de classificação.
- Utilizamos toda a base de dados para realizar os testes, enquanto a partição será realizada em train/test.
- Foi separada a base de dados para teste e treino. Ficou definido 30% para teste e 70% para treino.

MÉTODOS DE MACHINE LEARNING

- Foram aplicados os seguintes métodos:
  - **Modelo Decision Tree**: As Árvores de Decisão, ou Decision Trees, estabelecem regras para tomada de decisão. O algoritmo criará uma estrutura similar a um fluxograma, com “nós” onde uma condição é verificada, e se atendida o fluxo segue por um ramo, caso contrário, por outro, sempre levando ao próximo nó, até a finalização da árvore.
  - **Modelo Random Forest**: o algoritmo Random Forest Classifier gera várias árvores de decisão de maneira aleatória, formando uma floresta, onde cada árvore será utilizada na escolha do resultado final, em uma espécie de votação para qual Decision Tree obteve a melhor performance.
  - **Modelo de Logistic Regression**: a regressão logística é uma técnica de análise de dados que usa matemática para encontrar as relações entre dois fatores de dados. Em seguida, essa relação é usada para prever o valor de um desses fatores com base no outro. A previsão geralmente tem um número finito de resultados, como sim ou não. Apesar de ter o nome "Regressão" ela é utilizada para tarefas com algoritmos de "Classificação".

RESULTADOS E DISCUSSÃO

» Depois da aplicação dos 3 métodos, a performance dos modelos escolhidos quanto a acurácia, foi a do Random Forest com os seguintes scores:

```
[148] # Verificando score de acurácia dos dados de treino
      cv_scores_rf['train_f1_macro'].mean()

      0.8154567525600896

[149] cv_scores_rf['train_roc_auc'].mean()

      0.8885975892440279

[150] # Verificando score de acurácia dos dados de teste
      cv_scores_rf['test_f1_macro'].mean()

      0.8154555550976236

[151] cv_scores_rf['test_roc_auc'].mean()

      0.8866860533141117
```

UNIASSELVI
GRADUAÇÃO E PÓS

RESULTADOS E DISCUSSÃO

» Após a otimização dos parâmetros conseguimos definir treinar o modelo otimizado e definir a função que define o Score do cliente com base em uma classificação de negócio do banco, ficando dessa forma:

```
[166] # Agora criamos uma função para classificar os clientes em Risco A, B ou C
      def classificar_cliente(score):
          if score < 300: return 'A'
          elif 300 <= score <= 600: return 'B'
          elif score > 600: return 'C'
      result['DECISION'] = result.SCORE.apply(classificar_cliente)
```

```
result
```

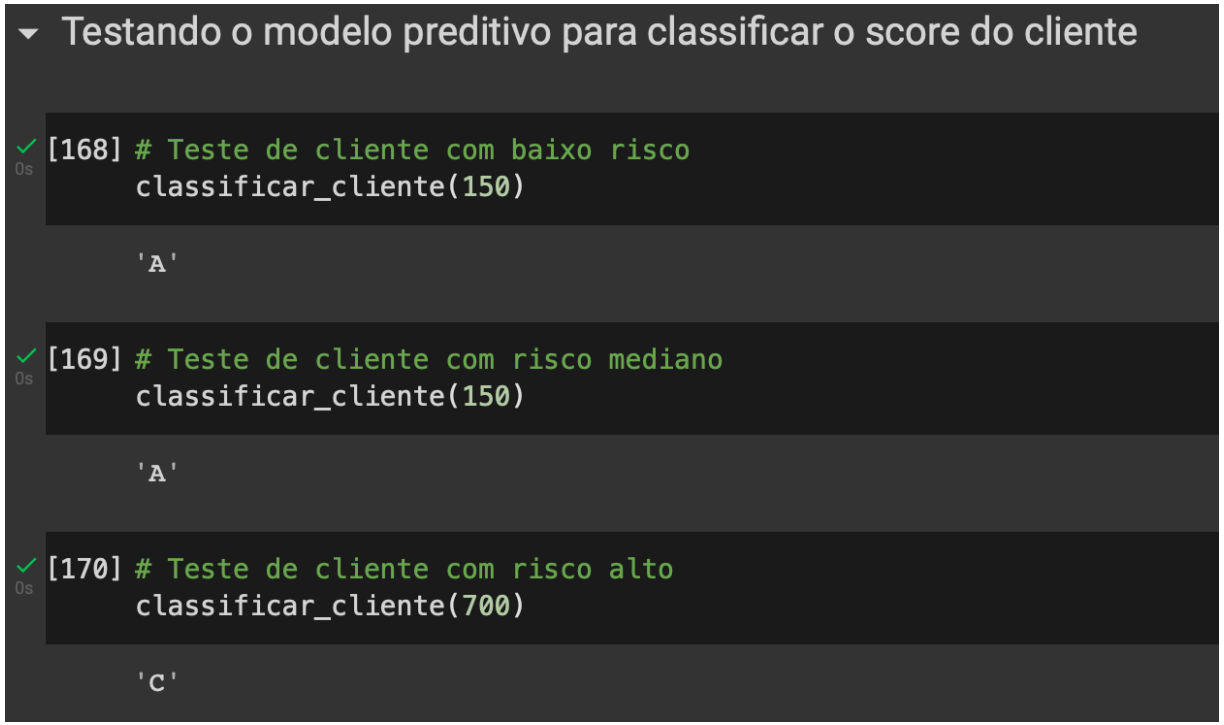
	id	SCORE	TARGET	DECIL	DECISION
0	45630330	102	0	1	A
1	56114485	600	0	5	B
2	32029382	615	0	8	C
3	97109191	969	1	8	C
4	28706788	117	0	2	A
...	...	...	...	...	...
74391	81523025	602	1	5	C
74392	10088950	132	0	3	A
74393	56613498	604	1	6	C
74394	26006330	112	1	1	A
74395	99525827	103	0	1	A

74396 rows × 5 columns

UNIASSELVI
GRADUAÇÃO E PÓS

RESULTADOS E DISCUSSÃO

» Por fim, realizamos o teste na função, fornecendo dados de scores dos clientes para que o classificador diga em que categoria de risco o cliente está para que o Banco tome uma decisão de negócio:



UNIASSELVI
GRADUAÇÃO E PÓS

CONSIDERAÇÕES FINAIS

» Apresente suas principais considerações acerca da pesquisa desenvolvida em seu Projeto.

- Como já explicado o modelo Random Forest foi o modelo que obteve melhor performance no projeto, após o balanceamento dos dados, e para tanto pode ser usado para análises preditivas considerando a mesma base de dados. No entanto cabe destacar que podem ser usados outros modelos, desde que previamente testados.
- Esse tipo de algoritmo já é bastante utilizado nos bancos públicos e privados no Brasil, cabendo aos cientistas de dados e engenheiros de machine learning darem manutenção e aprimorar por meio de featuring engineering esse modelos de classificação de risco.

UNIASSELVI
GRADUAÇÃO E PÓS

REFERÊNCIAS

- ESCOVEDO, T.; KOSHIYAMA, A. **Introdução a Data Science** – Algoritmos de Machine learning e métodos de análise. São Paulo: Casa do Código, 2020.
- HARRISON, M. **Machine learning** – Guia de Referência Rápida: Trabalhando com dados estruturados em Python. O'Reilly, Novatec, 2020.
- VANDERPLAS, J. **Python data science handbook**: Essential tools for working with data. O'Reilly Media, 2016.
- MITCHELL, T. M. et al. **Machine learning**. McGraw-Hill, 1997. Disponível em: http://profsite.um.ac.ir/~monsefi/machine-learning/pdf/Machine-Learning-Tom- -Mitchell.pdf. Acesso em: 2 out. 2022.
- GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow:** Concepts, Tools, and Techniques to Build Intelligent Systems. 2. ed. O'Reilly Media, 2019.
- ERBS, S. **Preparação e Análise Exploratória de Dados**. Indaial: UNIASSELVI, 2020.
- BRESSERT, E. SciPy and NumPy: an overview for developers. O'Reilly Media, 2012.
- KOERICH, A. **Aprendizagem de Máquina**. Apostila. Apresentação. Programa de Pós-Graduação em Informática, Pontifícia Universidade Católica do Paraná (PUCPR), 2012. Disponível em: http://www.ppgia.pucpr.br/~alekoe/AM/2012/ 0-Apresentacao-AM-2012.pdf. Acesso em: 20 set. 2022.
- PEDREGOSA, F. et al. **Scikit-learn: Machine learning in Python**. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011. Disponível em: https://www.jmlr. org/papers/volume12/pedregosa11a/pedregosa11a.pdf. Acesso em: 3 nov. 2020.

UNIASSELVI
GRADUAÇÃO E PÓS