Final Project 2.

Project Problem and Hypothesis

- Over the last couple years, wellness and mindfulness services are becoming more important and recognizable. More and more people are interested in improving their wellness using mindfulness and yoga exercises. Therefore, many of them are starting to use wellness web programs. For wellness companies it's important to understand a targeted user group and user engagement. In my project I will use Whil user data.

- *Problem***.** I would like to predict how likely Whil users who sign up for the trial period are to become paid customers based on the data provided when signing up and user engagement.
- *Model.* I am predicting how likely Whil users who signed up for trial period will become paid customers based on the data provided when signing up and user engagement. This is classification problem.
- *Impact.* The company will start to target users who are more likely to become paid customers.
- I believe user experience and activity variables will have the biggest impact on the model.
- *My hypotheses*: *experienced and active mindfulness users are more likely to convert into paid customers*. People tend more to follow mindfulness programs. They are easier than yoga and show good results, but usually only work for experienced users

Datasets. *Description of data set available, at the field level. (see table)*

Variable	***Type***	***Description***
gender	Categorical (str)	Gender (f = female, m = male)
experience_level	Categorical (str)	User experience with yoga and mindfulness: none, low, medium, high
estimate_birth_day	Continuous (datetime)	Users only provide the age, estimate_birth_day gets automatically calculated at the day of sign up.
focus	Categorical (str)	User's selection of practice program: 'yoga' or 'mindfulness'
sessions_completed	Continuous (int)	Amount of sessions completed during the trial period
series_completed	Continuous (int)	Amount of series completed during the trial period
subscribed	Categorical (int)	User subscription indicator (1 – subscribed, 0 – subscription cancelled, empty cell – not subscribed)
sec	Continuous (int)	Amount of seconds user spent on training during trial period

If I find out that there's not enough data for the model, I can collect more data about user activity: started sessions, favorite sessions, watched series, selected programs, etc.

Domain knowledge

- *Experience in the area.* I work for the company as a QA engineer and have a great knowledge about the product, features, issues, db and platform architecture. I collected and analyzed user data a few times, and created eligibility reports for company partners.
- I believe my knowledge of db and experience analyzing user data should help.
- *Other researches.* There were no similar researches performed by the company. In order to target users, there were a few marketing surveys made that showed different results over time.

Project Concerns

- *Concerns.* The data selected for the project only reflects a few parameters of users engagement. There are more data that can be used, so I hope to figure out during the EDA what is useful and what's not for building a model.
- *Assumptions and caveats to the problem.*
  - There are data about users moods, factors to train, location and user environment that's being collected through different analytics tools. Unfortunately, from my experience, it's not accurate and I won't be using it for this project.
- Risks.
  - Cost of my model being wrong can result in targeting wrong users and less revenue. Taking into consideration the fact that we don't have a model yet, the benefit should lead to revenue growth.
  - There's a possibility that data provided by users is not very accurate, we cannot guarantee users provide accurate data about themselves. Tracking and user activity data is collected by the company and correct.

Outcomes

- I hope to build a model that I can use to predict how likely a trial user purchases a subscription. I should be able to classify the users using binary values 0 (not subscribed) and 1(subscribed).
- I expect the model to be not very complicated. I am planning to break the dataset into test and train data to get a better model and avoid overfitting.
- I would consider the project successful if at least about 80% of users could be classified correctly.
- If the project is a bust, I will try to figure out if I missed anything or if there are more data that can help. Maybe redo the project for a different problem.