Projeto Aplicado - Aprendizagem de Máquina

Pedro Augusto A. F. de Melo Lucas Rabay Butcher

Abril de 2025

Introdução

Todos os arquivos e códigos necessários para subir o app (que ainda está em desenvolvimento) estão no GitHub do projeto (clicar aqui).

A tecnologia tem revolucionado a forma como funcionam os sistemas financeiros mundo afora, tornando o processo da transferência de renda cada vez mais prático e democrático, o que permite que pessoas pouco favorecidas e que residem em ambientes pouco acessíveis consigam ser alcançar serviços de suma importância para a sua vida. Nesse contexto, sabe-se da importância da disseminação da educação financeira para uma sociedade, que deve obter estes conhecimentos desde o ensino básico, para que problemas relacionados a esta transferência de renda sejam causados.

Entretanto, apesar de a relevância desse cenário ser de conhecimento geral, nota-se uma dificuldade de manter um controle e uma vida financeira saudável entre as pessoas, o que leva ao foco deste estudo. No intuito de identificar e entender a fundo os principais fatores que afetam o psicológico das pessoas com relação a dinheiro e às suas finanças em geral, desenvolve-se este trabalho, que utiliza de uma base de dados proveniente de uma pesquisa feita pelo Grupo do Banco Mundial para atingir ao objetivo proposto, por meio de uma exploração e modelagem deste conjunto de dados.

Ainda, dentre as potenciais contribuições que os desenvolvedores do projeto anseiam oferecer, pode-se citar o impacto que a identificação de variáveis que influenciam na preocupação com as finanças das pessoas pode trazer para órgãos públicos, que podem atuar com foco nestes fatores para garantir que estes problemas sejam sanados de forma muito mais eficiente, tratando o empecilho por um fator que realmente o impacta e impedindo que futuros problemas que são acarretados pela fragilidade econômica da população, como a falta de empreendedorismo e investimentos em geral. Além disso, ressalta-se o legado

para instituições que focam na educação financeira da população, que podem trazer um viés estatístico para atuar com foco em determinadas vertentes.

Revisão Literária

Neste tópico, alguns pontos relevantes para o entendimento do contexto e dos procedimentos utilizados para a modelagem dos dados são expostos, e é fruto de uma revisão da literatura realizada com o objetivo de obter uma melhor preparação para o desenvolvimento do projeto.

Preocupações financeiras e sofrimento

De acordo com resultados do estudo desenvolvido por Ryu e Lu Fan (2022), o sofrimento emocional de adultos dos estados unidos sofre uma relevante influência de preocupações financeiras, sendo ainda mais intensificada quando se tem a interação com outras variáveis socioeconômicas, e impactando de forma significativa na qualidade de vida dos cidadãos. Além disso, o trabalho cita a importância de variáveis que estão fora dos indicadores objetivos de estresse financeiro, que geralmente se apresentam como a incapacidade de honrar com obrigações ou de atingir objetivos financeiros. Ou seja, este e outros estudos recentes mostram que a percepção subjetiva do indivíduo sobre o seu dinheiro também impacta de forma significativa.

Portanto, tanto o entendimento das variáveis quanto a previsão de um possível nível de estresse para o futuro ganha importância, ainda mais em países emergentes ou subdesenvolvidos onde as preocupações financeiras atingem parte alarmante da população. No caso do Brasil, uma pesquisa realizada pela Associação Brasileira das Entidades Mercado Financeiro e de Capitais (AMBIMA) de 2024 mostra que mais da metade da população possuem algum tipo de preocupação ou estresse financeiro. Sob esse panorama, é reforçada a necessidade de estudos cada vez mais assertivos quanto à temática.

O uso da aprendizagem supervisionada no estudo do comportamento financeiro

Como apontado por Feng Zhu (2024), muitas técnicas de aprendizagem de máquina e conjuntos de dados com informações técnicas tem sido utilizadas por instituições financeiras e grandes casas de *research* para a gestão de risco ou para a precificação de ativos financeiros, mas pouco tem sido utilizado para o estudo de finanças pessoais. Assim, isso também pode ser aplicado para o tema do projeto, que utiliza de conjuntos de dados com informações sociodemográficas no escopo das finanças pessoais.

Introdução aos modelos utilizados

Os modelos de classificação, abordagem utilizada no projeto, é um tipo de modelagem do vasto campo da aprendizagem supervisionada, que utiliza conjuntos de dados com uma variável alvo rotulada e com natureza discreta. Tratando dos conhecimentos necessários

para a construção dos modelos, estes foram obtidos em sala e, como complemento para aplicações mais avançadas, foram utilizadas três principais obras, sendo elas:

- Estatística Básica (Bussab e Morettin);
- An Introduction to Statistical Learning (James at al);
- Fundamentals of Deep Learning: Designing Next Generation Machine Intelligence Algorithms (Boduma et al).

Assim, tomando como base tais referências, os códigos, análises e interpretações puderam ser feitas de forma muito mais robusta, podendo ser observado nos códigos anexados.

Fonte dos Dados

Nesta seção, é apresentada uma breve descrição dos dados que serão utilizados no projeto, seguindo as orientações solicitadas:

Os dados utilizados são provenientes do *Global Findex 2021*, disponibilizados pelo Banco Mundial por meio da World Bank Microdata Library. Esse conjunto de microdados reúne informações individuais acerca de inclusão financeira em diversos países, permitindo uma análise aprofundada sobre o acesso e uso de serviços financeiros em nível global.

Descrição das Variáveis

Variável	**Descrição**
`economy`	Nome da economia ou país no qual o respondente está inserido.
`female`	Indicador de gênero, onde valores específicos identificam se o respondente é do sexo feminino.
`inc_q`	Quintil de renda domiciliar dentro da economia, que classifica os respondentes em cinco grupos, do menor ao maior nível de renda.
`emp_in`	Indicador de inserção no mercado de trabalho, identificando se o indivíduo faz parte da força de trabalho.
`account`	Indica se o respondente possui alguma conta financeira, seja em instituição bancária ou através de serviços de dinheiro móvel.
`borrowed`	Informação sobre se o respondente realizou empréstimos no último ano, independentemente da fonte.
`saved`	Indica se o respondente realizou ações de poupança ou reservou parte de sua renda no período avaliado.
`receive_wages`	Registra se o respondente recebeu pagamento salarial e, se sim, de que forma (por conta ou em espécie).
`receive_transfers`	Identifica o recebimento de transferências governamentais.
`receive_pensions`	Informações sobre o recebimento de pensões governamentais.
`pay_utilities`	Indica se o respondente realizou pagamentos de contas de serviços (como luz, água ou outros) através de contas financeiras.
`anydigpayment`	Indicador de se o respondente fez ou recebeu pagamentos digitais no período.

`mobileowner`	Informa se o respondente é proprietário de um telefone celular.
`internetaccess`	Registra o acesso à Internet por parte do respondente.
`fin2`	Relaciona-se ao uso ou posse de cartões de débito (indicando, por exemplo, a existência de um cartão ATM).
`fin7`	Indica se o respondente possui um cartão de crédito.
`fin8b`	Refere-se à utilização do cartão de crédito, por exemplo, se o respondente quitou os saldos de cartão integralmente.
`fin44a`, `fin44b`, `fin44c`, `fin44d`	Capturam o nível de preocupação financeira com temas como: segurança na velhice, custos médicos, contas e educação. Cada uma dessas variáveis foi utilizada para construir a objetivo `financial_distress_level`, que possui as classes ***low worry, medium worry*** e ***great worry***.

Tabela 1 – Variáveis utilizadas no projeto e suas descrições.

Critérios de Inclusão e Exclusão Amostral

Inclusão

- Serão considerados apenas os respondentes para os quais estão disponíveis respostas válidas para todas as variáveis listadas.
- O conjunto de dados inclui indivíduos a partir dos 15 anos, conforme definido pelo Global Findex, o que garante a representatividade da população adulta.

Exclusão

- Respostas categorizadas como "não sei" ou "recusou responder" podem ser tratadas de forma diferenciada (por exemplo, consideradas como 0 na construção do score de `financial_distress_level`) ou excluídas, conforme os critérios pré-estabelecidos de análise.
- Respostas com valores faltantes (*missing values*) nas variáveis-chave serão avaliadas e, se necessário, excluídas ou imputadas por meio de técnicas apropriadas, a fim de assegurar a qualidade dos dados para o treinamento do modelo.

Essa abordagem garante que os dados utilizados permitam uma análise robusta dos fatores associados à preocupação financeira, bem como o desenvolvimento de modelos preditivos confiáveis.

Preparação dos Dados

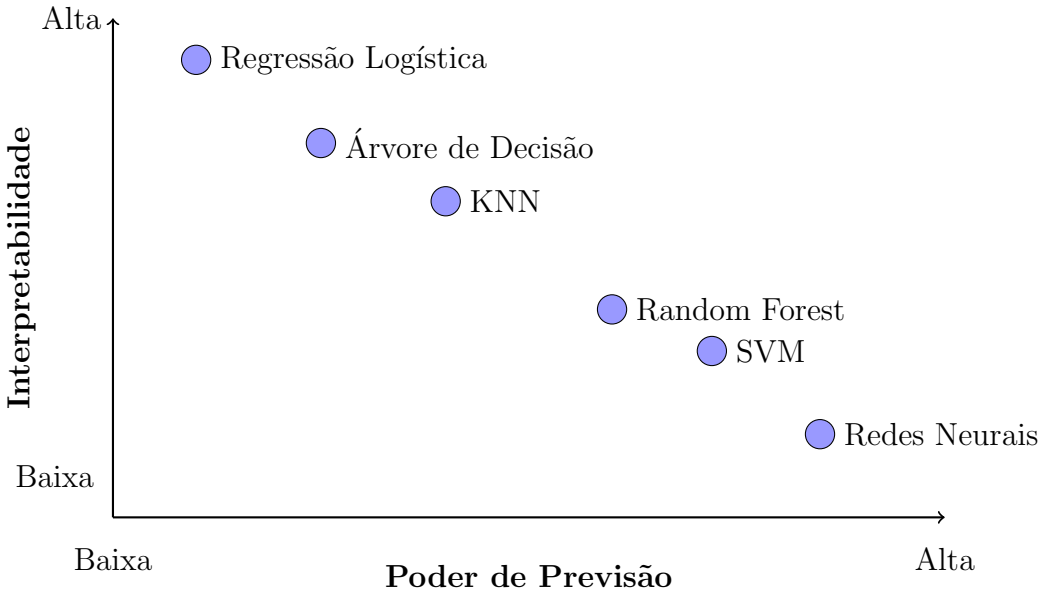
A etapa de preparação dos dados, que envolve desde a limpeza até as transformações e balanceamentos de classes são desenvolvidas no *notebook python* no GitHub do projeto. Tais etapas foram realizadas seguindo sempre um mesmo algoritmo:

- Primeiro se analisa o problema no conjunto de dados de forma resumida, por meio de métodos que permitem esta compactação e a identificação dos padrões nos dados;
- Posteriormente, se comprova uma ideia obtida no primeiro passo por meio da análise gráfica, onde são utilizadas bibliotecas que permitem a construção destes gráficos, sempre escolhendo os mais adequados para observar cada problemática desejada.
- Por último, as transformações necessárias são aplicadas, com mais testes visuais sendo realizados posteriormente para confirmar a melhoria em cada problema identificado.

Estratégia Empírica

A definição dos métodos de aprendizagem supervisionada escolhidos, em um primeiro momento, deu-se pelo critério do desejo de ampliar o conhecimento dos desenvolvedores do projeto em modelos que fossem menos interpretáveis, mas que tendem a ter um melhor ajuste aos conjuntos de dados e uma maior assertividade também, tendo em vista que alguns deles ainda não haviam sido estudados a fundo em sala de aula. Nesse contexto, a ideia foi utilizar alguns dos modelos menos interpretáveis que possuem uma "Acurácia"mais atrativa, tendo em mente este *trade-off*, que é apresentado no gráfico abaixo:

Trade-off entre Interpretabilidade e Poder de Previsão



Assim, optou-se por utilizar os seguintes modelos: Regressão Logística, *Decision Tree Classifier*, *Random Forrets Classifier* e uma Rede Neural. Ainda, é válido ressaltar a importância da definição de um *benchmark* para que se tenha uma *baseline* para a assertividade, complexidade e interpretabilidade dos modelos utilizados no projeto. Nesse sentido, a Regressão Logística, modelo mais básico e que permite a melhor análise e compreensão de como cada variável influencia na previsão, foi escolhida como modelo de referência.

Comparação, avaliação e seleção dos modelos

Primeiramente, é importante citar que existem dois processos que exigem a avaliação e escolha de um dentre alguns modelos:

- O primeiro é o processo da escolha do melhor modelo entre os do mesmo tipo (*e.g* a escolha do modelo com maior poder de previsão dentre os de *Random Forest Classifier*. Como a modelagem utiliza uma abordagem de classificação, para a escolha entre os modelos de um mesmo tipo, optamos por utilizar a matriz de confusão, acurácia (tendo em vista um bom balanceamento das classes previstas), precisão e *recall*.

- O segundo é a escolha do melhor modelo dentre os finais de cada tipo. Neste processo, levamos em consideração principalmente a curva **ROC** e a área sobre esta curva, a **AUC**, com a primeira levando em consideração os cálculos das taxas

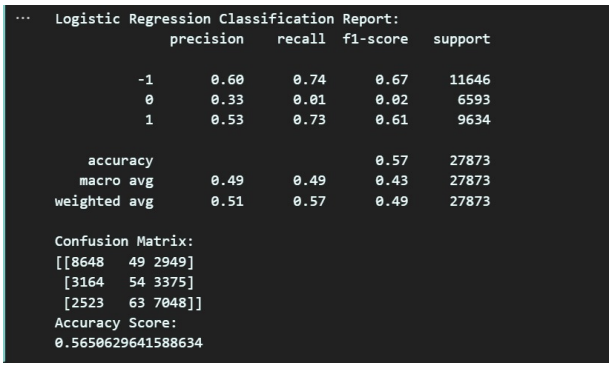


Figura 1 – Métricas de avaliação - Regressão Logística

de **verdadeiros positivos** e **falsos positivos** para a sua construção, levando em consideração os diferentes *tresholds* que podem ser aplicados para a classificação da variável objetivo.

Para esta primeira etapa do projeto, optou-se por utilizar apenas um modelo de cada tipo, com as mesmas variáveis. Assim, foi utilizada uma combinação dos métodos de avaliação para escolher o "*best model*".

Resultados

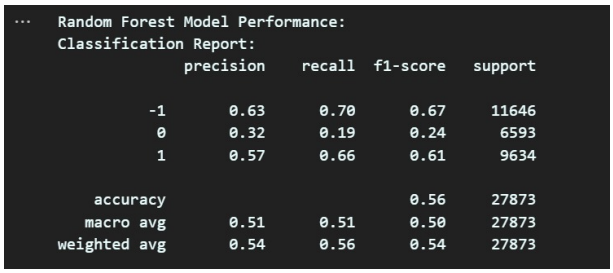
Após todo o processo padrão de um projeto de aprendizagem supervisionada, que inclui a coleta, pré-processamento, limpeza e transformações nos dados, com posterior aplicação dos modelos, os resultados são apresentados nesta seção.

Regressão Logística

A regressão logística alcança uma acurácia ligeiramente superior (0,57), mas falha drasticamente na classe média, com F1-score de apenas 0,02 e recall de 0,01 (54 corretos de 6593). Isso indica que o modelo praticamente não identifica a classe média, classificando-a como baixa (3164) ou alta (3375). Para as classes baixa e alta, o desempenho é bom (F1 de 0,67 e 0,61), mas a incapacidade de lidar com a classe média sugere que o modelo não captura bem a natureza ordinal dos dados, provavelmente por tratar as classes como independentes.

Random Forest

O modelo de floresta aleatória tem acurácia de 0,56, similar aos outros. Ele apresenta desempenho sólido para baixa e alta (F1 de 0,67 e 0,61), mas na classe média o F1-score é 0,24, melhor que a regressão logística (0,02), mas inferior à rede neural (0,37). O recall de 0,19 sugere que identifica mais instâncias médias que a regressão logística, mas ainda

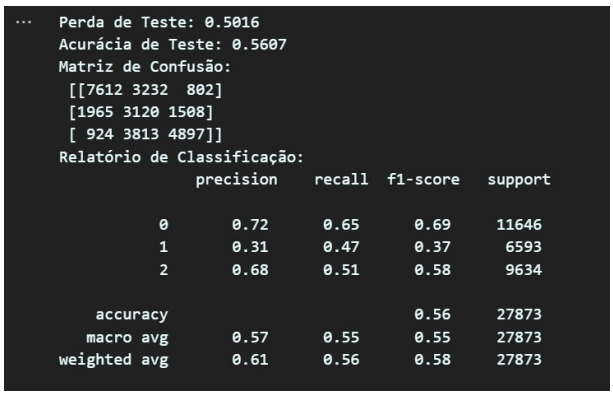


```
Random Forest Model Performance:
Classification Report:
              precision    recall  f1-score   support

          -1       0.63      0.70      0.67     11646
           0       0.32      0.19      0.24      6593
           1       0.57      0.66      0.61      9634

    accuracy                           0.56     27873
   macro avg       0.51      0.51      0.50     27873
weighted avg       0.54      0.56      0.54     27873
```

Figura 2 – Métricas de Avaliação - Random Forest Classifier



```
Perda de Teste: 0.5016
Acurácia de Teste: 0.5607
Matriz de Confusão:
 [[7612 3232  802]
 [1965 3120 1508]
 [ 924 3813 4897]]
Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.72      0.65      0.69     11646
           1       0.31      0.47      0.37      6593
           2       0.68      0.51      0.58      9634

    accuracy                           0.56     27873
   macro avg       0.57      0.55      0.55     27873
weighted avg       0.61      0.56      0.58     27873
```

Figura 3 – Métricas de Avaliação - Rede Neural

menos da metade em comparação com a rede neural. Isso indica que, embora capture algumas relações não lineares, o modelo não diferencia bem a classe média.

Rede Neural

O modelo de rede neural apresenta desempenho moderado com acurácia de 0,5607. Ele se destaca na classe média (F1 = 0,37), que parece ser a mais desafiadora, acertando 3120 de 6593 instâncias (recall de 0,47). Para as classes baixa e alta, o desempenho é sólido (F1 de 0,69 e 0,58, respectivamente), mas há confusão significativa entre classes adjacentes, como baixa sendo classificada como média (3232) e alta como média (3813). O test loss de 0,5016 sugere que o modelo poderia ser ajustado para melhorar a generalização.

Avaliação da Curva ROC e da AUC

A Curva ROC consiste no conjunto de pontos formados pela combinação da Taxa de Verdadeiros Positivos e pela Taxa de Falsos Positivos, apresentados como sensibilidade e especificidade. Já o AUC representa a área sob a curva ROC, e é o que de fato é utilizado como critério de escolha de modelos classificatórios, pois quanto maior a AUC, maior a probabilidade de o modelo classificar corretamente.

Considerações Finais

A partir do desenvolvimento do projeto, pode-se concluir que a integração das etapas de pré-processamento, transformação dos dados e modelagem resultou em uma análise robusta e multifacetada da preocupação financeira dos respondentes. Foram observados os seguintes pontos:

Limpeza e ETL Eficientes: A etapa de limpeza dos dados, que incluiu a correção de variáveis binárias, imputação de valores ausentes e transformação de variáveis (como a soma dos scores financeiros), foi essencial para garantir a integridade dos dados e viabilizar análises subsequentes.

Engenharia de Features e Criação do Rótulo: A conversão da variável financeira para um rótulo ordinal (worry level) possibilitou traçar uma classificação que reflete níveis distintos de preocupação financeira. Esse processo envolveu técnicas de transformação como a aplicação de funções lambda e one-hot encoding para variáveis categóricas (país).

Modelagem Preditiva: Foram exploradas abordagens diferentes para a resolução do problema, demonstrando que:

A regressão logística permite uma interpretação direta dos coeficientes e a identificação das principais features impactantes; O Random Forest capta relações não lineares, contribuindo para uma melhoria na performance preditiva; A implementação da rede neural para regressão ordinal oferta uma estratégia inovadora ao tratar a natureza ordinal dos rótulos, utilizando técnicas de aprendizado profundo com conversão dos rótulos para binários cumulativos.

Avaliação e Validação dos Modelos: O uso de métricas como acurácia, matriz de confusão e relatórios de classificação possibilitou a comparação dos modelos e evidenciou o potencial das abordagens aplicadas, além de ressaltar a importância de ajustes de hiperparâmetros e a definição de estratégias robustas de validação.

Limitações

Este estudo apresentou limitações quanto ao uso da base de dados, pois, no intuito de entregar um projeto original e com dados provenientes de fontes que não tenham os tratado, o potencial da pesquisa não foi totalmente explorado como deveria.

Perspectivas Futuras

Para as próximas estapas, a ideia é permanecer no mesmo tema, mas com uma possível mudança nos dados utilizados, para que haja uma melhora na interpretabilidade dos modelos.

Referências

BUSSAB, W. O.; MORETTIN, P. A. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017.

JAMES, G. *et al.* **An introduction to statistical learning: with applications in R**. Nova York: Springer, 2013.

BUDUMA, N.; LOCASCIO, N. **Fundamentals of deep learning: designing next-generation machine intelligence algorithms**. Beijing: O'Reilly Media, 2017.

ANBIMA. Mais da metade da população sente alto nível de estresse com as suas finanças, diz pesquisa da ANBIMA. Disponível em: ⟨https://www.anbima.com.br/pt_br/imprensa/mais-da-metade-da-populacao-sente-alto-nivel-de-estresse-com-as-suas-financas-diz-pesquisa-da-anbim htm⟩. Acesso em: 10 abr. 2025.

ZHANG, Y. *et al.* Unlocking financial literacy with machine learning: a critical step to advance personal finance research and practice. *Technology in Society*, v. 81, 2025. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0160791X24003452⟩. Acesso em: 10 abr. 2025.

RYU, S.; FAN, L. The relationship between financial worries and psychological distress among U.S. adults. *Journal of Family and Economic Issues*, v. 44, n. 1, p. 16–33, 2023. DOI: ⟨https://doi.org/10.1007/s10834-022-09820-9⟩.

LEVANTESI, S.; ZACCHIA, G. Machine learning and financial literacy: an exploration of factors influencing financial knowledge in Italy. *Journal of Risk and Financial Management*, v. 14, n. 3, p. 120, 2021. DOI: ⟨https://doi.org/10.3390/jrfm14030120⟩.