**Data Challenge**

The purpose of this challenge is to let you demonstrate the way you think and work. You shouldn’t spend more than 8 hours to complete the exercise.

The [dataset](https://github.com/totvslabs/datachallenge/raw/master/challenge.zip" \t "_blank) we are providing contains the orders made by customers in one of our applications. Here’s the description of each column:

* customer\_code: unique id of a customer;
* branch\_id: the branch id where this order was made;
* sales\_channel: the sales channel this order was made;
* seller\_code: seller that made this order;
* register\_date: date of the order;
* total\_price: total price of the order (sum of all items);
* order\_id: id of this order;
* quantity: quantity of items, given by item\_code, were bought;
* item\_total\_price: total price of items, i.e., quantity\* price;
* unit\_price: unit price of this item;
* group\_code: which group this customer belongs;
* segment\_code: segment this client belongs;
* is\_churn: if this client is set as a churn.

**Question 1 (10 Points)**

List as many use cases for the dataset as possible.

De forma mais ampla e generalizada, a base aparenta fornecer dados suficientes para uma análise de comportamento de consumo, possibilitando diversos casos nesta área como:

* Previsão de perda de clientes.
* Comportamento de consumo de clientes.
* Segmentação de perfis de consumo.
* Probabilidade de compra/recompra para direcionamento de campanhas.
* Diagnóstico de comportamento de venda e análise preditiva por canal e agente.
* Análise temporal e preditiva de vendas/faturamento.
* Análise de vendas fraudulentas.

**Question 2 (10 Points)**

Pick one of the use cases you listed in question 1 and describe how building a statistical model based on the dataset could best be used to improve the business this data comes from.

Como todo bom problema de ciência de dados, uma análise exploratória é essencial para entender os dados, avaliar sua saúde, possibilidade de uso e obter a clareza com os especialistas desta informação, e assim ter alguns primeiros cenários de necessidade do que é possível utilizar e para qual finalidade. Isso normalmente parte de uma ou algumas perguntas que se deseja responder, e neste caso, a pergunta que elenco para trabalho seria: **Há possibilidade de estimar a perda de um cliente?**

Com esta análise, também já pode possibilitar alguns insights que talvez ainda não tenham sido explorados, sem antes precisar investir em algoritmos, tempo, custo, para respostas que possam existir e já trazer algum tipo valor agregado. Em seguida, após este primeiro olhar citado, dar então um próximo passo em modelagem e aprendizado de máquina, com realização das seguintes etapas:

1. Planejamento inicial buscando uma pré-definição de caminhos e algoritmo, com o intuito de todos os envolvidos entender, conhecer o dado e alinhar como entrega possibilidades de caminho para experimentação. Este planejamento poderia ser uma sprint como meta “Definição de Experimento”.
2. Execução de análise exploratória:
   1. De cada variável individualmente (univariada).
   2. Dos pares de variáveis (bivariada) entendo suas dispersões e correlações.
3. Apresentação e discussão com especialistas de negócios para entender possíveis cenários estranhos nos dados, remoção de dúvidas e entendimento/alinhamento comum do significado de cada variável.
4. Análise multivariada e inferências, buscando visualizar suas multicolinearidades, existências de heterocedasticidades, comportamentos de erros e significâncias.
5. Indicar possibilidades de hipóteses e perguntas que possam ser utilizadas como experimento, e indicação de algoritmos para experimentação.
6. Planejamento de uma nova (ou mais de uma) sprint, com foco em execução dos algoritmos e apresentação de seus resultados de treino e teste.
   1. Seleção e definição das variáveis úteis.
   2. Padronização.
   3. Estruturar métricas.
   4. Aplicação dos algoritmos.
   5. Comparar seus resultados.
7. Apresentação e entrega dos resultados, com indicação de melhores soluções visualizadas e definição de um plano para validação.
8. Execução do processo de desenvolvimento (se necessário), implantação e validação.

Como proposta de experimentação neste challenge, minha sugestão seria treinamento de uma SVM, comparando com Regressão Logística e Regression Tree.

**Question 3 (20 Points)**

Implement the model you described in question 2, preferably in Python. The code has to retrieve the data, train and test a statistical model, and report relevant performance criteria. Ideally, we should be able to replicate your analysis from your submitted source-code, so please explicit the versions of the tools and packages you are using.

Disponível no arquivo 1\_SVM.py

Disponível no arquivo 2\_RegLogistica.py

Disponível no arquivo 3\_DecisionTreeRegression.py

**Question 4 (60 Points)**

1. Explain each and every of your design choices, you can use jupyter notebooks. (e.g., preprocessing, model selection, hyper parameters, evaluation criteria). Compare and contrast your choices with alternative methodologies.

Visando atingir a meta de entrega do desafio prototipado, como pré-processamento o único trabalho realizado foi a eliminação de dados nulos. A análise exploratória foi realizada, e decido para uma primeira execução não eliminar nenhuma feature ou tentar realizar alguma redução de dimensionalidade.

Em relação à escolha do modelo, houve o desejo de experimentar três cenários com usando Suppor Vector Machine, Regression + Decision Tree e Regressão Logística. Por uma simples questão de tempo de treinamento e computador disponível, decidido implantar primeiro a Regressão Logística, e se o tempo permitisse executar os demais. A SVM foi a única que não foi concluída, não por questão técnica de modelagem e codificação, mas por questão de disponibilidade atual para realização e o tempo de treinamento.

Para questão de extratificação (holdout / cross-validation), foi manual, separando 80-20 na ordem do dado como recebido, sem distribuição em folds ou realização de bootstrapping, também para ao máximo focar na execução e métricas dos algoritmos, visto que a SVM poderia não ser facilmente treinada com os dados.

Sobre avaliação dos modelos, foram selecionadas as métricas MAE, MSE e EVS, para o experimento com SVM, e para todos também seria avaliado Acurácia, inicialmente sem um ponto de corte nos scores.

Para a regressão logística, chegou-se nos seguintes valores, e o objetivo seria experimentar os outros modelos citados para comparação:

* MAE = 0,079
* MSE = 0,079
* EVS = 0,495
* Accuracy = 0,92

Para a decision regression tree, chegou-se nos seguintes valores:

* MAE = 0,00079
* MSE = 0,00079
* EVS = 0,995
* Accuracy = 0,9992

1. Describe how you would improve the model in Question 3 if you had more time.

Gostaria de ter avaliado o terceiro modelo com SVM, obtendo suas métricas e realizando as devidas comparações para sugerir um melhor, e após este cenário, com maior tempo de dedicação.

Como próximos passos e melhoria eu gostaria de realizar os seguintes pontos:

1. Estudar melhor as variáveis com especialistas, a fim de garantir seus usos e suas efetividades antes de modelar.
2. Aplicar holdout com uso de folds, a afim de não ter um teste viciado somente nos últimos 20% da base.
3. Treinar outras vezes a decision e regression tree, pois sem realizar nenhuma poda na árvore, fiquei com a preocupação do modelo estar com overfitting dados seus resultados nas métricas, e assim acabar sendo muito especialista.
4. Realizar uma quarta tentativa de modelo baseado em uma rede neural recorrente, para também comparar seus resultados, caso algum dos três modelados não se tornasse suficientemente bom para o negócio.

You can find the dataset for this challenge in the following url:  
<https://github.com/totvslabs/datachallenge/raw/master/challenge.zip>