Predicting Startup Success Using Data Science and Machine Learning Techniques

Equipe 4
Guilherme Schwarz, Júlia Cristina Moreira da Silva, Matheus Francisco Trevisan Del Zotto, Renan Belem Biavati
{guilherme.schwarz,s.moreira4,Matheus.del,renan.belem}@pucpr.edu.br

I. Introdução

Startups desempenham um papel central na economia moderna por sua capacidade de promover inovação, gerar empregos e transformar setores tradicionais [1]. No entanto, essas organizações também estão sujeitas a altos níveis de incerteza e risco, especialmente em seus estágios iniciais, nos quais a ausência de um histórico consolidado dificulta a avaliação de seu potencial de sucesso [2].

Diante desse cenário, cresce o interesse em utilizar técnicas de ciência de dados e aprendizado de máquina como suporte à tomada de decisão em contextos de investimento [3]. A análise de dados estruturados — como perfil da equipe fundadora, setor de atuação, investimentos captados e tempo de operação — tem se mostrado promissora na identificação de padrões associados ao desempenho futuro dessas empresas [4].

Este artigo propõe uma abordagem baseada em algoritmos de aprendizado supervisionado para modelar e prever o sucesso de startups. A partir de um conjunto de dados reais, são exploradas as variáveis mais relevantes, definidos protocolos de validação robustos e aplicadas métricas de avaliação que garantam a confiabilidade dos resultados. O objetivo é contribuir com ferramentas analíticas que apoiem investidores, aceleradoras e formuladores de políticas públicas na identificação de startups com maior potencial de crescimento.

II. Metodologia

A. Definição do Conjunto de Dados e da Tarefa

O conjunto de dados utilizado neste trabalho foi obtido a partir de um repositório público e contém informações históricas sobre startups, incluindo datas de fundação e encerramento, rodadas de investimento, localização, setor de atuação e marcos atingidos, entre outros. A tarefa definida para o projeto consiste na predição do sucesso de startups — representado pela variável *status*, que assume os valores *acquired* (adquirida) e *closed* (encerrada). O objetivo é prever esse desfecho com base nas características observadas ao longo do ciclo de vida das empresas.

B. Análise Exploratória e Pré-processamento dos Dados

Inicialmente, os dados foram carregados diretamente de uma fonte online utilizando a biblioteca `pandas`. Em seguida, realizou-se a inspeção geral das variáveis e a identificação de colunas irrelevantes, como `Unnamed: 0` e `Unnamed: 6`, que foram removidas. Também foi verificada a presença de valores ausentes, que foram tratados conforme o tipo e a importância da variável.

Em seguida, foi realizada engenharia de atributos para criar variáveis adicionais, a fim de enriquecer a base de dados e refletir aspectos relevantes para a tarefa preditiva:

- **`is_technology`**: variável binária que identifica startups pertencentes ao setor tecnológico, com base na categorização das áreas de atuação;
- **`age_at_closing`**: idade da startup no momento do encerramento, calculada em anos a partir da diferença entre as datas de fundação e encerramento.

C. Descrição Estatística e Visualização Inicial

Foi realizada uma análise descritiva das variáveis, incluindo o número total de instâncias, a quantidade de atributos, o número de classes e a distribuição destas (*Number of instances: 923; Number of features: 49; Number of classes: 2*). A variável `status` (sucesso ou fracasso) foi analisada por meio de gráficos de barras, evidenciando o desbalanceamento da base, com 64,7% de *Acquired* e 35,3% de *Closed*.

D. Análise Univariada

Foram analisadas, individualmente, ao menos 20 variáveis, entre categóricas, numéricas e temporais. Para as variáveis numéricas, foram investigadas medidas estatísticas como média, desvio padrão, assimetria (skewness) e curtose. Já para as variáveis categóricas, foram exploradas as frequências relativas, enquanto as variáveis temporais foram agregadas por ano para observar tendências sazonais.

E. Análise Multivariada

A análise multivariada teve como objetivo explorar possíveis correlações e padrões entre variáveis, a fim de investigar comportamentos relevantes no conjunto de dados. Foram testadas as seguintes hipóteses:

- **Hipótese 1**: Startups tecnológicas recebem mais investimento do que startups de outros setores.
- **Hipótese 2**: Startups que atingem o primeiro marco mais rapidamente tendem a ter maior sucesso.
- **Hipótese 3**: Crises econômicas impactam o número de fundações e encerramentos de startups, especialmente no setor de tecnologia.
- **Hipótese 4**: Startups deixam de operar pouco tempo após pararem de receber investimentos.

- **Hipótese 5**: Receber o primeiro investimento em uma fase inicial da vida da startup está relacionado a uma maior longevidade.

Effective Data Visualization

- **Visualização 1 – Impacto de crises econômicas nas startups:** A visualização aprimorada com histogramas sobrepostos ilustra a distribuição temporal de fundações e encerramentos de startups, considerando também os períodos de crise econômica. Observou-se um aumento significativo no número de encerramentos nos anos seguintes à crise de 2008, especialmente entre 2011 e 2013. Em contraste, a bolha da internet (2000) e a pandemia da COVID-19 (2020) não apresentaram efeitos evidentes no conjunto de dados disponível (limitado até 2013). Os padrões também foram confirmados para startups tecnológicas, sugerindo que nem mesmo esse setor, frequentemente considerado mais resiliente, ficou imune aos efeitos da recessão de 2008.
- **Visualização 2 – Tempo até o primeiro marco e sucesso:** A comparação da variável "idade ao atingir o primeiro marco"entre empresas adquiridas e encerradas foi realizada por meio de um boxplot. A distribuição mostrou que não há diferença significativa entre os grupos, com médias semelhantes e presença de outliers em ambos os casos. Notavelmente, algumas startups adquiridas atingiram seu primeiro marco apenas após 15 anos de operação. Os resultados não sustentam a hipótese de que alcançar marcos iniciais em menos tempo esteja associado ao sucesso.
- **Visualização 3 – Investimento em empresas tecnológicas vs. não tecnológicas:** A análise comparativa entre startups tecnológicas e não tecnológicas indicou que as primeiras tendem a receber investimentos médios mais elevados. Embora a distribuição das faixas de financiamento não tenha apresentado diferenças tão expressivas, a média de investimento foi significativamente maior nas empresas de base tecnológica. Esses resultados reforçam a ideia de que startups do setor de tecnologia atraem maior volume de capital, possivelmente devido ao seu potencial de escalabilidade e inovação.

As visualizações foram construídas com o auxílio das bibliotecas `matplotlib`, `seaborn`, `scipy` e `pandas`. As evidências extraídas dessas análises foram fundamentais para orientar as próximas etapas do projeto, especialmente a seleção de atributos relevantes para a modelagem preditiva.

III. Protocolo Experimental

A. Pré-processamento e Engenharia de Atributos

O conjunto de dados foi carregado a partir de uma URL do GitHub e passou por um processo de limpeza inicial. Além da remoção de colunas com excesso de valores nulos ou informações redundantes, foi realizada uma etapa adicional de engenharia de atributos. Para otimizar o custo computacional, especialmente do seletor RFE (Recursive Feature Elimination), que se mostrou excessivamente lento com o conjunto original de quase 50 atributos, foi feita uma seleção manual para remover variáveis consideradas menos significativas com base em análise preliminar. Este passo foi crucial para viabilizar a execução dos experimentos em tempo hábil.

As colunas de data foram convertidas para o formato `datetime`. Uma nova coluna binária, `is_technology`, foi criada para identificar startups do setor de tecnologia, e a coluna `age_at_closing` foi calculada como a diferença entre as datas de fundação e de fechamento da empresa.

B. Preparação para Treinamento

Para o treinamento dos modelos, a coluna `status` foi definida como a variável-alvo (y). As demais colunas foram designadas como variáveis preditoras (X), após a aplicação da codificação de variáveis categóricas usando a função `pd.get_dummies`. A variável-alvo `status` foi, por sua vez, codificada numericamente com o `LabelEncoder`.

C. Validação Cruzada

Para lidar com o desbalanceamento de classes e avaliar a performance dos modelos de forma robusta, foi empregada a estratégia de validação cruzada `Repeated Stratified K-Fold`. Essa abordagem foi configurada com 5 splits e 3 repetições, visando obter uma estimativa mais estável do desempenho do modelo e reduzir a variância dos resultados. Para garantir a reprodutibilidade dos experimentos, um `random_state` foi fixado em todas as etapas.

D. Métricas de Avaliação

As métricas utilizadas para comparar o desempenho dos modelos foram:

- **F1-Score (weighted):** Métrica principal, que calcula a média ponderada do F1-score para ambas as classes, sendo crucial em problemas com dados desbalanceados.
- **Accuracy:** Proporção de previsões corretas em relação ao total.
- **Precision (weighted):** Capacidade do modelo de identificar corretamente as instâncias positivas, ponderada pela frequência das classes.
- **Recall (weighted):** Capacidade do modelo de encontrar todas as instâncias positivas, também ponderada pela frequência das classes.
- **Training Time:** Tempo, em segundos, gasto para treinar e avaliar o modelo em cada configuração.

E. Modelos Preditivos e Hiperparâmetros Avaliados

Foram avaliados cinco modelos de aprendizado de máquina supervisionado, representando diferentes abordagens para problemas de classificação:

- **Logistic Regression:** Um modelo linear simples, utilizado como baseline.
- **Random Forest:** Um modelo de *ensemble* baseado em árvores de decisão, robusto a *overfitting* e capaz de capturar interações não lineares.

- **Gradient Boosting:** Outro modelo de *ensemble* que constrói preditores sequencialmente, focando nos erros dos modelos anteriores.
- **Support Vector Machine (SVM):** Um modelo poderoso para classificação, especialmente em espaços de alta dimensão.
- **K-Nearest Neighbors (KNN):** Um classificador baseado em distância, simples de implementar e eficaz para conjuntos de dados menores.

Para otimizar o desempenho dos modelos e lidar com as características do conjunto de dados, foram exploradas combinações de técnicas de balanceamento de classes e seleção de atributos.

Técnicas de Balanceamento e Seleção de Atributos

- **Balanceamento:**
  - **SMOTE (Synthetic Minority Oversampling Technique):** Técnica de *oversampling* para aumentar o número de instâncias da classe minoritária.
  - **RandomUnderSampler:** Técnica de *undersampling* para reduzir o número de instâncias da classe majoritária.
- **Seleção de Atributos:**
  - **SelectKBest:** Seleciona os 10 melhores atributos com base no teste estatístico ANOVA F-score.
  - **RFE (Recursive Feature Elimination):** Elimina recursivamente os atributos menos importantes com base nos pesos do modelo, selecionando os 10 mais relevantes.

Cada modelo foi testado em combinação com cada técnica de balanceamento e cada método de seleção de atributos. Essa abordagem resultou em um total de 5 modelos $\times 2$ samplers $\times$ 2 seletores $= 20$ configurações distintas, todas avaliadas sob o mesmo protocolo de validação cruzada. Os hiperparâmetros dos modelos foram mantidos como os padrões da biblioteca *scikit-learn* para esta etapa inicial de avaliação.

IV. Resultados nos Dados de Teste

Os resultados obtidos na avaliação dos diferentes modelos de Machine Learning no conjunto de dados de startups são apresentados a seguir. A Tabela I resume o desempenho de uma seleção das 20 combinações testadas, ordenadas pelo F1-score ponderado (F1), e ilustra os principais trade-offs encontrados.

Os resultados mostram que diversas combinações de modelos alcançaram F1-scores ponderados de 0.99, indicando um desempenho excelente na classificação. As métricas de Acurácia, Precisão e Recall também foram consistentemente altas para essas combinações de melhor desempenho.

V. Discussão e Análise dos Resultados

A análise dos resultados revela insights importantes sobre a predição de sucesso de startups com base neste conjunto de dados:

- **Desempenho Elevado:** A maioria das combinações testadas apresentou um desempenho de classificação muito alto, com F1-scores ponderados próximos a 0.99. Isso sugere que os atributos selecionados, combinados com as técnicas de balanceamento e os modelos utilizados (exceto nos casos com PCA), são eficazes na distinção entre startups adquiridas e fechadas.
- **Impacto da Seleção de Atributos:** A seleção de atributos teve um impacto notável no desempenho. O método PCA (Principal Component Analysis) resultou consistentemente nos piores F1-scores para todos os modelos com os quais foi combinado, conforme destacado na Tabela I. Isso indica que a projeção linear realizada pelo PCA pode ter removido informações cruciais para a classificação. Em contraste, SelectKBest e RFE mantiveram o alto desempenho em métricas de classificação, embora com custos computacionais distintos.
- **Eficiência Computacional:** Conforme antecipado na metodologia e evidenciado na Tabela I, o custo computacional foi um fator distintivo. Combinações que utilizaram RFE com modelos como RandomForest e GradientBoost foram significativamente mais lentas, com tempos de execução superiores a 600 segundos, tornando-as inviáveis para cenários que exigem agilidade. Por outro lado, combinações como 'Logistic Regression + RUS + SKB' e 'KNN + RUS + SKB' atingiram F1-scores excelentes com tempos de execução inferiores a 1 segundo, demonstrando um ótimo trade-off.
- **Balanceamento de Classes:** As técnicas de balanceamento (SMOTE e RandomUnderSampler) parecem ter sido bem-sucedidas em mitigar o impacto do desbalanceamento, permitindo que os modelos alcançassem altas pontuações em métricas que consideram ambas as classes (como o F1-score ponderado).
- **Modelos:** Modelos como Logistic Regression, SVM, RandomForest e GradientBoost apresentaram alto desempenho quando combinados com SelectKBest ou RFE. O KNN, embora ligeiramente inferior em alguns casos, também obteve ótimos resultados com o SelectKBest, apresentando-se como uma alternativa rápida e eficaz.

VI. Conclusão

Em suma, os resultados demonstram a viabilidade de construir modelos preditivos de alta performance para o sucesso de startups com base neste conjunto de dados. A escolha da combinação ideal dependerá do equilíbrio desejado entre desempenho preditivo e eficiência computacional. Para este projeto, pipelines mais simples e rápidos, como aqueles que utilizam Logistic Regression ou KNN com RandomUnderSampler e SelectKBest, apresentaram o melhor trade-off. As principais limitações encontradas foram a ineficiência computacional do RFE com modelos complexos e a perda de desempenho associada ao uso do PCA. As implicações práticas desses achados residem na possibilidade de utilizar modelos preditivos eficazes para auxiliar investidores e empreendedores na avaliação do potencial de novas startups.

Tabela I
DESEMPENHO DE COMBINAÇÕES SELECIONADAS, ORDENADAS PELO F1-SCORE PONDERADO.

Grupo	Modelo	Balanceamento	Seleção	F1-Score	Acurácia	Precisão	Recall	Tempo (s)
Melhor Custo-Benefício	**Log. Regression**	RUS	SKB	**0.99**	0.99	0.99	0.99	**0.63**
	KNN	RUS	SKB	0.98	0.98	0.98	0.98	0.72
	SVM	SMOTE	SKB	0.99	0.99	0.99	0.99	0.95
	SVM	RUS	SKB	0.99	0.99	0.99	0.99	1.12
Ineficientes (Alto Custo)	RandomForest	RUS	RFE	0.99	0.99	0.99	0.99	632.37
	GradientBoost	RUS	RFE	0.99	0.99	0.99	0.99	802.90
	RandomForest	SMOTE	RFE	0.99	0.99	0.99	0.99	805.28
	GradientBoost	SMOTE	RFE	0.99	0.99	0.99	0.99	**1534.82**
Pior Desempenho	**Log. Regression**	RUS	PCA	0.82	0.81	0.82	0.81	0.76
	SVM	RUS	PCA	0.81	0.81	0.82	0.81	1.16
	KNN	SMOTE	PCA	0.78	0.79	0.78	0.79	1.06
	KNN	RUS	PCA	**0.75**	0.75	0.76	0.75	0.75

Abreviações: RUS (RandomUnderSampler); SMOTE (Synthetic Minority Oversampling Technique); SKB (SelectKBest); RFE (Recursive Feature Elimination); PCA (Principal Component Analysis).

REFERÊNCIAS

[1] Blank, S., and Dorf, B. (2012). *The Startup Owner's Manual*. K&S Ranch.

[2] Ries, E. (2011). *The Lean Startup*. Crown Business.

[3] Guzman, J., and Stern, S. (2015). Where is Silicon Valley? *Science*, 347(6222), 606–609.

[4] Younis, M., Zhao, Y., and Younis, S. (2020). Predicting Startup Success Using Machine Learning Techniques. *Journal of Entrepreneurship Education*.