Uma imagem com texto, design gráfico, Gráficos, Tipo de letra

Descrição gerada automaticamente

ISAG - Instituto Superior Administração e Gestão

Pós-Graduação em Data Science e Business Intelligence

**Projeto Final**

**Modelo de Previsão de Pagamentos em Atraso:**

Identificação Proativa de Clientes em Risco

Porto, maio de 2024

Realizado por: Sara Helena M. Oliveira Silva

Orientado por: Ivo Nogueira e Rita Faria

**Índice**

1. Compreensão do Negócio *(Business Understanding)*
   1. Motivação Projeto
   2. Objetivos
   3. Entendimento do contexto do negócio
   4. Benchmarking (o que existe no mercado que faça esta previsão)
   5. Identificação do melhor método de desenvolvimento (diferentes métodos disponíveis para desenvolver o projeto, como modelos de machine learning, análise estatística ou abordagens baseadas em regras)
   6. Justificação da escolha do método (porque o método escolhido é o mais apropriado? Como pode alcançar os objetivos? Alinhado com as tendências e práticas)
2. Compreensão dos dados *(Data Understanding)*
   1. Identificação das fontes de dados
   2. Descrição dos dados (variáveis e seu significado bem como tipo e valores possíveis)
   3. Exploratory Data Analysis (EDA)
   4. Distribuições, tendências e relações
3. Preparação dos Dados *(Data Preparation)*
   1. Missing values e outliers
   2. Os dados estão bem distribuídos?
   3. Análise de Correlação
   4. Clustering
   5. Tratamento de variáveis categóricas
   6. Normalização
   7. Feature enginnering: é pertinente criar novas variáveis? (atraso médio de pagamento, razão de ausência de pagamento)
4. Modelação *(Modelling)*
   1. Analisar que modelo é mais adequado para o problema
   2. Divisão dos dados entre treino e teste
   3. Treino dos modelos (ajuste de parâmetros)
   4. Comparação de resultados (MLFlow?)
   5. Checar se é necessário fazer regularização.
   6. Otimização
5. Avaliação / Resultados
   1. Perceber que métricas são mais apropriadas para avaliação do modelo
   2. Avaliar desempenho do modelo
6. Implementação do modelo
   1. Como empresas podem implementar o modelo em ambiente de produção
   2. Atualização automática dos dados e do modelo
   3. Implementar sistema de monitorização contínua para controlo do modelo tanto para a empresa como para developers (carta de controlo)
7. Conclusão (importância da ética)
8. Anexos

1. **Compreensão do Negócio *(Business Understanding)***

O atraso no pagamento de faturas é um desafio comum e transversal a muitas empresas que, por consequência, pode causar problemas sérios de liquidez e impactar negativamente o desempenho financeiro das organizações. Perante este cenário, a capacidade de identificar proactivamente os clientes mais propensos a efetuarem o pagamento das suas faturas torna-se num desafio que, em caso de correta previsão, torna-se fundamental para a saúde financeira de uma empresa e para um aumento na satisfação dos seus clientes.

* 1. **Motivação do Projeto**

Tendo em consideração este enquadramento, este projeto foi motivado pela vontade desenvolver um modelo de previsão de pagamentos em atraso que ajude as empresas a identificar antecipadamente os clientes que apresentam maior probabilidade de não efetuarem o pagamento das suas faturas dentro do prazo estabelecido.

Considero que este modelo, que será desenvolvido com técnicas de análise de dados e machine learning, tem um enorme potencial de se tornar uma ferramenta imprescindível e crucial para colmatar os riscos associados ao não pagamento de faturas para além de ser uma ferramenta que pode ajudar no trabalho aborrecido e cansativo das equipas que fazem gestão de cobranças nas organizações.

A metodologia adotada para o desenvolvimento deste projeto segue as orientações do processo CRISP-DM (*Cross-Industry Standard Process for Data Mining*) que engloba etapas de compreensão do negócio, compreensão e preparação dos dados, modelação e avaliação do modelo bem como proposta de como implementar o modelo em produção nas empresas.

* 1. **Objetivos**

O principal objetivo deste projeto é identificar clientes que são propensos a não efetuarem ou a atrasarem a efetuar o pagamento das faturas dos produtos ou serviços que usufruíram. Para este objetivo ser cumprido com sucesso, serão realizadas análises ao histórico de pagamentos dos clientes, identificação de padrões de gastos e segmentação de clientes por scores de risco.   
Desta maneira, as empresas podem tomar medidas preventivas e proactivamente ajudar os clientes com maior probabilidade de atrasarem ou não efetuarem o pagamento.

Todas as etapas necessárias para desenvolver este projeto serão implementadas utilizando a linguagem de programação *Python* e bibliotecas como Pandas, Numpy,

* 1. **Entendimento do contexto do negócio**
  2. **Benchmarking (o que existe no mercado que faça esta previsão)**
  3. **Identificação do melhor método de desenvolvimento (diferentes métodos disponíveis para desenvolver o projeto, como modelos de machine learning, análise estatística ou abordagens baseadas em regras)**
  4. **Justificação da escolha do método (porque o método escolhido é o mais apropriado? Como pode alcançar os objetivos? Alinhado com as tendências e práticas)**

1. **Compreensão dos dados (Data Understanding)**

Nesta primeira fase do projeto, compreender os dados disponíveis é muito importante uma vez que servirão de base para o desenvolvimento do modelo proposto.

Durante esta etapa de compreensão de dados, é crucial estudar a estrutura do conjunto de dados (*dataset)* escolhido: que variáveis estão disponíveis, de que tipo são bem como o seu significado e que valores possíveis podem ter. Assim que este estudo inicial for realizado, será realizada uma análise exploratória dos dados para verificar distribuição das variáveis, deteção de valores omissos (*missing values*), *outliers* bem como analisar possíveis tendências e relações entre as variáveis.

* 1. **Identificação das fontes de dados**

O *dataset* selecionado foi extraído da plataforma *Kaggle*. Trata-se de uma plataforma amplamente reconhecida por ter à disposição conjuntos de dados com elevada qualidade para diversos projetos na área de *data science* e *business inteligence*.

Relativamente ao tipo de ficheiro, o dataset está no formato CSV *(Comma-separated values)*, um formato de arquivo de texto muito utilizado para armazenar dados tabulares.

* 1. **Descrição dos dados**

Este *dataset* contém informações sobre o comportamento de clientes relativamente ao pagamento da fatura de um serviço que adquiriu. Cada entrada (*data point)* representa uma fatura emitida com datas de emissão compreendidas entre 01-01-2023 e 09-09-2023, identificada pelo número da fatura *(invoiceNumber)* e respectivo cliente através do seu *customerID.*

O *dataset* é composto por 12 variáveis e 2466 visualizações *(data points).* As variáveis que compõem o *dataset*, bem como o seu tipo e o que representam, são:

|  |  |  |
| --- | --- | --- |
| **Nome da variável** | **Tipo** | **Significado** |
| countryCode | int64 | Código do país do cliente. |
| customerID | object | Identificação Única do Cliente. |
| PaperlessDate | datetime | Data em que cliente consentiu fatura digital. |
| invoiceNumber | int64 | Número da fatura. |
| InvoiceDate | datetime | Data da fatura. |
| DueDate | datetime | Data Vencimento da fatura. |
| InvoiceAmount | float64 | Valor total da fatura. |
| Disputed | object | Indica se fatura foi contestada pelo cliente.  Valores Possíveis: (‘Yes’ (fatura contestada) ou ‘No’ (fatura não contestada) |
| SettledDate | datetime | Data em que fatura foi paga ou contestação resolvida. |
| PaperlessBill | object | Indica se cliente recebe fatura digital. Valores Possíveis: (‘Paper’ (fatura em papel) ou ‘Electronic’ (fatura eletrónica). |
| DaysToSettle | int64 | Número de dias que cliente demorou para efetuar pagamento. |
| DaysLate | int64 | Número de dias que fatura ficou em atraso em relação à data de vencimento. |

Estes dados servirão de base para desenvolver este projeto. Para além disso, são importantes para passarmos à fase da análise exploratória dos dados onde se vai analisar a distribuição das variáveis, o comportamento de pagamento dos clientes, perceber se existem padrões de atraso que possam indicar um maior risco de atraso nos pagamentos, entender se existe diferença entre pagamento com fatura eletrónica ou em papel entre outras análises.

* 1. **Exploratory Data Analysis (EDA)**

- Missing Values: Durante esta análise inicial, analisou-se a presença de valores omissos mas conclui-se que não existem nem valores omissos nem observações idênticas ou duplicadas.

- *Outliers:* Existem 259 *outliers* no *dataset* (*boxplot* das variáveis onde se pode verificar a presença de *outliers* nos anexos)*.* Estes *outliers* foram detetados recorrendo ao método do intervalo interquartil (IQR) que permitiu identificar observações que se encontram significativamente distantes dos quartis das respetivas variáveis.

Embora tenha identificado presença de *outliers* no *dataset,* devido ao contexto específico do modelo que vou desenvolver, decidi manter estes valores atípicos uma vez que considero importante existir uma variedade de situações que podem influenciar o comportamento dos clientes. Desta forma, acredito que o modelo vai ser capaz de lidar com estes casos extremos e prever o atraso dos pagamentos numa grande variedade de cenários.

Apesar disto, considero fazer testes com o *dataset* com e sem os *outliers* para perceber se a performance melhora com a exclusão destes.

* 1. **Distribuições, tendências e relações**

**ANEXOS**

**Uma imagem com diagrama, Retângulo, file, captura de ecrã

Descrição gerada automaticamente**