Avaliação de Modelos de Inteligência Computacional para Classificação de Ativos de Renda Variável do Mercado Financeiro Brasileiro Utilizando Análise Fundamentalista

Aluno: Amadeu Chacar
Curso: Bacharelado em Engenharia de Computação
Orientador: Prof. D.Sc. Fábio Duncan de Souza
Coorientador: Prof. D.Sc. Fernando Luiz de Carvalho e Silva
Campos dos Goytacazes, novembro de 2022.

Sumário

1. Introdução

- Aumento de 92% no número de investidores em 2020 com relação a 2019.
- Investidores iniciantes.
    - Internet.
    - Longo prazo.
- Estudo de ativos, análise fundamentalista e seus indicadores.
- Grande quantidade de ativos disponíveis.
    - Mais de 400 em 2021.

1. Introdução

- Grande quantidade de dados disponíveis.
- Crescimento do número de produtos baseados em inteligência artificial.
    - Mais da metade das patentes entre 2013 e 2019, desde 1956.
- Aplicação da inteligência artificial no mercado financeiro.
- Utilização de I.A. para auxílio em análise fundamentalista.
- Alto custo de esforço e tempo para análise manual.

1.1. Justificativa

- Grande quantidade de empresas para um iniciante analisar individualmente.
- Redução de custo de esforço e tempo.
- Busca no Periódicos CAPES.
    - *Machine learning* e análise fundamentalista.
    - Trabalhos utilizando análise técnica.
- Entendimento do comportamento de modelos de I.C. neste contexto.

1.2. Objetivos

- Propor uma avaliação de desempenho de modelos de *machine learning* para pré-seleção de ativos do mercado financeiro brasileiro.
    - Utilização de dados fundamentalistas (balanços patrimoniais, DREs e indicadores).
    - Visando uma posterior utilização por investidores iniciantes.

1.2.1. Objetivos Específicos

- Desenvolver modelos de inteligência computacional para pré-selecionar empresas do mercado financeiro brasileiro utilizando como dados os balanços patrimoniais e indicadores fundamentalistas de cada ativo.
- Comparar o desempenho dos modelos e definir quais possuem melhor performance.
- Realizar simulações com cada modelo a fim de identificar se as empresas selecionadas geraram lucro.

2. Fundamentação teórica

- Introdução ao mercado financeiro.
- Mineração de dados e inteligência computacional.

2.1. Introdução ao Mercado Financeiro

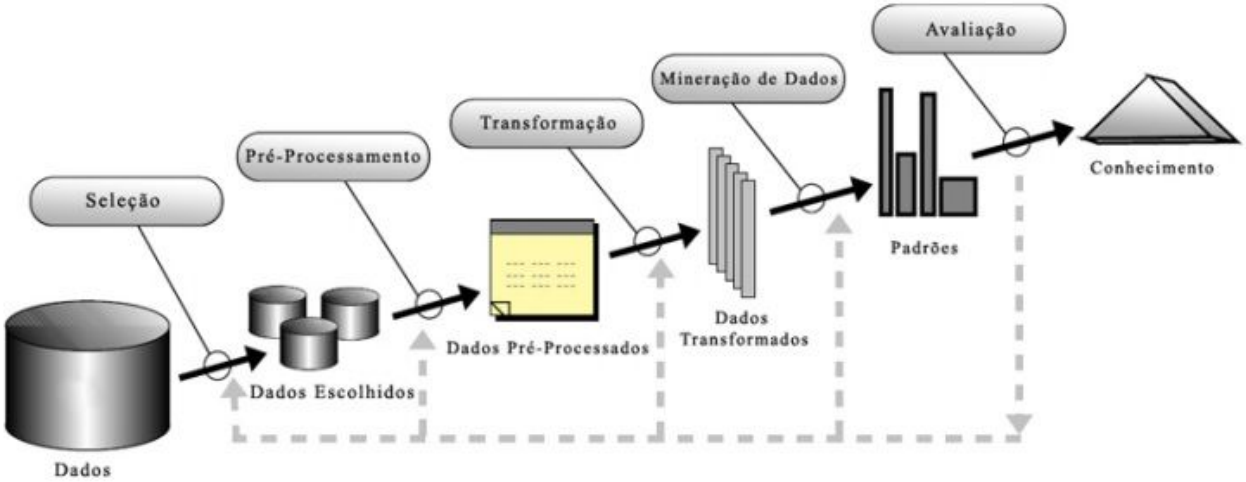
- Bolsa de valores e ações.
- Características de uma ação.
- Métodos de análise financeira.
    - Análise técnica.
    - Análise fundamentalista.

2.1. Introdução ao Mercado Financeiro

- Análise fundamentalista.
  - Longo prazo.
  - Fundamentos econômicos do mercado, balanços patrimoniais, DREs, etc.
  - Análise qualitativa
    - Qualidade da gestão e governança.
  - Análise quantitativa.
    - Saúde financeira, indicadores fundamentalistas.
      - P / L, P / VPA, DY, etc.

2.2. Tecnologias Aplicadas

- Mineração de dados.
    - KDD - *Knowledge Discovery in Databases*

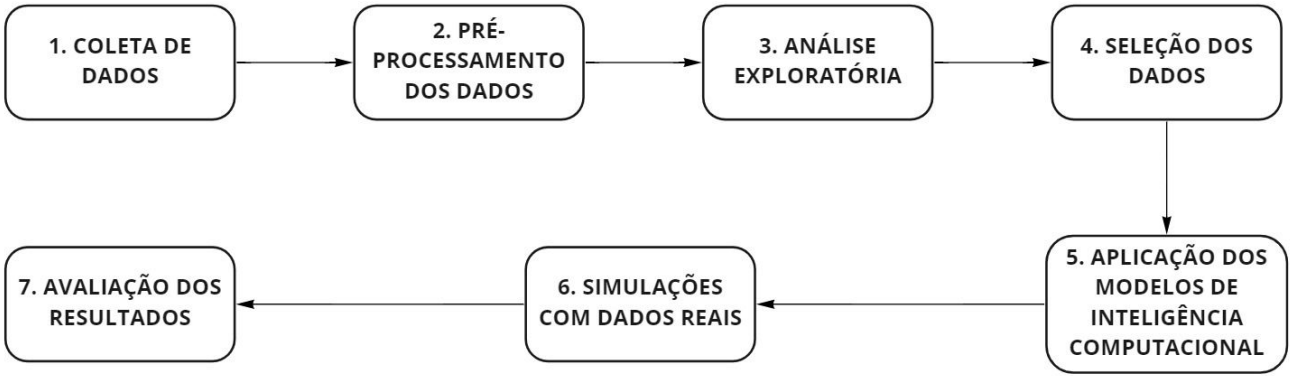


2.2. Tecnologias Aplicadas

- Inteligência computacional.
    - Aplicação de técnicas humanas e da natureza para resolução de problemas.
    - Velocidade da resolução de problemas.
    - Variedade de modelos.

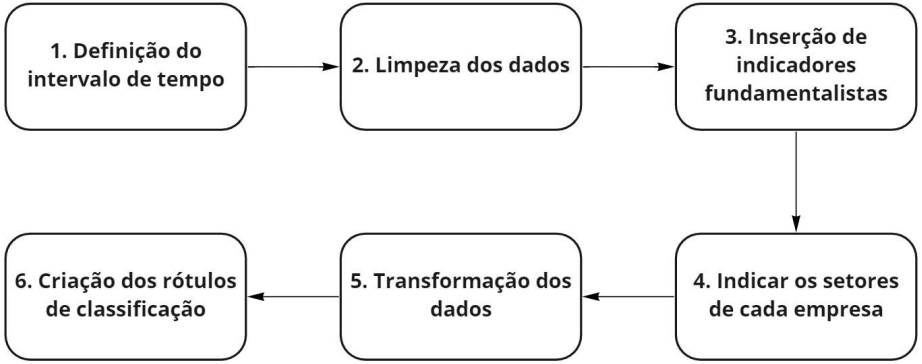
3. Metodologia

- Etapas para realizar a avaliação dos modelos de *machine learning* no contexto do trabalho.



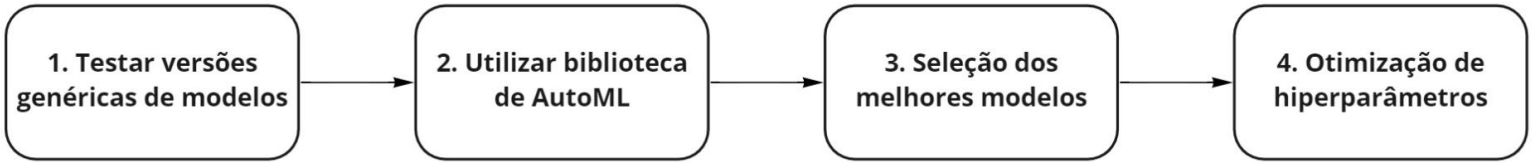
3. Metodologia

- Pré-processamento dos dados.



3. Metodologia

- Aplicação dos modelos de *machine learning*.



4. Desenvolvimento

- Aplicação dos passos explicitados na metodologia.
- Python.
  - Pandas
  - NumPy
- Jupyter Notebook.

4.1. Coleta de Dados

- Coleta manual em agosto de 2021.
- Fundamentus.
  - Balanços patrimoniais e DREs.
- Yahoo! Finanças.
  - Histórico de cotações.
- Total de 898 ativos coletados.

4.1. Coleta de Dados

YDUQ3	30/06/2021	31/03/2021	31/12/2020	30/09/2020	30/06/2020	31/03/2020	31/12/2019	30/09/2019	30/06/2019
Ativo Total	9641204.736	9642667.008	9265265.664	9411001.344	9294304.256	7635795.968	5512492.032	5740264.96	5564189.184
Ativo Circulante	3075841.024	3067963.904	2736397.056	2994115.072	3047205.12	3332005.12	1475683.968	1664539.008	1618936.96
Caixa e Equivalentes de Caixa	1212306.944	1278475.008	28407	25176	20415	10392	12251	10071	19436
Aplicações Financeiras	755078.016	771812.992	1604868.992	1895389.056	1886955.008	2535168	596860.992	855699.008	698840
Contas a Receber	964782.976	863148.992	890150.976	873300.992	955827.968	641708.992	759622.016	714601.024	813102.976
...	...	...	...	...	...	...	...	...	...
IR Diferido	-1472	15187	37923	17439	49837	27825	-5763	13165	-7007
Participações/Contribuições Estatutárias	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Reversão dos Juros sobre Capital Próprio	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Part. de Acionistas Não Controladores	0	0	0	0	0	0	0	0	0
Lucro/Prejuízo do Período	116465	43225	-102641	112476	-79542	167888	58059.04	152511.008	194772.992

4.1. Coleta de Dados

	Date	Open	High	Low	Close	Adj Close	Volume
0	2016-01-04	17.730000	17.730000	17.209999	17.209999	14.775684	13206900.0
1	2016-01-05	17.250000	17.520000	17.110001	17.480000	15.007492	10774200.0
2	2016-01-06	17.360001	17.480000	17.200001	17.309999	14.861543	7739100.0
3	2016-01-07	17.170000	17.320000	16.850000	16.850000	14.466606	15316400.0
4	2016-01-08	16.930000	17.200001	16.930000	17.070000	14.655489	10684000.0
...	...	...	...	...	...	...	...

4.2. Pré-processamento dos Dados

- Definição do intervalo de tempo.
  - 2016 a 2021.
  - Restaram 427 ativos, redução de mais de 50%.
- Limpeza dos dados.
  - Empresas com dados que não seguiam os padrões foram removidas.
    - Cerca de 9%.
  - Colunas com muitos dados vazios.
    - De um total de 7411 campos, foram removidas as com 7400 vazios ou mais.

4.2. Pré-processamento dos Dados

- Inserção dos indicadores fundamentalistas.
  - Preço sobre Lucro (P / L);
  - Preço sobre Valor Patrimonial por Ação (P / VPA);
  - *Dividend Yield* (DY);
  - Payout;
  - Lucro por Ação (LPA).

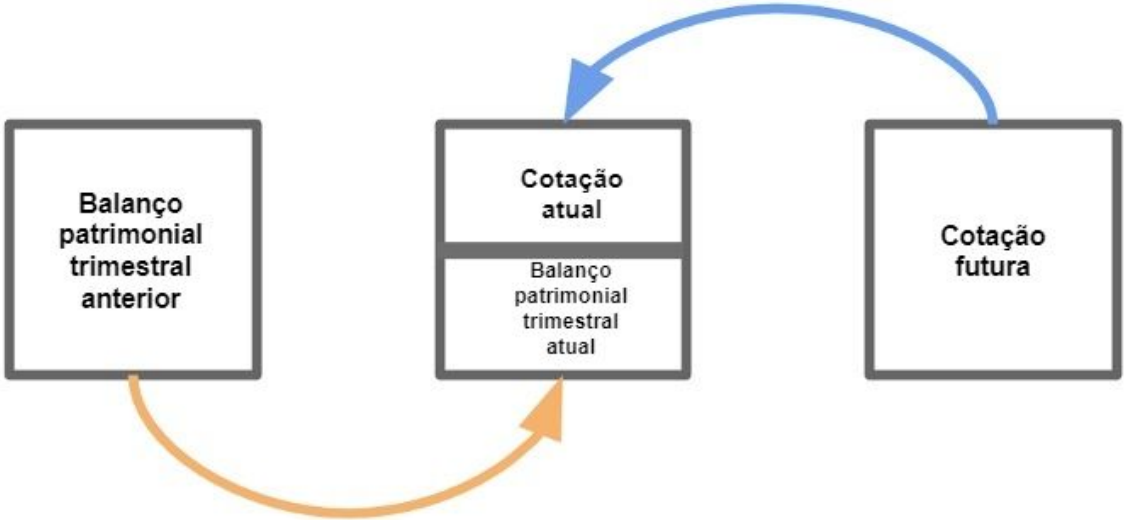
4.2. Pré-processamento dos Dados

- Indicação dos setores de cada empresa.
  - Bens industriais;
  - Comunicações;
  - Consumo cíclico;
  - Consumo não-cíclico;
  - Financeiro;
  - Materiais básicos;
  - Petróleo, gás e biocombustíveis;
  - Saúde;
  - Tecnologia da informação;
  - Utilidade pública;
  - Outros.

4.2. Pré-processamento dos Dados

- Indicação dos setores de cada empresa.
  - Para cada trimestre foi adicionada a média da valorização das empresas do mesmo setor.
- Transformação dos dados.
  - Dados absolutos foram relativizados.
  - Dados fundamentalistas (indicadores, balanço e DRE) relativizados com o passado.
  - Dados de cotações relativizados com o futuro.

4.2. Pré-processamento dos Dados



4.2. Pré-processamento dos Dados

- Criação dos rótulos de classificação.
  - “Decisão”.
  - Comparação entre a variação da cotação da empresa com a variação da média das empresas do mesmo setor.
  - **BUY (Valor 2) ->** cotação da empresa superar em 3% ou mais a média das empresas do setor.
  - **SELL (Valor 0) ->** cotação da empresa for menor em 3% ou menos que a média das empresas do setor.
  - **HOLD (Valor 1) ->** cotação, com relação a média das empresas do mesmo setor, entre 3% e -3%.

4.2. Pré-processamento dos Dados

	Ativo Total	Ativo Circulante	Caixa e Equivalentes de Caixa	Aplicações Financeiras	Decisao
2016-09-30	0.032006	0.077012	0.271493	0.040486	0
2016-12-31	0.059805	0.194882	0.081210	0.029842	2
2017-03-31	-0.037064	-0.108406	-0.082087	-0.964979	1
2017-06-30	0.041991	0.081957	0.211207	-0.114410	2

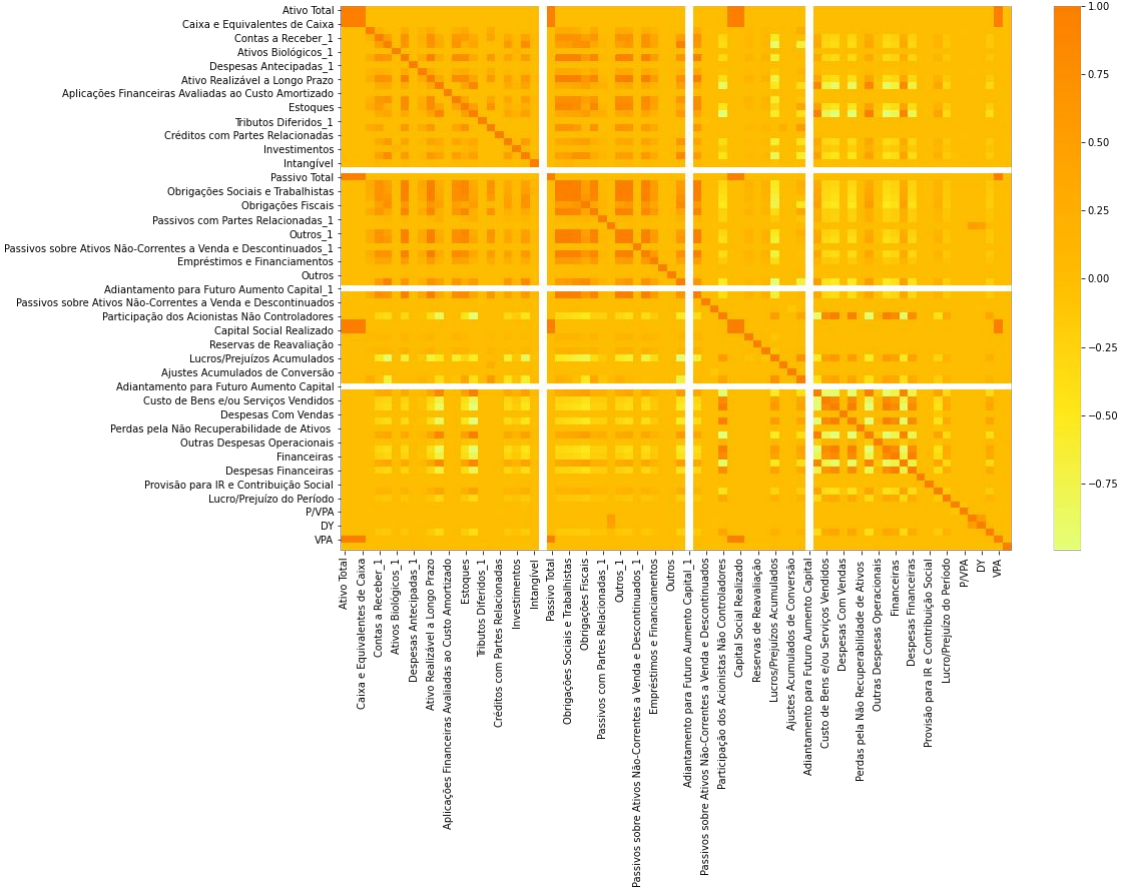
4.2. Pré-processamento dos Dados

- Junção de todas as tabelas das empresas em uma única.
    - 78 colunas.
    - 6143 linhas.

4.3. Análise Exploratória

- Melhor entendimento dos dados tratados.
- Bibliotecas Python como Matplotlib, Seaborn e Plotly.
- Quantidade de cada rótulo atribuída nos dados.
- Correlação de Pearson entre as tabelas.

4.3. Análise Exploratória


Ativo Total
Caixa e Equivalentes de Caixa
Contas a Receber_1
Ativos Biológicos_1
Despesas Antecipadas_1
Ativo Realizável a Longo Prazo
Aplicações Financeiras Avaliadas ao Custo Amortizado
Estoques
Tributos Diferidos_1
Créditos com Partes Relacionadas
Investimentos
Intangível
Passivo Total
Obrigações Sociais e Trabalhistas
Obrigações Fiscais
Passivos com Partes Relacionadas_1
Outros_1
Passivos sobre Ativos Não-Correntes a Venda e Descontinuados_1
Empréstimos e Financiamentos
Outros
Adiantamento para Futuro Aumento Capital_1
Passivos sobre Ativos Não-Correntes a Venda e Descontinuados
Participação dos Acionistas Não Controladores
Capital Social Realizado
Reservas de Reavaliação
Lucros/Prejuízos Acumulados
Ajustes Acumulados de Conversão
Adiantamento para Futuro Aumento Capital
Custo de Bens e/ou Serviços Vendidos
Despesas Com Vendas
Perdas pela Não Recuperabilidade de Ativos
Outras Despesas Operacionais
Financeiras
Despesas Financeiras
Provisão para IR e Contribuição Social
Lucro/Prejuízo do Período
P/VPA
DY
VPA
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

4.3. Análise Exploratória

- Descarte de colunas com correlação de Pearson maior que 0,8 ou menor que -0,8.
- 27 colunas descartadas, restando 51.

4.4. Seleção dos dados

- Redução do número de colunas.
- *Feature selection* utilizando *Extra Trees Classifier*.
    - Parâmetro *feature importances*.
    - Dez colunas mais importantes para a classificação.
    - Descarte das demais colunas.
- Normalização (StandardScaler da biblioteca Sklearn) e separação dos dados de treino e teste

4.4. Seleção dos dados

P/VPA	0.048986
P/L	0.041931
LPA	0.033950
DY	0.029078
Obrigações Fiscais	0.026876
Financeiras	0.026175
Estoques_1	0.025328
Tributos a Recuperar	0.024958
Outros Ativos Circulantes	0.024792
Despesas Com Vendas	0.024453

4.4. Seleção dos dados

	Estoques_1	Tributos a Recuperar	Outros Ativos Circulantes	Obrigações Fiscais	Despesas Com Vendas	Financeiras	P/L	P/VPA	DY	LPA	Decisao
0	-0.011016	-0.033460	-0.037767	-0.038725	0.080915	0.025797	0.015865	-0.021737	-0.039389	-0.000196	0
1	-0.025936	-0.042931	-0.042693	-0.041154	0.080915	0.027443	0.013109	-0.036060	-0.039389	7.527497	2
2	-0.026638	-0.032441	-0.040249	-0.035474	0.080915	0.031635	0.015340	-0.028014	-0.039389	-0.019925	2
3	-0.028158	-0.035007	-0.041271	-0.036348	0.080915	0.009097	0.017182	-0.007593	-0.039389	-0.024033	0
4	-0.024555	-0.039936	-0.042392	-0.044370	0.080915	0.027734	0.013487	-0.026166	-0.039389	0.022231	0
...	...	...	...	...	...	...	...	...	...	...	...

4.5. Aplicação dos Modelos de Inteligência Computacional

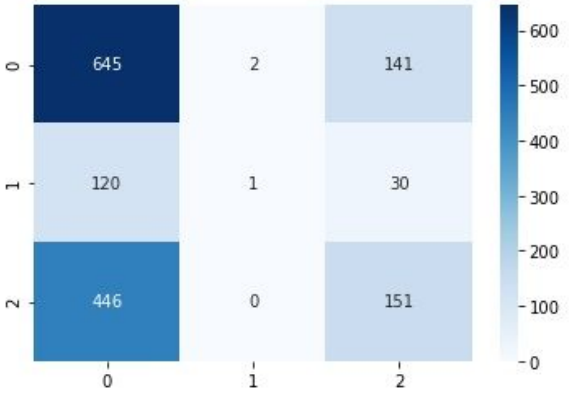
- Desempenho avaliado por métricas.
  - Acurácia (*accuracy*);
  - Precisão (*precision*);
  - Revocação (*recall*);
  - Média F1 (*f1-score*).
- Métricas consideradas sobre a rotulação de compra (*buy*, valor 2).

4.5. Aplicação dos Modelos de Inteligência Computacional

- Testes com modelos genéricos.
  - Cada modelo com seus parâmetros de *accuracy, precision, recall* e *f1-score* e matriz de confusão.
  - Cada modelo foi executado 25 vezes.

AdaBoost

	precision	recall	f1-score	support
0	0.53	0.82	0.65	788
1	0.33	0.01	0.01	151
2	0.47	0.25	0.33	597
accuracy			0.52	1536
macro avg	0.44	0.36	0.33	1536
weighted avg	0.49	0.52	0.46	1536



4.5. Aplicação dos Modelos de Inteligência Computacional

Nome do modelo	Média de *f1-score*	Desvio padrão de *f1-score*
AdaBoost Classifier	0.33	0.0
Decision Tree Classifier	0.4368	0.008818163074019449
Dummy Classifier	0.38119999999999993	0.021414014102918676
Extra Trees Classifier	0.4796000000000001	0.009991996797437435
Gradient Classifier	0.3403999999999999	0.0019595917942265336
KNN	0.40	0.0
Logistic Regression	0.01	0.0
Naive Bayes	0.07	0.0
Random Forest Classifier	0.46280000000000016	0.008255906976220102
Rede neural MLP	0.02280000000000001	0.008726969691708572
SVM	0.02	0.0

4.5. Aplicação dos Modelos de Inteligência Computacional

- Testes com *Machine Learning* Automatizado.
  - Biblioteca PyCaret.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.5679	0.6485	0.4271	0.5472	0.5401	0.1899	0.1973	0.5590
et	Extra Trees Classifier	0.5651	0.6567	0.4286	0.5421	0.5371	0.1861	0.1949	0.5480
catboost	CatBoost Classifier	0.5614	0.6324	0.4150	0.5319	0.5341	0.1820	0.1871	7.9990
lightgbm	Light Gradient Boosting Machine	0.5593	0.6311	0.4182	0.5358	0.5311	0.1737	0.1800	0.6920
xgboost	Extreme Gradient Boosting	0.5579	0.6370	0.4209	0.5368	0.5341	0.1793	0.1837	1.8720
gbc	Gradient Boosting Classifier	0.5407	0.5984	0.3759	0.5075	0.4858	0.1050	0.1200	1.7670
dummy	Dummy Classifier	0.5198	0.5000	0.3333	0.2702	0.3555	0.0000	0.0000	0.0160
ridge	Ridge Classifier	0.5195	0.0000	0.3339	0.4344	0.3604	0.0017	0.0054	0.0100
lda	Linear Discriminant Analysis	0.5195	0.5072	0.3339	0.4344	0.3604	0.0017	0.0054	0.0180
lr	Logistic Regression	0.5193	0.5102	0.3338	0.4368	0.3611	0.0017	0.0048	0.0630
ada	Ada Boost Classifier	0.5174	0.5494	0.3558	0.4544	0.4597	0.0578	0.0670	0.1820
svm	SVM - Linear Kernel	0.5128	0.0000	0.3337	0.3961	0.3693	-0.0015	-0.0028	0.0290
dt	Decision Tree Classifier	0.5053	0.5768	0.4246	0.5104	0.5074	0.1494	0.1496	0.0300
knn	K Neighbors Classifier	0.5000	0.5596	0.3732	0.4686	0.4752	0.0769	0.0793	0.0680
qda	Quadratic Discriminant Analysis	0.1272	0.5129	0.3348	0.4329	0.0781	0.0004	0.0005	0.0120
nb	Naive Bayes	0.1253	0.5130	0.3351	0.4329	0.0735	-0.0002	-0.0014	0.0190

Nome do modelo	Média de *f1-score*	Desvio padrão de *f1-score*
AdaBoost Classifier	0.459428	0.00591550640266748
CatBoost Classifier	0.5225279999999999	0.008231428551594193
Decision Tree Classifier	0.48999200000000004	0.00824393498276157
Dummy Classifier	0.35501599999999994	0.005581016394887223
Extra Trees Classifier	0.535728	0.006880437195411348
Extreme Gradient Boosting Machine	0.5273199999999999	0.007649365986799164
Light Gradient Boosting Machine	0.5202319999999999	0.009013532936645873
Linear Discriminant Analysis	0.35894	0.005943803496078922
Logistic Regression	0.3596640000000001	0.006215585571770371
Gradient Boosting Classifier	0.47642799999999996	0.006288689529623804
KNN Classifier	0.46768800000000005	0.024397414945030556
Naive Bayes	0.07069199999999999	0.004755369175994645
Quadratic Discriminant Analysis	0.07512	0.0042727040618325066
Random Forest Classifier	0.5323280000000001	0.00648114310905105
Ridge Classifier	0.35897600000000013	0.0059094690116794765
SVM - Linear Kernel	0.369392	0.008572183852438074

4.5. Aplicação dos Modelos de Inteligência Computacional

- Seleção dos melhores modelos.
  - CatBoost Classifier;
  - Extra Trees Classifier;
  - XGBoost;
  - Random Forest Classifier;
  - Light GBM.

4.5. Aplicação dos Modelos de Inteligência Computacional

- Otimização dos hiperparâmetros dos modelos selecionados.
  - Melhorar o desempenho dos modelos.
  - Busca profunda (*Grid Search*).
  - Busca randômica (*Random Search*).
- Biblioteca Sklearn
  - *RandomizedSearchCV* e pontuação *f1-weighted*.
  - Escolha de valores conforme documentação.

4.5. Aplicação dos Modelos de Inteligência Computacional

- CatBoost Classifier

Parâmetro	Resultado encontrado pela busca randômica
depth	10
iterations	500
learning_rate	0.3
l2_leaf_reg	1
border_count	32
thread_count	1

4.5. Aplicação dos Modelos de Inteligência Computacional

- Extra Trees Classifier

Parâmetro	Resultado encontrado pela busca randômica
criterion	*gini*
max_depth	48
min_samples_split	14
min_samples_leaf	1
max_features	*auto*
n_estimators	4

4.5. Aplicação dos Modelos de Inteligência Computacional

- XGBoost

Parâmetro	Resultado encontrado pela busca randômica
max_depth	20
min_child_weight	4
subsample	0.75
learning_rate	0.2
n_estimators	128
colsample_bytree	0.7

4.5. Aplicação dos Modelos de Inteligência Computacional

- Random Forest Classifier

Parâmetro	Resultado encontrado pela busca randômica
bootstrap	*True*
criterion	*gini*
max_depth	70
max_features	*auto*
min_samples_leaf	1
min_samples_split	2

4.5. Aplicação dos Modelos de Inteligência Computacional

- Light GBM

Parâmetro	Resultado encontrado pela busca randômica
learning_rate	0.1
num_leaves	222
subsample	1
colsample_bytree	0.3
max_depth	90
min_child_samples	1
bagging_fraction	0.3

4.5. Simulações com Dados Reais

- Entrada: dados do primeiro trimestre de 2021 (março).
- Saída: compras, vendas e *holds*.
- Carteira fictícia montada com as compras.
- Análise do desempenho em 1 ano.

5. Resultados e Discussões

- Resultados das simulações.
- Análise do tamanho e do desempenho das carteiras fictícias geradas
    - Intervalo de 1 ano.
- Total de empresas pós pré-processamento: 370
- Rendimento médio destas 370 empresas no mesmo período:

  0,8140677563218336.

5. Resultados e Discussões

- CatBoost Classifier.

Variável	Média	Desvio padrão
Quantidade de ações na carteira	129,28	23,95916526091842
Variação no tamanho da carteira	-0,65059459412	0,0647545007039251
Rendimento da carteira	0,8335264608323817	0,008929500020533337

5. Resultados e Discussões

- Extra Trees Classifier.

Variável	Média	Desvio padrão
Quantidade de ações na carteira	152,32	54,39317604258829
Variação no tamanho da carteira	-0,5883243238399999	0,14700858376219764
Rendimento da carteira	0,823604278208304	0,02484047594974296

5. Resultados e Discussões

- XGBoost.

Variável	Média	Desvio padrão
Quantidade de ações na carteira	92,64	30,6723067277308
Variação no tamanho da carteira	-0,7496216211999998	0,08289812627196926
Rendimento da carteira	0,842463645688401	0,015400618707479096

5. Resultados e Discussões

- Random Forest Classifier.

Variável	Média	Desvio padrão
Quantidade de ações na carteira	181,32	30,632949580476247
Variação no tamanho da carteira	-0,5099459453999999	0,08279175563506316
Rendimento da carteira	0,8176090798871962	0,013965998337635089

5. Resultados e Discussões

- Light GBM.

Variável	Média	Desvio padrão
Quantidade de ações na carteira	129,44	59,99071928223565
Variação no tamanho da carteira	-0,6501621616399998	0,1621370791632075
Rendimento da carteira	0,8049425733600178	0,026937296701084346

5. Resultados e Discussões

- Nenhum modelo apresentou um rendimento muito superior à média do mercado.
- Redução do tamanho sempre acima de 50%.
- XGBoost teve o melhor desempenho.
- CatBoost Classifier teve o segundo melhor desempenho.
- Light GBM com rendimento médio inferior à média do mercado.

6. Conclusão

- Objetivo de pré-seleção de empresas possivelmente lucrativas utilizando modelos de *machine learning* atingido.
- Avaliação dos modelos realizada.
    - Carteiras com rendimentos semelhantes à média do mercado nacional
    - Tamanho reduzido em, no mínimo, 50%
    - Esforço e tempo reduzido para futuras análises por parte do investidor.

6.1. Trabalhos Futuros

- Estratégia de remoção de empresas com lucro negativo no pré-processamento.
- Simulações para encontrar o valor ideal do percentual utilizado para rotulação dos trimestres em *buy, hold* e *sell*.
- Adição de mais dados e indicadores, como o índice de Graham.
- Maior otimização dos hiperparâmetros dos modelos de *machine learning*.
- Integração do projeto a uma aplicação real, automatizando a coleta de dados e etapas como pré-processamento.
- Aplicar junto à metodologia proposta métodos de auxílio multicritério à decisão, como o *Anarchy Hierarchy Process* (AHP).