ENTREGA DA SEMANA 3 - Esboço da Apresentação de Pitch

1. Objetivos [semana 1]

Problema:

A Instituição Sirius, ao expandir sua oferta de cursos de MBA, enfrenta o desafio de atrair alunos de diversas localidades. O desafio é determinar quais municípios têm o maior potencial para atrair estudantes para seus programas. Gastos mal direcionados em marketing e esforços de atração em locais com baixo retorno potencial podem resultar em perda de recursos e oportunidades perdidas em locais mais promissores.

Questão-chave:

Como selecionar estrategicamente municípios com alto potencial para atração de alunos para os cursos de MBA da Instituição Sirius, utilizando dados para embasar a decisão, de forma a maximizar o ROI (retorno sobre investimento) e minimizar riscos e gastos desnecessários?

Objetivos:

- Desenvolver um Modelo Preditivo: Criar um modelo que possa prever a probabilidade de uma cidade ter um alto potencial de atração de alunos para os cursos de MBA com base em uma série de características demográficas, socioeconômicas e outros indicadores relevantes.

- Identificação de Cidades-Chave: Utilizando o modelo, identificar as cidades que possuem a maior probabilidade de sucesso na atração de estudantes e, por consequência, as que deveriam ser o foco dos esforços de marketing e atração.

2. Acesso aos dados

Nossos dados provêm de duas fontes principais:

1. Dataset Público de Cidades:

- Descrição: Este dataset oferece uma visão ampla sobre várias cidades brasileiras, destacando variáveis que podem influenciar sua produtividade e, possivelmente, seu potencial para atrair alunos para cursos de MBA.

- Origem: Base de dados pública de indicadores socioeconômicos.

- Características: Inclui variáveis como densidade demográfica, PIB municipal, taxas de urbanização, entre outras.

2. Dataset Sirius - Alunos:

- Descrição: Informações sobre os alunos da Instituição Sirius, incluindo demografia, formação educacional e histórico na instituição.

- Origem: Fornecido diretamente pela Instituição Sirius.

- Características: Contém detalhes como cidade de origem, nível de escolaridade, situação atual como aluno, entre outros.

Processo de Aquisição:

- Dataset Público: Acesso direto através de plataformas de dados abertos do governo ou instituições de pesquisa relevantes. Estes dados são geralmente atualizados anualmente e são de domínio público.

- Dataset Sirius: A equipe de TI da Instituição Sirius forneceu os dados em um formato tabular. Antes da entrega, foi realizado um processo de anonimização para garantir a privacidade e segurança dos dados dos alunos. O dataset foi verificado para assegurar que não contém informações pessoalmente identificáveis.

Integração dos Dados:

A integração entre os dois conjuntos de dados foi realizada usando as colunas 'cidade' e 'estado'. Esta integração permitiu uma comparação direta entre as características gerais das cidades e a presença (ou ausência) de alunos da Instituição Sirius nesses locais.

Desafios e Soluções:

- Qualidade dos Dados: Como sempre, ao trabalhar com dados reais, enfrentamos desafios relacionados à qualidade, como valores ausentes ou inconsistências. Implementamos processos de limpeza e pré-processamento para mitigar esses problemas.

- Tamanho do Dataset: Dada a grande quantidade de cidades e variáveis, foi necessário um trabalho cuidadoso de seleção e filtragem para focar nas informações mais relevantes para nossos objetivos.

- Privacidade: Garantir a privacidade dos alunos foi primordial. Ao receber o dataset da Instituição Sirius, garantimos que nenhuma informação pessoalmente identificável estava presente.

Em resumo, o acesso e integração dos dados foram etapas cruciais deste projeto, permitindo-nos obter insights valiosos e fundamentar nossas análises em informações concretas e confiáveis.

3. Divisão dos dados [semana 1]

Método de Divisão:

Para garantir que nosso modelo seja generalizado e performe bem em dados não vistos, dividimos o conjunto de dados em dois subconjuntos: treino e teste. Isso nos permite treinar nosso modelo no conjunto de treino e depois avaliar seu desempenho no conjunto de teste.

Estratégia:

Estratificação: Dada a natureza desbalanceada dos dados, optamos por uma divisão estratificada para garantir que ambas as classes (cidades com alunos da Sirius e cidades sem alunos da Sirius) estivessem bem representadas em ambos os subconjuntos.

Tamanho da Divisão: O conjunto de dados foi dividido em 80% para treino e 20% para teste. Este é um split comum que geralmente fornece um bom equilíbrio entre ter dados suficientes para treinar o modelo e ainda ter uma quantidade significativa de dados para testar.

Desbalanceamento:

Identificamos um desbalanceamento no dataset, onde as cidades que tinham alunos da Sirius eram muito menos frequentes do que as que não tinham.

Para tratar este desequilíbrio, optamos por usar o método SMOTE (Synthetic Minority Over-sampling Technique). Esta técnica gera exemplos sintéticos da classe minoritária. Após a aplicação do SMOTE, obtivemos um conjunto de dados balanceado, com número igual de exemplos em ambas as classes.

4. Exploração dos dados [semana 1]

A análise exploratória de dados (EDA) é fundamental para entender as características dos dados, identificar tendências, anomalias e padrões. Essa compreensão é crucial antes de qualquer modelagem. A seguir estão os principais pontos que exploramos:

Visão Geral dos Dados:

- Tipos de Dados: Verificamos os tipos de dados de cada coluna para garantir que estavam no formato correto para modelagem.

- Valores Ausentes: Analisamos o dataset em busca de valores ausentes para decidir sobre estratégias de tratamento ou remoção.

- Distribuição dos Dados: Analisamos a distribuição dos dados, especialmente a variável alvo (se a cidade tem ou não alunos da Sirius), para entender o equilíbrio entre as classes.

Análise Univariada:

- Histogramas e Boxplots: Utilizamos esses gráficos para entender a distribuição de cada variável numérica.

- Gráficos de Barras: Para variáveis categóricas, visualizamos a frequência de cada categoria.

Análise Bivariada:

- Correlação: Calculamos a correlação entre as variáveis numéricas para entender suas relações. Isso ajudou a identificar qualquer multicolinearidade que poderia afetar o desempenho do modelo.

- Tabelas de Contingência: Para variáveis categóricas, analisamos como elas se relacionavam com a variável alvo.

Insights Iniciais:

- Cidades com Alunos da Sirius: Identificamos que cidades com determinadas características (como densidade demográfica, PIB municipal e nível educacional) tendem a ter mais alunos da Sirius.

- Cidades Sem Alunos da Sirius: Observamos que algumas cidades, apesar de terem bons indicadores socioeconômicos, ainda não tinham representação na base de alunos da Sirius. Estas cidades são potenciais alvos para futuras campanhas.

Conclusões da EDA:

A EDA revelou várias características importantes sobre nossos dados. Ficou claro que existem padrões e tendências que podem ser explorados para prever com eficácia quais cidades têm maior potencial para atração de alunos. Também identificamos áreas onde dados adicionais ou mais detalhados poderiam ser úteis, como informações mais específicas sobre a economia local ou o perfil demográfico da população jovem.

5. Qual modelo implementado [semana 2]

Após a análise exploratória de dados e o pré-processamento, avançamos para a fase de modelagem. Aqui está uma descrição concisa do modelo que implementamos:

Seleção do Modelo:

- XGBoost: Optamos por utilizar o XGBoost, um algoritmo de boosting que se mostrou altamente eficiente e eficaz em uma variedade de tarefas de aprendizado de máquina. Este modelo é conhecido por sua capacidade de lidar com grandes conjuntos de dados, lidar com dados faltantes e por sua flexibilidade para ser afinado.

Justificativa para a Escolha:

- Desempenho: Em nossos testes preliminares, o XGBoost superou outros modelos tradicionais, como regressão logística, florestas aleatórias e máquinas de vetores de suporte em termos de métricas de desempenho.

- Flexibilidade: O XGBoost oferece uma ampla gama de hiperparâmetros que podem ser ajustados para melhorar o desempenho e evitar o overfitting.

- Tratamento de Classes Desbalanceadas: Como observamos um desequilíbrio nas classes da variável alvo, o XGBoost, com seus parâmetros específicos para tratamento de desbalanceamento, se mostrou uma excelente escolha.

Configuração do Modelo:

- Parâmetros: Foram utilizados parâmetros específicos para maximizar o desempenho, como a taxa de aprendizado, profundidade máxima da árvore.

- Validação Cruzada: Para garantir que o modelo não estivesse apenas se adaptando bem aos dados de treino, utilizamos a técnica de validação cruzada.

Hiperparâmetros:

Realizamos uma busca de hiperparâmetros usando a técnicas de Grid Search para encontrar a melhor combinação que otimizasse nosso desempenho.

Racional para a Escolha do Modelo:

O XGBoost foi escolhido devido à sua capacidade de lidar eficientemente com datasets desequilibrados, sua flexibilidade em tunar múltiplos parâmetros, e seu histórico comprovado em oferecer desempenho de alto nível em competições de ciência de dados e aplicações práticas. A métrica de avaliação, o "logloss", foi selecionada por ser apropriada para problemas de classificação, especialmente quando a probabilidade das predições é relevante.

O modelo mostrou-se extremamente eficaz, alcançando uma acurácia de 100% nos dados de teste, indicando um ótimo desempenho na diferenciação entre as classes. Porém, vale mencionar que tamanha precisão pode também levantar suspeitas quanto ao overfitting, que é quando o modelo se ajusta demais aos dados de treino e pode ter problemas ao generalizar para novos dados. Isso seria um ponto a ser investigado em etapas futuras.

6. Prós e contras do modelo [semana 2]

Prós:

1. Desempenho de ponta: O XGBoost é frequentemente reconhecido por seu desempenho superior em comparação com outros algoritmos em competições de aprendizado de máquina, como as do Kaggle.

2. Paralelização: O XGBoost foi projetado para ser altamente eficiente. Ele utiliza a paralelização para construir as árvores, tornando o processo de treinamento mais rápido.

3. Flexibilidade: Pode ser usado para problemas de regressão, classificação, ranking e previsão de preferência do usuário.

4. Trata dados faltantes: O XGBoost tem uma maneira inerente de lidar com valores faltantes.

5. Regularização: O algoritmo tem regularização L1 (Lasso) e L2 (Ridge) integrada, que previne o overfitting.

6. Capacidade de Pruning: Ao contrário de outras árvores de decisão, o XGBoost pode 'podar' suas árvores, removendo ramos que não proporcionam poder de previsão.

7. Cross-validation integrada: O XGBoost permite que o usuário execute uma validação cruzada fold-by-fold após cada iteração do processo de boosting, fornecendo uma visão clara de qual iteração proporciona o melhor desempenho.

Contras:

1. Complexidade: Devido à sua ampla gama de hiperparâmetros, pode ser complexo e demorado afinar o modelo para obter o melhor desempenho.

2. Recursos computacionais: Embora seja eficiente, o XGBoost pode ser intensivo em termos de recursos computacionais, especialmente com grandes conjuntos de dados e/ou quando é usado com uma grande quantidade de árvores.

3. Overfitting: Sem uma configuração adequada, o modelo pode se ajustar demais aos dados de treinamento. É essencial usar técnicas como validação cruzada e regularização para mitigar isso.

4. Interpretabilidade: Enquanto o XGBoost oferece importâncias de características e outras ferramentas, como SHAP, para interpretação, ele não é tão intuitivo ou fácil de visualizar quanto uma única árvore de decisão ou uma regressão linear.

5. Tamanho do modelo: Os modelos XGBoost podem se tornar grandes e levar a problemas de memória, especialmente quando o número de árvores é alto.

7. Conclusão - achados e previsões [semana 3/4]

Ao final do projeto, após a implementação do modelo XGBoost, ajuste fino dos hiperparâmetros e considerando as variáveis explicativas disponíveis no dataset, foi possível identificar as cidades que apresentam o maior potencial para atração de alunos para os cursos de MBA da Instituição Sirius.

Os achados indicam uma tendência clara: cidades de grande e médio porte, que muitas vezes são capitais de estado ou centros regionais, apresentam maior probabilidade de atração de alunos. Esse comportamento pode estar associado a diversos fatores, como maior densidade populacional, concentração de indústrias e empresas, e, consequentemente, um maior número de profissionais buscando qualificação.

Top 10 cidades com maior probabilidade de atração de alunos:

1. Campinas, São Paulo - Probabilidade: 99.98%

2. Curitiba, Paraná - Probabilidade: 99.95%

3. Florianópolis, Santa Catarina - Probabilidade: 99.95%

4. São Paulo, São Paulo - Probabilidade: 99.94%

5. Belo Horizonte, Minas Gerais - Probabilidade: 99.94%

6. Brasília, Distrito Federal - Probabilidade: 99.93%

7. Goiânia, Goiás - Probabilidade: 99.92%

8. Recife, Pernambuco - Probabilidade: 99.90%

9. Porto Alegre, Rio Grande do Sul - Probabilidade: 99.88%

10. Balneário Camboriú, Santa Catarina - Probabilidade: 99.87%

Estes resultados fornecem insights valiosos para a instituição, possibilitando estratégias de marketing e investimento mais direcionadas. Por exemplo, ações promocionais, parcerias com empresas locais ou até mesmo a consideração para abertura de novos campi podem ser estratégias consideradas para as cidades identificadas.

O modelo, contudo, não se restringe apenas a estas dez cidades. A lista completa, que foi gerada, pode ser usada para estratégias segmentadas de acordo com o potencial de cada cidade. Isso permite que a Instituição Sirius otimize seus esforços e recursos, buscando atingir o público-alvo de maneira mais eficiente e eficaz.

Com isso, concluímos que a análise baseada em dados, aliada à aplicação de técnicas avançadas de aprendizado de máquina, pode oferecer soluções concretas e alicerçadas em informações confiáveis, tornando-se uma ferramenta indispensável no processo decisório da instituição.

8. O que o grupo faria de diferente?

Refletindo sobre a trajetória do projeto e os resultados alcançados, identificamos algumas áreas de melhoria e abordagens alternativas que poderiam ter sido consideradas:

1. Maior Integração com a Área de Negócios: Ter uma colaboração mais próxima com a equipe de marketing e vendas da Instituição Sirius poderia ter nos proporcionado insights mais profundos sobre os desafios e as expectativas do negócio.

2. Diversificação de Modelos: Embora o XGBoost tenha apresentado ótimos resultados, poderíamos ter experimentado uma variedade maior de modelos, como redes neurais ou técnicas de ensemble, para comparar o desempenho.

3. Análise Geoespacial: Considerando que estamos lidando com cidades, uma análise geoespacial poderia ter sido benéfica. Isso nos permitiria visualizar clusters geográficos de interesse e entender padrões regionais.

4. Coleta de Mais Dados: Ainda que os dados disponíveis tenham sido suficientes, sempre há espaço para incorporar mais informações. Por exemplo, dados socioeconômicos mais detalhados, informações sobre infraestrutura local ou até mesmo pesquisas de satisfação de ex-alunos.

5. Engenharia de Recursos: Poderíamos ter investido mais tempo criando e selecionando recursos que poderiam influenciar a atração de alunos, como a proximidade de empresas de grande porte ou a presença de outras instituições de ensino.

Em resumo, embora estejamos satisfeitos com os resultados alcançados, reconhecemos que sempre há espaço para melhorias. Cada projeto é uma oportunidade de aprendizado e, com os insights adquiridos, estamos confiantes de que abordagens futuras serão ainda mais bem-sucedidas.