***Thanks for your interest in joining the Data Science team at MiBolsillo!***

**Background**

At MiBolsillo, we are all about financial health (that means we optimize the way our users spend, save, borrow and plan for the future). But it’s exceptionally tough to change consumers’ behavior around their finances. As a matter of fact, one of the most common areas we see this pattern is the difficulty consumers have in budgeting and saving money.

However, there are a number of strategies to help consumers save money that are seeing success at scale nowadays driven by data. Among the most interesting approaches are **financial automation** and exploiting the concept of the **“gambler’s fallacy”** for the saver’s benefit. Both seek to offer solutions that address - and, in some cases, embrace our irrational behavior when it comes to microeconomics.

Specifically, by analyzing your spending habits, an algorithm can dynamically project the trend a given user is following and send him/her a timely behavioral intervention to curb that behavior before he/she defaults. That same algorithm could also calculate the amount the user can afford to part with without materially impacting his or her weekly cash flow needs. That “unmissed” money could transferred into a savings account for example and automatically and invisibly save money for the user.

The logical extension of this is the concept of **“**[**self driving money**](https://a16z.com/2019/10/01/google-maps-for-money/)**,”** in which consumers benefit from:

* Not only saving automatically, but also
* Being refinanced into the credit cards, loans, and mortgages with the most favorable terms,
* Offered products with characteristics match their own irrational behavior. For instance, a high-limit credit card might be offered to the consumer who considers it a source of security, but not the consumer who views it as an irresistible temptation to spend.

At MiBolsillo, we perform flow-based assessments to determine the monthly consumption trends, savings and repaying capacity of a customer to create a financial health index. Our final users are small & underbanked individuals, micro & small business owners in Latin America who are running a bootstrapped and consistently profitable business, yet they remain excluded from formal credit. Formal institutions have shied to lend to lower-middle income group because the cost-benefit analysis of lending and collections do not offset the cost of originating and recovery. There is no cost-effective measure to monitor income/solvency and ensure timely repayment which is why it's important to prove to those partners (financial institutions) we're providing them with financially-fit potential clients.

**The goal**

Understanding that automated finance is less about the ability to decrease paperwork and prices, and more about the ability to impact our users' choices and cognitive biases at scale, the goal of this challenge is **to build a theoretical categorisation engine (a predictive experiment)** to send the right behavioral interventions to our users in order for them to keep an optimal level of financial health based on their credit card consumption trends.

**The dataset**

This data set belongs to real **credit card** statements for 20+ users from a mid-size bank in Brazil. The data belongs to the period of Jan 2019 to April 2020. We have a transaction narrative column for categorisation. This is not ideal, but enough to solve the problem at hand with adequate accuracy.

Columns are of transaction data: [ Portuguese to English]

id = id

celular = phone number

safra\_abertura = branch number

cidade = city

estado = state

idade = age

sexo = gender

limite\_total = total credit card limit

limite\_disp = current available limit

data = date

valor = amount

grupo\_estabelecimento = category of expense

1. SERVIÇO = any service other than restaurants
2. FAMARCIAS = drugstores
3. M.O.T.O = online purchases
4. ARTIGOS ELETRO = electronics
5. VAREJO = convenience store
6. LOJA DE DEPART = department store
7. VESTUARIO = clothing
8. HOSP E CLINICA = hospital or clinic
9. SEM RAMO = not specified/other expenses
10. SUPERMERCADOS = supermarket
11. MAT CONSTRUCAO = construction material
12. POSTO DE GAS = gas station
13. RESTAURANTE = restaurant
14. CIA AEREAS = airlines
15. MOVEIS E DECOR = furniture & decoration
16. JOALHERIA = jewelry
17. TRANS FINANC = financial transfers
18. AGENCIA DE TUR = tourism agency
19. HOTEIS = hotels

cidade\_estabelecimento = city where the purchase was made

pais\_estabelecimento = country where the purchase was made

**Questions to answer:**

1. **What would be the suggested steps to make this Supervised Learning Model? Explain in detail when possible.** Ex: Start by creating training data, train the model, run experiments, visualize dataset output, validate the model and the substeps these phases would require **(no need of data execution).**
2. **What are some of the challenges you oversee from the dataset shared?**
3. **Do you consider a need to apply any preprocessing on the training dataset? If so, why?**
4. In order to send personalized behavioral interventions in the way of in-app push notifications to our users, we need to classify users in 3 profiles: (i) those who want to better manage their credit card expenses; (ii) those who want to avoid going over their credit limit and (iii) those who want to renegotiate their credit card debt [for which this specific data set has limited information]. How would you execute data from any one of the indicators (highlighted in yellow on [slide 4](https://docs.google.com/presentation/d/13uBlgjAgziVfgM1ImqxZbElezfJv_NH2I551RcNQX58/edit?usp=sharing)) for 1 of the profiles mentioned and across one of the 4 time frames referred on that same slide to analyze user consumption trends **(needs execution, provide visual aids where possible).**
5. **If your were the client "Banco PAN" how would you envision a dashboard where all this data is collected, which 3 features from the users' spending behavior do you think they'd be interested in taking a look at and why?** **(needs execution, provide visual aids where possible).**

**To help out!**

\*Dates in portuguese follow the pattern dd/mm/yyyy

\*\*Examples of behavioral intervention

(+) *Bravo! this month (May) you reduced in two-thirds your restaurant expenses, cooking for most of your meals is paying off!*

(-) I just saw you s*pent more than your average on online shopping this month, click here to give you a hand on your budget.*