Tecnologias Avançadas de Dados - Bank Marketing Campaigns

Filipa Capela nº 2018297335/ Tiago Conceição nº 2021167993

December 14, 2022

1 Introdução

Atualmente muitas empresas tem que inovar as suas estratégias para obtenção de novos clientes. Algumas dessas estratégias passam por criação de campanhas de marketing. Uma campanha de marketing de uma empresa é uma tarefa efetuada com o objetivo de promover algum tipo de serviço, produto ou marca. Com a divulgação desta campanha, e a direcioná-la a potenciais clientes, o objetivo é de incrementar o número de vendas.

As campanhas de marketing podem ser promovidas de diversas formas, nomeadamente através de publicidade distribuídas através de vários canais de informativos, telefone ou até email. No entanto para uma campanha de marketing ter sucesso, é necessário realizar um estudo intensivo ao público alvo, para que a população selecionada seja mais propícia à adesão das campanhas.

Na atualidade, os bancos têm que inovar para alcançar novos clientes para os seus serviços. Desta forma, e no contexto do projeto de análise ao banco Portugal, a principal via de divulgação usada foram os meios telefónicos.

Após uma primeira análise, uma das conclusões retiradas foi a ineficiência desta campanha, obtendo apenas 11% de adesão. Tendo em conta todos os recursos gastos, não só a nível monetário mas como também a nível de tempo gasto pelos trabalhadores, vemos que as campanhas desta maneira acabam por não compensar.

Desta forma a equipa Insightia, visa melhorar os resultados obtidos para as futuras campanhas, realizando um estudo mais aprofundado de quais as características que levam a população a aceitar.

2 Informação Inicial

Neste capítulo, vão ser abordados os temas fundamentais para a realização do projeto, começando pela a apresentação inicial e finalizando com o poster informativo.

2.1 Pitch

Hoje em dia, das distintas formas que existem para um banco fazer dinheiro, 3 destacam-se sendo elas:

- Capital de interesse;
- Mercados de capitais;
- Rendimento baseado em taxas.

O capital de interesse é a principal fonte de rendimento do banco pois utilizam dinheiro de outras contas que não necessitam do dinheiro no momento atual. Em troca, recebem juros de acordo com a quantidade de dinheiro que emprestam. Outra fonte de comissões é através do mercado de capitais, que consiste no empréstimo de dinheiro a empresas e investidores quando querem abrir um negócio, ou quando querem promover o crescimento do negócio, querendo assim o retorno do capital com juros. Por fim, o rendimento baseado em taxas ocorre quando os bancos cobram taxas pelos seus serviços. Quando, uma entidade quer abrir uma conta, o banco pode cobrar taxas mensais para manter a conta aberta.

Tendo em conta estas fontes de rendimento a abertura de uma conta a prazo,faz com que o banco ganhe dinheiro através das taxas.

Assim, foi realizado um estudo sobre uma campanha de marketing do banco Portugal para a abertura de uma conta a prazo.

Estes dados foram retirados do website kaggle [1], uma fonte que contém uma grande variedade de datasets com diferentes temas, que permite a obtenção e a recolha de dados disponível a todos.

Realizada uma primeira análise, chegou-se à conclusão que efetivamente a campanha efetuada não obteve os melhores resultados.

Como é possível verificar na imagem 1, a quantidade de pessoas que aderiram à campanha corresponde ao número irrisório de 11% e as que decidiram não aderir equivale ao valor de 89%. Dado que a população total abordada, foi de 41.118 mil pessoas, apenas 4640 decidiram aderir, enquanto os restantes 36.548 não aceitaram como se pode analisar na figura 2

Dado o grande tamanho da campanha efetuada, e a sua duração, também se conclui que houve um grande desperdício de recursos. Como não foi disponibilizado dados relativos ao investimento executado pelo banco, só se consegue realizar uma análise à duração em dias mensal gasta na população estudada que rejeitou a campanha.

Observando a imagem 3, pode-se retirar que efetivamente ao longo do ano de 2014 a quantidade de tempo gasto foi bastante elevada. Nomeadamente nos meses de junho e agosto, o tempo gasto pelo departamento responsável pela campanha, corresponde a 75 dias. Uma vez realizados os cálculos, o tempo investido nesta campanha totaliza 411 dias, ou seja durou mais 46 dias que a duração de um ano regular.

Realizada a primeira abordagem, ao projeto e tendo em conta todos os seus dados analisados até ao momento, surgem algumas questões que a equipa Insightia vai tentar responder, tais como:

- Porque é que o banco obtém resultados tão maus, quando se trata de adesão à campanha;
- É possível fazer uma análise demográfica aprofundada, para compreender se os diferentes atributos afetam a tomada de decisão;
- Compreender se os dias da semana e os meses têm resultados diferentes em conformidade com a campanha, e porquê;
- É possível fazer um modelo de previsão, para ajudar os bancos a poupar tempo e recursos e, ao mesmo tempo, aumentar os rendimentos

As questões reunidas visam obter alguma informação útil, para o banco e para a equipa de marketing. Por exemplo, obter detalhes sobre a população que aderiu e tentar compreender que aspetos afetam a decisão. Também realizar uma análise sobre as taxas e os seus valores ao longo dos meses.

Uma vez recolhidos os dados fulcrais para avançar com o projeto em si, é proposto uma pipeline de solução que visa ajudar no problema.

Como é possível averiguar na imagem 5, a solução passa por vários passos fundamentais como a parte de obtenção dos dados, que como referido anteriormente já está definido, uma primeira análise do problema também já foi observada anteriormente. De seguida vai ser realizado um pré-processamento dos dados, que é uma etapa fulcral para qualquer tipo de projeto do mesmo género, posteriormente vai ser realizado uma extração dos dados para obter gráficos e demonstrações visuais para obtenção de respostas. Por fim, é escolhido um modelo de previsão que pretende obter uma antecipação do comportamento da pessoa estudada.

Dada a proposta de solução, os resultados esperados passam por trazer mais clientes para aderir à campanha, e ao mesmo tempo reduzir a quantidade de recursos gastos.

2.2 Poster Informativo

Ao ser efetuado uma segunda exposição de resultados desta vez em formato de poster, foi necessário repensar que dados compensariam demonstrar. Normalmente um poster é composto por informação mais concreta do caso de estudo a ser analisado, dado que já foi feita uma primeira analise sobre os dados, é necessário fazer um estudo mais profundo dos dados para expor.

Assim, após um estudo mais profundo dos dados a ser analisados, as seguintes imagens foram obtidas.

Na imagem 6 consegue-se afirmar que efetivamente a educação apresenta um fator decisivo na adesão da campanha, pois ao olhar com mais atenção para o gráfico radar as adesões são maioritariamente de pessoas com a educação de grau universitário.

Já na figura 7, conclui-se que durante o período de um ano os meses com maior adesão foram os meses de Maio até Agosto, e a partir dai sofreu uma grande queda.

Finalmente, foi realizada uma verificação sobre os valores das taxas e concluise que no mês de Julho a taxa euribor e a taxa de empregabilidade foram atingidos conforme se pode vê na figura 4. Para além disso, é notável que existe uma certa correlação entre ambas as taxas, seguindo a mesma tendência. É importante também mencionar, que a taxa euribor em Portugal remete tipos de taxas de juros aplicados a empréstimos, pelos bancos da zona euro, para além disso também está associado ao credito habitação. No caso da employment variation rate, ou em português indicador de empregabilidade define como é que naquele mês a taxa de empregabilidade

Tendo em conta as análises feitas até ao momento, verificou-se que a exposição destes gráficos não é a melhor, pois existem ferramentas que demonstram os dados de uma melhor forma. No entanto, para isso seria necessário passar por um processo que é demonstrado na figura 8.

A equipa Insightia decidiu para a parte do ETL usar o programa pentahoo, pois é uma ferramenta mesmo destinada para esse fim, desta forma pode-se simplificar o processo de extrair, transformar e carregar os dados. E para a exposição dos dados gerados no ETL, foi decidido usar o Tableau, por causa da sua interface e capacidade de criar visualizações.

Por fim, era importante demonstrar um mockup de como irá ser a demonstração dos dados, tal como na aparece na imagem 9

Como é possível ver na imagem 9, a visualização vai estar dividida em 3 grandes partes, uma dedicada relacionada com a informação pessoal, outra para a vida profissional, e outra para as taxas finalizando com uma secção de previsão. Nesta secção vai ser permitido a inserção de dados da pessoa para ser realizada uma previsão se vai aderir ou não à campanha.

3 Data Warehouse

Antes de se realizar qualquer tipo de tarefa no projeto, existe alguns conceitos que têm de ser definidos. Nomeadamente o star scheme (esquema estrela) do projeto, é uma abordagem que se usa com muita frequência no tema de data warehouses, pois permite expor os dados de uma maneira simples e de forma redundante. Um star scheme é composto por uma tabela de facto que expressa o evento principal da ocorrência, este modelo também compõe diversas tabelas de dimensão que são as características desses eventos. A tabela de fatos é composta por chaves que redireciona para as diferentes tabelas de dimensão. No contexto do nosso projeto o evento principal da ocorrência que se evidência e que faz sentido é a campanha, que contém:

- Last contact month: *months*;
- Last contact day of week: *day_of_week*;

- Number of contacts performed in the campaign: *n_contacts_campaign*;
- Number of contacts performed in previous campaign: *n_contacts_previous*;
- Previous Outcome: *poutcome*;
- Outcome: *outcome*;
- Duration of the previous call: *duration*;
- Days from previous contact: *pdays*;

Definir as características do evento principal, ou seja as tabelas de dimensão, foi mais desafiante pois, devido há pouca variedade de dados só se alcançar duas tabelas na primeira versão. Esta primeira versão apenas era composta pela tabela facto e duas tabelas de dimensão, sendo elas: Tabela que remete para os dados da pessoa.

- Age: *age*;
- Marital: *marital*;
- Credit: *credit*;
- Housing loan: *housing*;
- Personal Loan: *loan*;
- Education: *education*;
- Job: *job*.

Tabela das taxas.

- Employment Variation Rate: *emp_var_rate*;
- Consumer price index: *cons_price_idx*;
- Consumer confidence index: *cons_conf_idx*;
- Euribor: *euribor3m*.

Resumindo, nesta primeira versão apenas contínhamos uma tabela facto e duas tabelas de dimensão, o que não torna um esquema de estrela. Foi necessário repensar a estrutura do esquema, ao qual posteriormente pensou-se em dividir os dados da tabela dimensão que continha os dados da pessoa numa tabela que menciona apenas o aspeto da vida profissional da pessoa. Mas como os dados que remetem para a carreira da pessoa são apenas dois, optou-se por fazer uma mini-dimensão chamada vida profissional. Esta é composta por:

- Education: *education*;
- Job: *job*.

Ao criar esta tabela de dimensão garantimos que o esquema em estrela é obtido e que ao mesmo tempo obtemos granularidade dos dados, obtendo desta maneira a seguinte imagem 10.

Assim e já com o star scheme definido, vai ser mais simples dar respostas as questões definidas na secção 2.1, com as tabelas assim definidas. Uma vez que os dados mais relevantes demográficos estão evidenciados, será mais fácil entender através de visualizações algumas tendências e assim responder à pergunta dos diferentes atributos demográficos. Outra questão que facilmente será respondida com o facto dos dados se encontrarem desta forma, é a resposta para entender de que forma é que os dias da semana e os meses obtém resultados diferentes, a tabela das taxas vai permitir a resposta a essa pergunta. Por fim e com os dados todos reunidos e analisados vai ser possível ver que características são mais adequados para o modelo de predição, e também o porque do banco obter tão maus resultados.

4 ETL

Em qualquer projeto que envolva big data, é necessário passar por um processo de ETL (Extract, Load and Transform). Que significa extrair os dados pretendidos, transforma-los de modo a que seja possível inseri-los numa base de dados, e por fim carregar os dados para a base de dados. A aplicação selecionada para a tarefa de ETL foi o Pentaho Data Integration, por ser uma aplicação, com umas capacidades bastante desenvolvidas para as tarefas que se compromete a realizar. Posto isto, foi necessário estabelecer os passos fulcrais para a execução do ETL, e estes passos são compostos pelo tratamento de dados, processamento das tabelas de dimensão e por fim o processamento da tabela de facto. Na parte de tratamento de dados, foi realizado um upload dos dados da csv do projeto. Posteriormente através das capacidades do pentaho foi executada uma limpeza dos dados: removemos as colunas que não eram necessárias para o contexto deste trabalho: *contact*(contém a forma como o cliente foi contactado, que poderia ser *cellular* ou *telephone*). Também fizemos um filter das colunas, isto é, removemos os valores em todas as colunas a *NAN* ou a *unknown*. Fizemos também uma alteração dos valores da coluna, comparativamente com os valores das colunas que apareciam no dataset inicial do kaggle. Por exemplo, substituímos valores que à partida não eram intuitivos: trocámos *default* por *credit* e também nomes de atributos que continham "." no meio da palavra. Estes casos conteciam nomeadamente nos nomes dos atributos das taxas: trocámos *emp.var.rate* por *emp_var_rate*, entre outros. Seguidamente, fizemos um mapper value para alguns atributos, dado que, como referido anteriormente, continham "." no meio da palavra, causando posteriormente conflitos quando utilizávamos o tableau. Estes casos aconteciam nomedamente nos atributos da *education* e do *job*. Desta forma, os dados vão para a base de dados todos normalizados. O próximo passo passa por fazer a divisão das colunas do dataset inicial entre as diferentes tabelas dimensão de acordo com 3 e a tabela de facto. À medida que íamos criando as tabelas, tanto as das dimensões como a facto, colocávamos

uma primary key em cada tabela, uma ser possível interligá-las. Seguidamente, as tabelas eram armazenadas numa base de dados postgreSQL.

5 OLAP

Como em todos os projetos de Big Data existe uma fase de OLAP (Online Analytic Processing). Com o OLAP obtém-se a capacidade de manipular e analisar um grande volume de dados através de diversas perspetivas. Uma vez que neste projeto contamos com várias dimensões e com múltiplas características estas podem ser repartidas em apresentação, rastreio ou análise. OLAP extrái dados de múltiplos conjuntos de dados relacionais e reorganiza-os num formato multidimensional que permite um processamento muito rápido e uma análise muito perspicaz. Tal como tinha sido mencionado na secção de 3, contamos com uma tabela de facto e várias tabelas de dimensão formando nesta forma um cubo. E há diversas possibilidades de operações que podem ser executadas desta forma podendo alterar os dados os seus parâmetros de pesquisa.

6 Tableau

Finalmente, chegou a altura de apresentar a ferramenta escolhida para obter a exposição de dados. Esta ferramenta é conhecida por facilitar a visualização de dados para fim de *business inteligence* e através da sua simples interface. Uma das grandes vantagens desta ferramenta é que é bastante intuitiva. Em relação ao relacionamento entre tabelas, um requisito essencial no tableau para haver algum tipo de ligação entre elas. No entanto foi ai que notámos um outro grande desafio que foi o de como relacionar as tabelas, uma vez que não tínhamos hipótese através de argumentos de relacionar as diferentes tabelas, foi ai que se decidiu usar uma coluna que iria funcionar como index para relacionar. Uma vez já, com as relações estabelecidas chegou a altura de criar as visualizações, e decidimos tentar replicar ao máximo o mockup mostrado na figura 9. Portanto dividir em 3 principais campos, sendo que o primeiro era focado na pessoa e nas características que a compõem, o segundo campo era para as carreiras profissionais, e o terceiro para a variação das taxas aplicadas ao longo do ano. Todas estas visualizações estão relacionadas com as outras tabelas, até usam colunas das outras tabelas. No entanto ao realizar a visualização da vida profissional surgiu outro desafio que foi mudar um dos eixos dinâmicamente neste caso em concreto o eixo dos x, que foi resolvido recorrendo ao uso de parâmetros que quando é acionado através de uma checkbox ou dropdown menu, o eixo muda. Porém os desafios não ficaram por aqui, na visualização das taxas como inicialmente tínhamos exposto no mockup um gráfico de linhas seria uma solução irrealista pois a sua visualização seria bastante complicada de retirar informações e não iria conseguir mostrar as variações de valores ao longo dos meses. Desta forma e após uma pesquisa, e conforme as capacidades do tableau foi decidido a utilização de um boxplot acompanhada com um gráfico de barras que demonstra

o numero de aceitações por mês , pois assim conseguiríamos resolver o facto dos clareza dos dados e a variação. Visto que as visualizações já estavam todas as definidas, chegou a parte de mostrar as variações e possibilidades que os dados poderiam ter, portanto para cada uma das secções definidas, existem alguns filtros que vão permitir as diferentes perspetivas de dados. Um dos problemas encontrados quando se fechava e abria novamente o tableau é que para além do tempo de espera que se estava sujeito para abrir cada visualização, por vezes existia erro na ligação ao server postgreSQL. Resultando assim, num erro que não permitia progredir.

7 Machine Learning Model

Para o nosso OLAP ser mais completo, decidimos utilizar modelos preditivos para que, o cliente possa selecionar um conjunto de dados relativos a uma subscrição e verificar se à partida, uma pessoa aderir à campanha e abrir uma conta a prazo ou não. Assim sendo, selecionámos duas opções de modelos: o Random Forest e a Regressão Logística, dado que são bastante adequados para problemas de classificação. Antes de aplicarmos os modelos, necessitámos de normalizar os dados numéricos e converter os dados categóricos a dados numéricos. Posto isto, dividimos o dataset normalizado em dois conjuntos: conjunto de treino e conjunto de teste. Para cada um dos modelos aplicámos no conjunto de treino o *HalvingGridSearchCV()* da bibiloteca scikit-learn, pois é mais rápida que o GridSearch. Este método permite a otimização dos hiperparâmetros devido à utilização do k-fold cross validation. Assim sendo, para o modelo Random Forest testámos no *HalvingGridSearchCV()* os seguintes parâmetros:

- *n_estimators:* [100,200,400,500]
- *max_depth:* [4,6,8,None]
- *min_samples_split:* [2, 5, 8]
- *min_samples_leaf:* [1, 2,10]
- *bootstrap:* [False,True]
- *max_samples:* [500,1000,5000,9000,None]

Para o modelo Regressão Logística utilizámos os seguintes parâmetros:

- *penalty:* ['l1','l2','elasticnet',None]
- *C:* [0.1,1,5]
- *solver:* ["lbfgs", "liblinear","sag", "saga"]

Após a execução destes parâmetros, para o Random Forest os melhores parâmetros obtidos foram : *bootstrap*= True, *max_depth*= None, *max_samples*= None, *min_samples_leaf*= 10, *min_samples_split*= 5, *n_estimators*= 400 e para

	Random Forest	Regressão Logística
Recall	0.889	0.890
Precision	0.258	0.191
Accuracy	0.621	0.703

Table 1: Métricas nos Modelos

a Regressão Logística, C= 1, *penalty*= 'l1', *solver*: 'liblinear' Seguidamente, treinámos ambos os modelos com os melhores parâmetros obtidos referidos anteriormente. De seguida, testámos os modelos com o conjunto de teste e aplicámos as métricas: recall, precision e accuracy, para verificar qual foi o modelo que obteve melhores resultados.

Posto isto, a tabela 1 contém as métricas obtidas por ambos os modelos. Após uma breve análise da tabela, verificamos que o modelo Regressão logística contém uma accuracy maior que o Random Forest. No entanto, a regressão logística tem um recall maior que o Random Forest mas tem uma precision menor que o Random Forest. Neste caso, iremos ter mais em consideração a recall, dado que esta métrica considera os falsos negativos, isto é, positivos que são classificados como negativos. Para este caso, nós queremos saber o número máximo de clientes que são capazes de aderir à campanha, para se gastar mais recursos do banco para convencer o cliente aderir, e não desperdiçar recursos em clientes que à priori não irão aderir. Assim sendo, iremos aplicar o modelo Random Forest. Dado isto, tento um modelo preditivo com a possibilidade de input de valores por parte no cliente, a ideia inicial seria aplicar este modelo no tableau, no entanto não foi possível. Apesar disso, temos um ficheiro em python, com menu com a possibilidade de inputs para obter a decisão do modelo, isto é, se a pessoa tem probabilidade de aderir, será útil para o cliente.

8 Análise de resultados

No que toca a análise de dados foram obtidos várias conclusões. Apesar de já algumas conclusões já terem sido retiradas e expostas anteriormente, algumas ainda surgiram. Nomeadamente, uma das conclusões que só conseguimos visualizar no tableau, foi que de toda a população estudada apenas 3 pessoas efetivamente tinham um crédito, isso pode demonstrar que a equipa de marketing, tentou abordar pessoas para a campanha que não tinham algum tipo de crédito. Outra conclusão retirada é que pessoas entre os 20-30 anos estão mais propensas a aceitar a campanha. Mais algumas das conclusões retiradas é que independentemente do estado matrimonial da pessoa ou seja (divorciado, casado ou solteiro) o grupo dos 30 anos continua a ser o que mais aceita a campanha. Referindo agora a taxas podemos ver que:

- Employment Variation Rate varia entre -3.4 e 1.4;
- Cons price index não tem grande variação encontrando-se no intervalo de 92,2 a 94,767;

	Person by detail	Taxes	Professional Career	Painel
Elapsed Time	0.01764s	0.015s	0.01s	0.02s
Start Time	178.7s	360.288s	1603s	1466s

Table 2: Tempos do primeiro carregamento de cada gráfico

- Cons conf index varia entre -50,80 a -26,90;
- Euribor varia entre os valores 0,60 a 5.

Uma outra conclusão retirada, é que da população analisada se não tiver algum tipo de crédito esta tem cerca de mais 321 % de aceitar do que se tivesse algum tipo de dependência. Ao reunir esta informação, vai ser mais fácil ajudar na decisão por parte da equipa responsável, a decidir a quem abordar. Variando os parâmetros no tableau consegue-se entender como cada um diverge na decisão, permitindo assim visualizar as suas diferenças. No entanto se a equipa não tiver acesso à ferramenta, conclui-se que se a abordagem for destinada ao grupo etário de 20-30 e que independentemente do seu estado conjugal obterá bons resultados, isto também se não tiver qualquer tipo de crédito, empréstimo. E no que toca as taxas, se o mês revelar ter pouca variação desses valores, existe mais suscetibilidade a aceitarem.

8.0.1 Análise de Tempos

Quando abrimos o tableau, o first load do projeto foi cerca de 0.40s. Seguidamente a tabela 2 contém os tempos do primeiro carregamento de cada gráfico individualmente e do painel com o conjunto de todos gráficos. Seguidamente, os tempos em cada gráfico quando alteramos valores nos filtros, estão presentes nos gráficos 11, 12, 13 e por fim 14. Como é de esperar, como se trata de um trabalho de big data, em que é necessário ir buscar os dados uma base de dados, é natural seja mais demorado a renderizar os dados, nomeadamente a executar queries, quando se altera parâmetros dos filtros no tableau.

References

[1] Bank Marketing Campaigns Dataset - `https://www.kaggle.com/datasets/volodymyrgavrysh/bank-marketing-campaigns-dataset`

9 Apêndice

Por questões de conveniênica, seguidamente apresentamos o link do tableau público: `https://public.tableau.com/app/profile/filipa2385/viz/BankMarketingCampaign_16710474338200/Painel1?publish=yes`

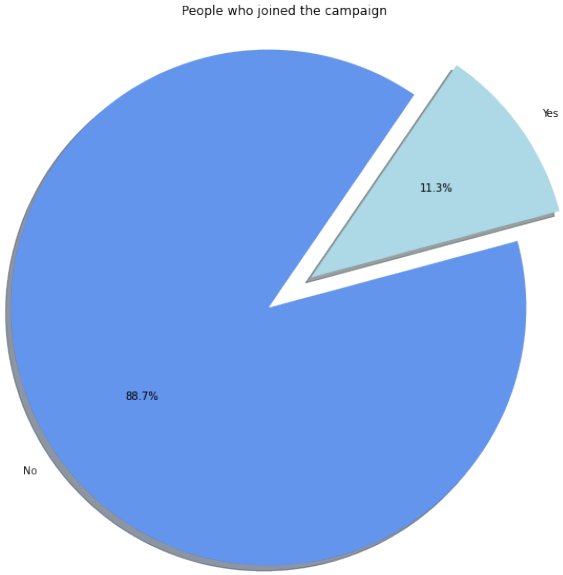

Figure 1: Percentagem de adesão

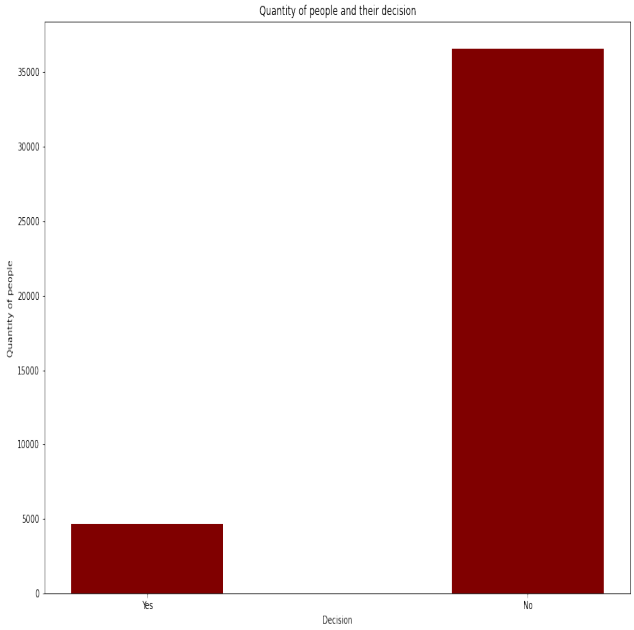

Figure 2: Quantidade de pessoas aderente

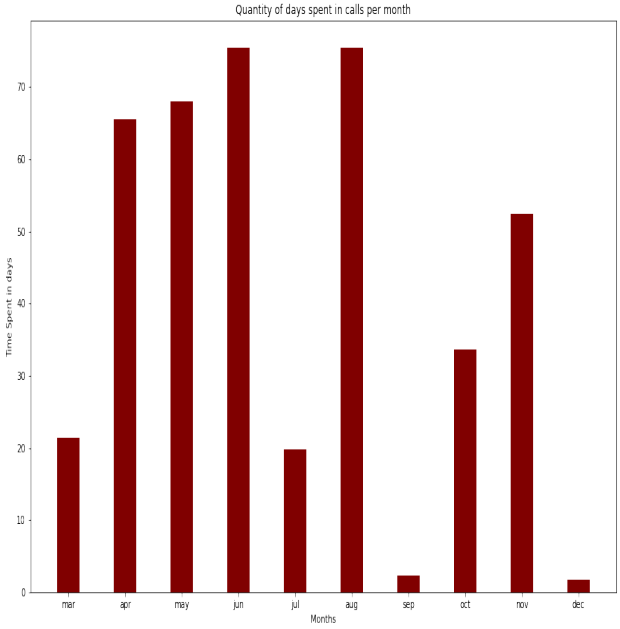

Figure 3: Tempo gasto durante o ano

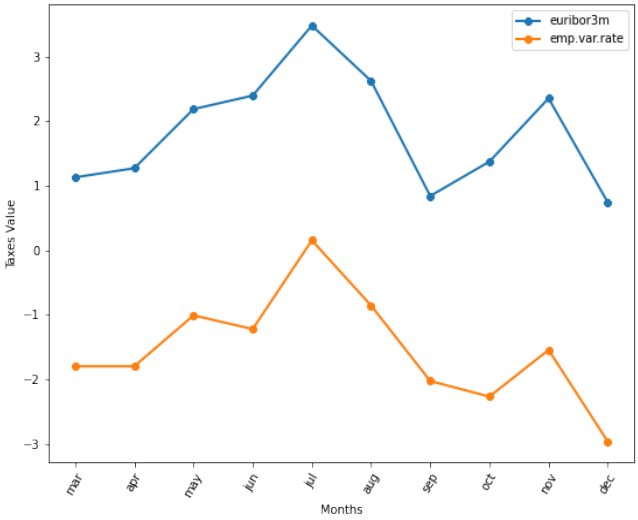

Figure 4: Taxas por mês

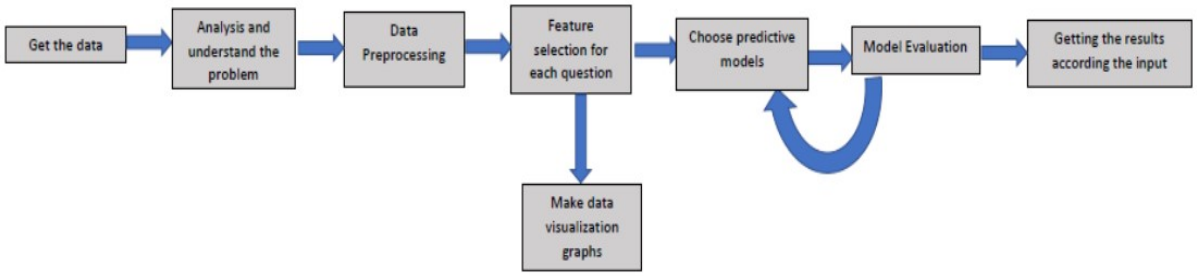

Figure 5: Pipeline de solução

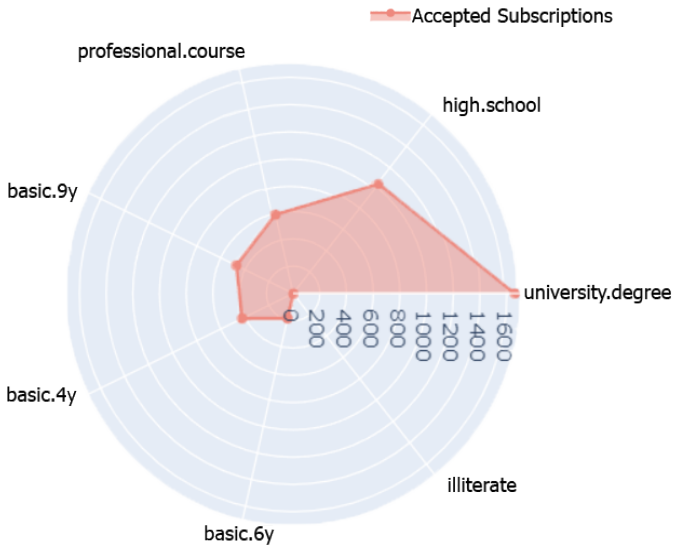

Figure 6: Educação

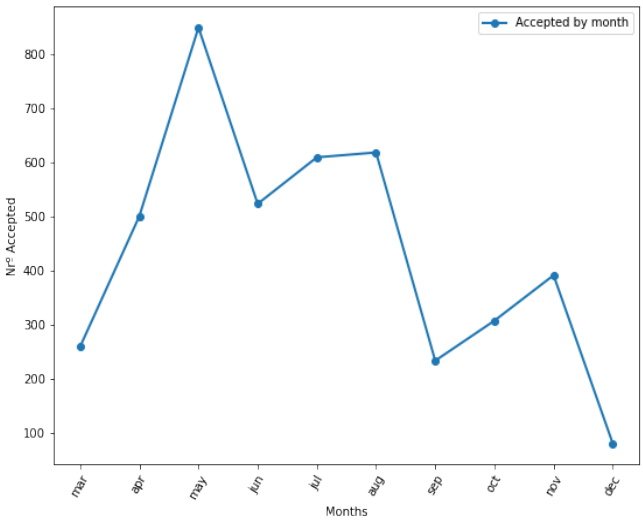

Figure 7: Adesões por mês

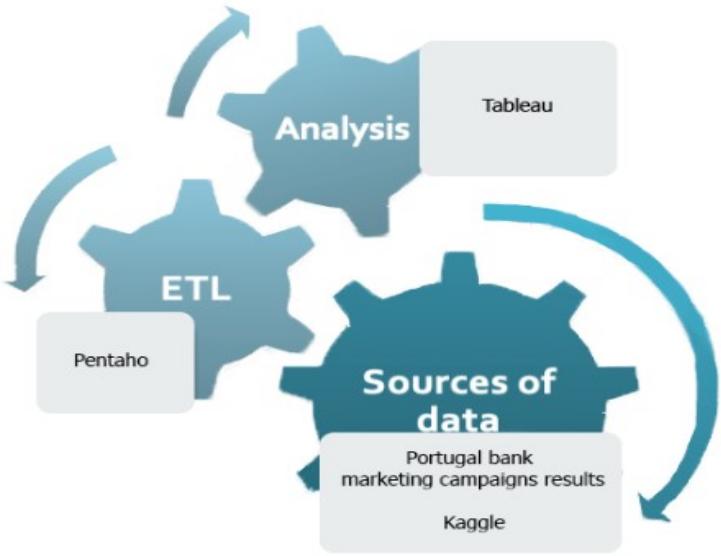

Figure 8: Workflow

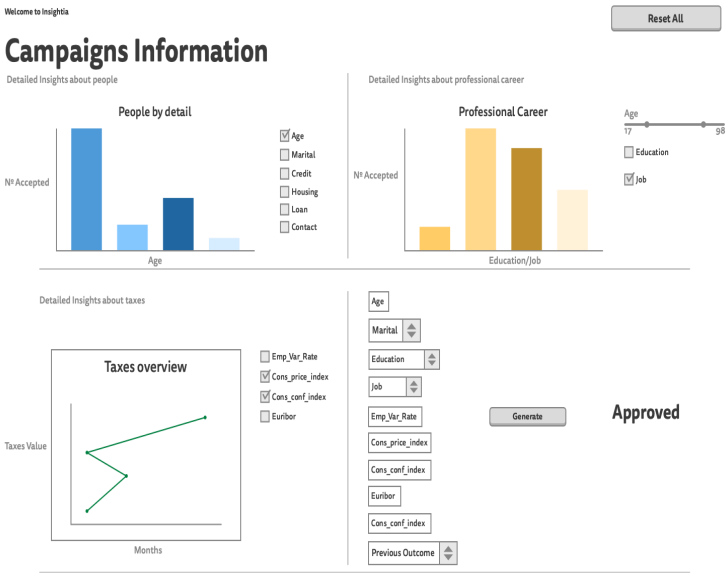

Figure 9: Mockup

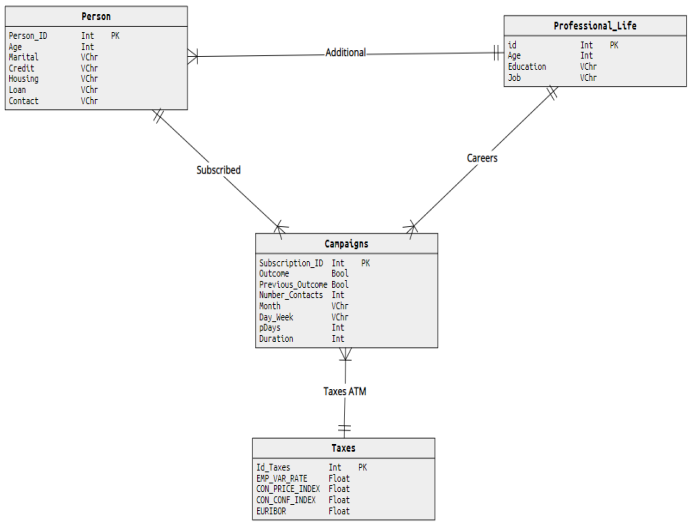

Figure 10: Star Scheme

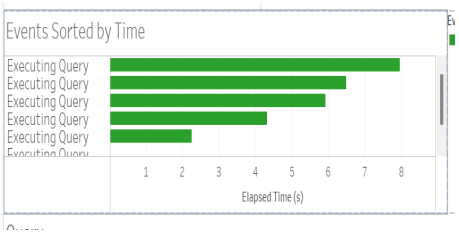

Figure 11: Gráfico do tempo de renderização das queries das Persons by detail

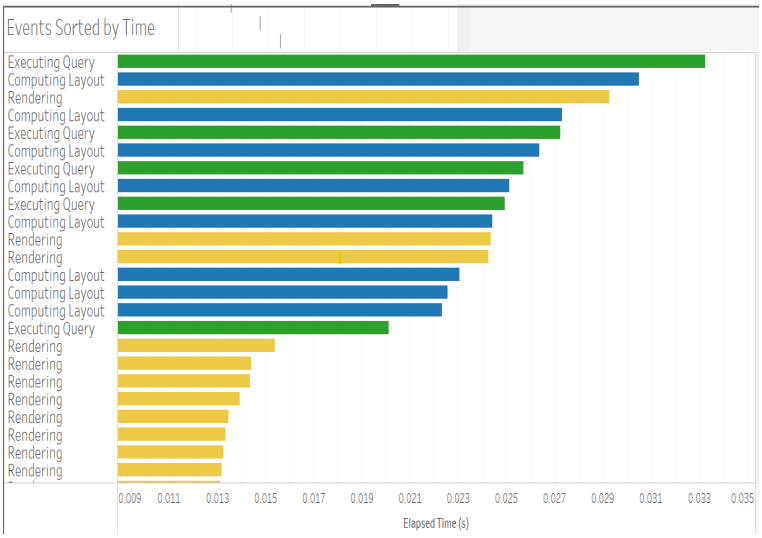

Figure 12: Gráfico do tempo de renderização das queries das Taxes

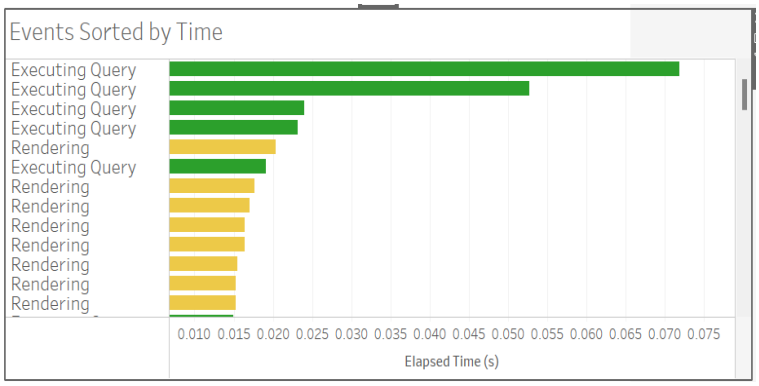

Figure 13: Gráfico do tempo de renderização das queries do Professional Career

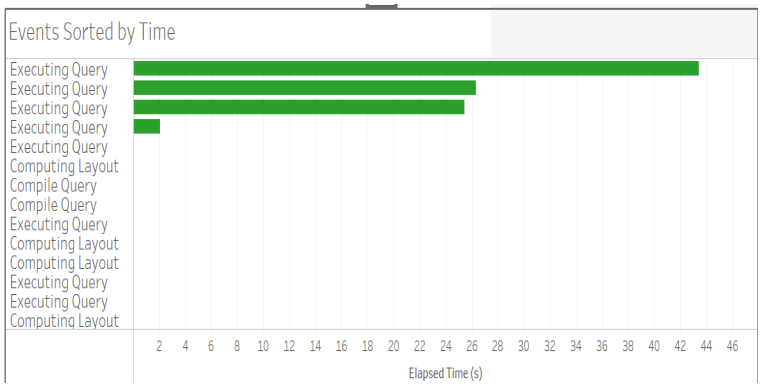

Figure 14: Gráfico do tempo de renderização das queries no overview