INTERFACE TECNOLÓGICA

CIÊNCIA DE DADOS APLICADO À LOGÍSTICA

DATA SCIENCE APPLIED TO LOGISTICS

Lucas Confortini Batista –lucas.batista12@fatec.sp.gov.br
Faculdade de Tecnologia de Taquaritinga (Fatec) – Taquaritinga – SP – Brasil

Marcus Rogério de Oliveira – marcusrogerio@gmail.com
Faculdade de Tecnologia de Taquaritinga (Fatec) – Taquaritinga – SP – Brasil

DOI: 10.31510/infa.v19i1.1397
Data de submissão: 10/03/2022
Data do aceite: 25/05/2022
Data da publicação: 30/06/2022

RESUMO

A quantidade de dados cresce de maneira exponencial e é importante que a partir destes possam ser abstraídos conhecimentos e informações com intuito de gerar vantagens competitivas. Em vista disso, a ciência de dados surgiu com o intuito de extrair essas informações e interpretá-las com modelos matemáticos, estatísticos e algoritmos de inteligência artificial. Diante disso, o presente artigo com a utilização de modelos matemáticos e estatísticos tem por objetivo a aplicação de modelos de regressão na logística, a fim de obter predição de custos. Dentro desse contexto, este artigo apresenta a atuação da ciência de dados aplicada à área de logística, com introdução a termos, métodos e experiência prática, enfatizando a partição de predição de custos. A metodologia proposta abrange inicialmente uma pesquisa bibliográfica descritiva e posteriormente, analisa mais de 30 mil coeficientes aplicados a diferentes modelos de regressão. Os resultados permitem identificar a influência dos coeficientes no lucro líquido e a comparação de precisão dos modelos de regressão aplicados à predição de custos.

Palavras-chave: Ciência. Dados. Regressão. Logística. Predição.

ABSTRACT

The amount of data grows exponentially and it is important that knowledge and information can be abstracted from them, in order to generate competitive advantages. In view of this, data science emerged with the aim of extracting this information and interpreting it with mathematical, statistical models and artificial intelligence algorithms. In this way, this article, with the use of mathematical and statistical models, aims to apply regression models in logistics, in order to obtain cost prediction. Within this context, this article presents the performance of data science applied to the area of logistics, with an introduction to terms, methods and practical experience, emphasizing the partition of cost prediction. The proposed methodology initially covers a descriptive bibliographic research and later, analyzes more than 30 thousand coefficients applied to different regression models. The results allow identifying the influence of coefficients on net income and comparing the accuracy of regression models applied to cost prediction.

Keywords: Data. Science. Regression. Logistics. Prediction.

1. INTRODUÇÃO

Em um mundo cada vez mais centralizado em torno da tecnologia da informação, enormes quantidades de dados são produzidos e armazenados todos os dias (Nelli, 2015).

Os dados, na verdade, não são informações, pelo menos em termos de sua forma. No fluxo sem forma de bytes, à primeira vista é difícil entender sua essência, se não estritamente o número, palavra, ou tempo que eles relatam. A informação é, na verdade, o resultado do processamento, que levando em conta um determinado conjunto de dados, extrai algumas conclusões que podem ser utilizadas de várias maneiras. Esse processo de extração informação dos dados brutos é precisamente a análise de dados (Nelli, 2015).

O objetivo da análise de dados é justamente extrair informações que não são facilmente dedutíveis, mas que, quando compreendido, leva à possibilidade de realizar estudos sobre os mecanismos dos sistemas que os produziu, permitindo assim a possibilidade de fazer previsões de possíveis respostas desses sistemas e sua evolução no tempo (Nelli, 2015).

A análise de dados tornou-se uma verdadeira disciplina que conduz ao desenvolvimento de modelos reais de geração de metodologias. O modelo é, na verdade, a tradução para uma forma matemática de um sistema. Uma vez que há um raciocínio matemático, lógico ou uma forma capaz de descrever as respostas do sistema sob diferentes níveis de precisão, pode-se então fazer previsões sobre seu desenvolvimento ou resposta a certas entradas. Assim, o objetivo da análise de dados não é o modelo, mas a bondade de seu poder preditivo (Nelli, 2015).

O poder preditivo de um modelo depende não apenas da qualidade das técnicas de modelagem, mas também na capacidade de escolher um bom conjunto de dados sobre o qual construir toda a análise de dados (Nelli, 2015).

A ciência de dados envolve princípios, processos e técnicas para entender fenômenos por meio da análise (automatizada) de dados (Provost e Fawcett, 2013).

Ciência de dados é um termo cada vez mais utilizado para designar uma área de conhecimento voltada para o estudo e a análise de dados, onde busca-se extrair conhecimento e criar novas informações. É uma atividade interdisciplinar, que concilia principalmente duas grandes áreas: Ciência da Computação e Estatística. [...](Guerra *et al*, 2018).

A ciência de dados é um amálgama de métodos analíticos com o objetivo de extrair informações dos dados. Esta descrição também se encaixa nas estatísticas e mineração de dados (Steele *et a*l, 2016).

2. FUNDAMENTAÇÃO TEÓRICA

O artigo vigente tem como objetivo desenvolver um modelo capaz de realizar a predição de custos para logística.

Para associar a pesquisa e a busca por um resultado que se encaixe na expectativa do objetivo do trabalho, é preciso estabelecer uma base conceitual. Para isso, os tópicos a seguir apresentam a fundamentação teórica. Além de servir como base para esclarecimento aos leitores e muitos conceitos que facilitarão o entendimento da pesquisa.

2.1 Ciência de Dados

A Ciência de Dados é a área da computação responsável pelos modelos matemáticos, estatísticos e de inteligência artificial, os quais são aplicados para extração de informação de uma determinada quantidade de dados (Gorelik, 2019).

Não existe apenas uma forma de estruturar e aplicar os conhecimentos da Ciência de Dados. A forma de aplicação varia bastante conforme a necessidade do projeto ou do objetivo que se busca alcançar (Guerra *et al*, 2018).

No centro desta tecnologia está uma combinação de matemática (especificamente estatística), ciência da computação (especialmente manipulação de dados e aprendizado de máquina) e domínio ou conhecimento de negócios. O conhecimento do domínio ou de negócios é crucial para o cientista de dados entender quais problemas precisam ser resolvidos, quais dados são relevantes e como interpretar os resultados (Gorelik, 2019).

A Ciência de Dados localiza-se na interseção da ciência da computação, estatística e domínios de aplicação substantivos. Da ciência da computação vem o aprendizado de máquina e as tecnologias de computação de alto desempenho para lidar com a escala. Das estatísticas vem uma longa tradição de análise exploratória de dados, teste de significância e visualização (Skiena, 2017).

O Big Data está cada vez mais presente, mais difundido e mais importante. Nesta tecnologia há um enorme potencial, como melhor compreensão dos problemas, oportunidades

para prever - e até mesmo para moldar - o futuro. A Ciência de Dados é o principal meio para descobrir e explorar esse potencial. Esta área fornece maneiras de lidar e se beneficiar do Big Data: ver padrões, descobrir relacionamentos e dar sentido a imagens e informações incrivelmente variadas (EMC, 2015).

Os desafios colocados pelo que é frequentemente chamado de Big Data decorrem da diluição predominante de conteúdo, volume e a escassez de projeto e controle. A ciência de dados se desenvolveu em resposta à oportunidade e necessidade de criar valor a partir desses dados (Steele *et al*, 2016).

A ideia da Ciência de Dados é fazer recomendações sobre ações a serem tomadas com base em informações factuais representadas como dados (Gorelik, 2019).

2.2 Ferramentas de Ciência de dados

A utilidade do Python para a ciência de dados decorre principalmente da grande variedade de bibliotecas, por exemplo: NumPy para manipulação de dados homogêneos baseados em array; Pandas para manipulação de dados heterogêneos e rotulados; SciPy para tarefas comuns de computação científica; Matplotlib para visualizações de qualidade de publicações e, Scikit-Learn para aprendizado de máquina, dentre outras (VanderPlas, 2017).

NumPy (abreviação de Numerical Python) fornece uma interface eficiente para armazenar e operar em buffers de dados densos. De certa forma, os arrays NumPy são como o tipo de lista interno do Python, mas os arrays NumPy fornecem armazenamento e operações de dados muito mais eficientes à medida que os arrays aumentam de tamanho. As matrizes NumPy formam o núcleo de quase todo o ecossistema de ferramentas de ciência de dados em Python (VanderPlas, 2017).

Pandas fornece um objeto DataFrame, o qual é uma estrutura construída em matrizes NumPy que oferece uma variedade de funcionalidades de manipulação de dados. DataFrames são essencialmente arrays multidimensionais com rótulos de linha e coluna anexados e, muitas vezes, com tipos heterogêneos e/ou dados ausentes. Além de oferecer uma interface de armazenamento conveniente para dados rotulados, o Pandas implementa várias operações de dados poderosas familiares aos usuários de estruturas de banco de dados e programas de planilha (VanderPlas, 2017).

Matplotlib é uma biblioteca de visualização de dados multiplataforma construída em arrays NumPy, ou seja, é responsável pela geração de gráficos. Um dos recursos do Matplotlib

é sua capacidade de funcionar bem com muitos sistemas operacionais e back-ends gráficos. O Matplotlib suporta dezenas de back-ends e tipos de saída, o que significa que ele funciona independentemente de qual sistema operacional seja utilizado ou qual formato de saída seja necessário. Essa abordagem multiplataforma, tem sido uma das vantagens do Matplotlib (VanderPlas, 2017).

Outras bibliotecas importantes para a ciência de dados com Python são o XGBoost e o LightGBM. Ambas implementam algoritmos de aprendizado de máquina sob a estrutura Gradient Boosting (Aumento de gradiente).

O Gradient Boosting (Aumento de gradiente) funciona construindo árvores de forma serial, onde cada árvore tenta corrigir os erros da anterior. Este modelo engloba o GBDT (Gradient boosted decision trees - Árvores de decisão impulsionadas por gradiente) também conhecido como GBM (Gradient boosting machine - Máquina de aumento de gradiente), no qual as árvores impulsionadas por gradiente geralmente usam árvores muito rasas, de profundidade de um a cinco, o que torna o modelo menor em termos de memória e torna as previsões mais rápidas (Müller e Guido, 2017).

O XGBoost executa modelos com aumento de gradiente que são escaláveis e aprende computação paralela e distribuída rápida sem sacrificar a eficiência da memória. Não só isso, mas é um aprendiz de conjunto. O XGBoost usa o boosting para aprender com os erros cometidos nas árvores anteriores (Nokeri, 2022).

Algoritmo LightGBM (Light Gradient Boosting Machine) é um algoritmo do tipo GBDT (Gradient Boosting Decision Tree) e geralmente é usado em classificação, ordenação, regressão e suporta treinamento paralelo. O algoritmo usa árvores e funções de perda que permite aproximações sucessivas em cada etapa das iterações. Em seguida, ele treina uma árvore de decisão para minimizar a aproximação de segunda ordem (Minastireanu e Mesnita, 2019).

2.3 Modelos estatísticos e aprendizado de máquina

A seguir serão apresentados alguns exemplos de modelos estatísticos e suas aplicações.

O modelo preditivo da ciência de dados requer técnicas de regressão, as quais formam a base da análise preditiva. Estas técnicas buscam criar uma equação matemática, a qual servirá de matriz para apresentar as interações entre as diferentes variáveis. Dependendo das

circunstâncias, diferentes técnicas de regressão podem ser usadas para realizar análises preditivas (Zhang, 2017).

Alguns exemplos de técnicas de regressão são Linear Regression (Regressão Linear), Robust Regression (Regressão Robusta) e Ridge Regression (Regressão de cume).

O modelo de regressão linear calcula a relação entre as variáveis dependentes e independentes usando uma linha reta (linha de regressão). É normalmente em forma de equação. A regressão linear pode ser usada onde a variável dependente tem um intervalo ilimitado (Zhang, 2017).

A regressão robusta surgiu com o intuito de eliminar parâmetros que causam mudanças significativas nas estimativas. Uma abordagem comum é usar uma métrica alternativa que seja menos sensível a grandes valores discrepantes. Uma forma de realizar essa métrica é utilizando a função de Huber, a qual usa os resíduos quadrados quando são “pequenos” e a diferença simples entre os valores observados e previstos quando os resíduos estão acima de um limite (Kuhn e Johnson, 2013).

Uma vez que uma regressão pode produzir coeficientes muito grandes a regressão de cume é aplicada. Sua função é restringir o crescimento incontrolável dos coeficientes de regressão aplicando uma penalidade, ou seja, é adicionado um parâmetro de complexidade para suprimir os coeficientes (Zinoviev, 2016).

O aprendizado de máquina é um método de análise de dados que automatiza a construção de modelos analíticos possui a capacidade de treinar sistemas ou algoritmos para obter informações de um conjunto de dados. Existem muitos tipos diferentes de aprendizado de máquina, e eles variam de redes neurais ao modelo de regressão (Zhang, 2017).

3 PROCEDIMENTOS METODOLÓGICOS

O presente artigo apresenta como característica o desenvolvimento baseado em pesquisa de caráter quantitativo, mensurando os critérios para a análise de um procedimento experimental de uma rede de transporte de produtos primários visando os efeitos de diversas variáveis para a viagem com a finalidade de estipular a predição de custos.

3.1 Aplicação do Modelo na logística

A aplicação da pesquisa experimental irá considerar as variáveis coletadas em uma pesquisa de campo e realizar a geração de dados para a predição de custos, na qual o nome da empresa e informações de registros não podem ser citados devido a questões de segurança.

A coleta de variáveis foi feita com o gerente da empresa, o qual disponibilizou os dados básicos para a realização de uma viagem, a seguir serão apresentadas as variáveis com suas respectivas descrições.

Tabela 1 - Planilha de variáveis

Sigla	Descrição	Cálculo
VF	Valor do frete por tonelada	Não
VO	Valor do óleo diesel	Não
DP	Distância percorrida em quilômetros	Não
P	Pedágio cobrado por eixo	Não
E	Quantidade de eixos por caminhão	Não
VP	Valor total de pedágio	VP = P * E
CT	Capacidade da carga do caminhão transportada	Não
PC	Porcentagem da comissão do motorista	Não
GM	Gasto médio por quilômetro	Não
GT	Gasto total de combustível	GT = (DP/ GM) * VO
DE	Data de embarque carga	Não
DV	Duração da viagem	Não
TP	Tipo do produto transportado	Não
CM	Comissão motorista	CM = (VF*CC)*PC
LL	Lucro líquido	LL= (VF*CT)-(GT+CM+VP)

Fonte: Autor (2022)

Para iniciar a aplicação foi instalada a API python hvplot, responsável pela plotagem dos dados, a função desta será gerar gráficos baseados nos cálculos realizados. Em seguida a importação das bibliotecas necessárias para o treinamento dos modelos e cálculos estatísticos.

```
>> !pip install hvplot
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import hvplot.pandas
%matplotlib inline
```

A seguir os registros contidos no csv, o qual possui dados relacionados à logística, são carregados e lidos. Nessa etapa as tuplas com valores nulos são eliminados.

```
df = pd.read_csv('/content/frota-3anos-0.zip',compression='zip')
df = df.dropna()
```

Com a utilização da biblioteca sklearn do python as colunas com valores string serão caracterizadas numericamente. No contexto deste trabalho as colunas data de embarque e tipo do produto recebem categorias numéricas para auxiliar na interpretação. Posteriormente a tabela é convertida para uma matriz a qual é percorrida e são identificadas as colunas correlatas, o intuito desta etapa é eliminar as colunas que não possuem influência no resultado final. Após essa análise as colunas correlatas são excluídas.

```
correlation_matrix = encoded_df.corr()
for i in range(len(correlation_matrix.columns)):
    for j in range(i):
        if abs(correlation_matrix.iloc[i, j]) > 0.08:
            colname = correlation_matrix.columns[i]
            correlated_features.add(colname)
encoded_df.drop(labels=correlated_features, axis=1, inplace=True)
```

É importante ressaltar que a tabela df representa a versão original do csv, enquanto a tabela encoded_df é uma versão otimizada para uso nos cálculos. A seguir são criados dois novos data sets a partir dos anteriores, ambos são divididos para treino e teste.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

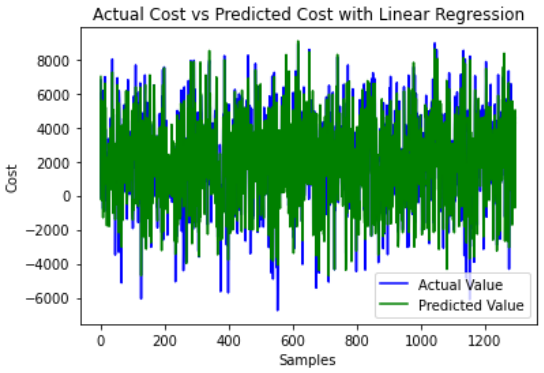
A partir desses primeiros estágios inicia-se a predição baseada nos modelos de regressão. O primeiro modelo executado foi o de Regressão linear.

```
from sklearn.linear_model import LinearRegression
lin_reg = LinearRegression(normalize=True)
lin_reg.fit(X_train,y_train)
```

Com base no resultado é realizada a predição e o gráfico é gerado.

```
pred = lin_reg.predict(X_test)
```

Gráfico 1 - Gráfico do custo atual vs predição de custo com Regressão Linear



Fonte: Autor (2022)

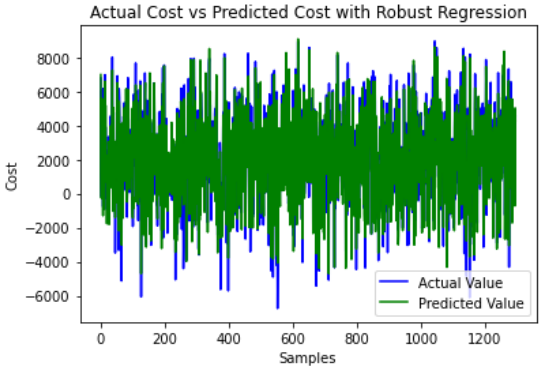
Posteriormente calculam-se as métricas de avaliação da qualidade da precisão do modelo, as quais são responsáveis pelo poder comparativo, pois a partir desse resultado será possível identificar o modelo mais preciso.

As métricas de avaliação consistem em erro absoluto médio (*Mean Absolute Error - MAE*), erro quadrado médio (*Mean Squared Error - MSE*), erro quadrático médio (*Root Mean Squared Error - RMSE*) e raiz quadrada (*Root squared - R²*).

O segundo modelo executado foi o de Regressão Robusta.

```
from sklearn.linear_model import RANSACRegressor
model = RANSACRegressor(base_estimator=LinearRegression(),
max_trials=100)
model.fit(X_train, y_train)
list_Y_test = list(y_test)
list_pred = list(pred)
```

Gráfico 2 - Gráfico do custo atual vs predição de custo com Regressão Robusta

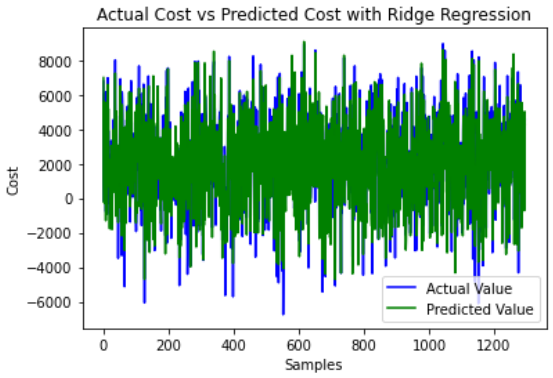


Fonte: Autor (2022)

Por fim, o último método executado foi o Regressão de Cume.

```
from sklearn.linear_model import Ridge
model = Ridge(alpha=100, solver='cholesky', tol=0.0001,
random_state=42)
model.fit(X_train, y_train)
pred = model.predict(X_test)
```

Gráfico 3 - Gráfico do custo atual vs predição de custo com Regressão de Cume



Fonte: Autor (2022)

4 RESULTADOS E DISCUSSÃO

Com base nos resultados obtidos, houve a comparação dos coeficientes de influência das variáveis no lucro líquido.

Figura 2 - Correlação dos coeficientes

de	tp	vf	vo	dp	vp	ct	gm	dv
-1.690508	6.670925	1828.021835	-553.542115	-1004.160801	-574.123953	950.962995	702.662203	-0.154062

Fonte: Autor (2022)

A partir dos resultados obtidos foi possível, além da predição de custos aplicada à logística, uma comparação dos três modelos aplicados anteriormente, na qual considerando-se a análise das métricas e busca por menores índices, o modelo de Regressão Linear mostrou-se mais eficiente.

Figura 3 - Métricas de avaliação de precisão

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Linear Regression	481.524936	414750.867213	644.011543	0.940985	0.941158
1	Robust Regression	481.485857	470008.759789	685.571849	0.933122	0.937688
2	Ridge Regression	485.888670	427774.976675	654.045088	0.939132	0.941158

Fonte: Autor (2022)

5 CONSIDERAÇÕES FINAIS

A finalidade deste artigo envolveu apresentar a atuação da ciência de dados no contexto logístico usufruindo de princípios, processos e técnicas para ampliar o processo preditivo e a partir do desenvolvimento e descrições dos acontecimentos foi possível elaborar uma análise comparativa entre alguns modelos.

É importante ressaltar que nem todos os modelos preditivos existentes foram comparados, entretanto, os modelos apresentados possuem maior aplicação entre as empresas devido a profissionais e consultorias manipularem os registros de avaliação da métrica, e serem as modelagens mais abordadas e plausíveis de intervalos ilimitados (Zhang, 2017).

Como síntese da conclusão do experimento nota-se que o nível de correlação do valor do frete é o principal fator determinante no custo da viagem, a capacidade da carga transportada e o gasto médio por quilômetros influenciam taxativamente no valor da viagem, corroborando para decisões estratégicas ao gerenciamento logístico.

REFERÊNCIAS

1. Apresenta a documentação da extensão SciPy. Disponível em: <**https://docs.scipy.org/doc/scipy/tutorial/general.html**>. Acesso em 22 fevereiro de 2022

2. Apresenta a documentação das bibliotecas do Python. Disponível em: <**https://docs.python.org/3/library/index.html**>. Acesso em 22 fevereiro de 2022

3. Apresenta a documentação do LightGBM. Disponível em: <**https://lightgbm.readthedocs.io/en/latest/index.html**>. Acesso em 22 fevereiro de 2022

4. Apresenta a documentação do XGBoost. Disponível em: <**https://xgboost.readthedocs.io/en/stable/index.html**>. Acesso em 22 fevereiro de 2022

5. Crickard Paul. **Data Engineering with Python: Work with massive datasets to design data models and automate data pipelines using Python**. Packt, 2020.

6. Densmore James. **Pipelines Pocket Reference Moving and Processing Data for Analytics**. O'REILLY, 2021.

7. EMC Education Services. **Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data**. WILEY, 2015.

8. Gorelik Alex. **The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science**. O'Reilly, 2019.

9. Guerra Saulo, Oliveira Paulo Felipe e McDonnell Robert. **Ciência de Dados com R Introdução**. IBPAD, 2018.

10. Kuhn Max e Johnson Kjell. **Applied Predictive Modeling**. Springer, 2013.

11. Kukreja Manoj. **Data Engineering with Apache Spark, Data Lake, and Lakehouse: Create scalable pipelines that ingest, curate, and aggregate complex data in a timely and secure way**. Pack, 2021.

12. Mailund Thomas. **Beginning Data Science in R Data Analysis, Visualization, and Modeling for the Data Scientist**. APRESS, 2017.

13. Manyika James, Chui Michael, Brown Brad, Bughin Jacques, Dobbs Richard, Roxburgh Charles, e Hung Byers Angela. **Big data: The next frontier for innovation, competition, and productivity**. McKinsey Global Institute, 2011.

14. Minastireanu E. A. e Mesnita G. .**Light GBM Machine Learning Algorithm to Online Click Fraud Detection**. IBIMA, 2019.

15. Müller C. Andreas e Guido Sarah. **Introduction to Machine Learning with Python**. O'REILLY, 2017.

16. Nelli Fabio. **Python Data Analytics, Data Analysis and Science Using Pandas, matplotlib, and the Python Programming Language**. APRESS, 2015.

17. Nokeri C. Tshepo. **Data Science Solutions with Python: Fast and Scalable Models Using Keras, PySpark MLlib, H2O, XGBoost, and Scikit-Learn**. APRESS, 2022.

18. Provost Foster e Fawcett Tom. **Data Science for Business What you need to know about Data Mining and Data-Analytic Thinking**. O'REILLY, 2013.

19. Provost Foster, Fawcett Tom. **Data Science for Business: What you need to know about data mining and data-analytic thinking**. O'Reilly, 2013.

20. Skiena S. Steven . **The Data Science Design Manual**. Springer, 2017.

21. Steele Brian, Chandler John e Reddy Swarna. **Algorithms for Data Science**. Springer, 2016.

22. VanderPlas J. **Python Data Science Handbook**. O'REILLY, 2017.

23. Zhang Arthur. **Data Analytics Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life**. 2017.

24. Zinoviev Dmitry. **Data Science Essentials in Python Collect → Organize → Explore → Predict → Value**. The Pragmatic Programmers, LLC, 2016.