Bootcamp Capstone 项目说明

Capstone项目都有哪些Track？

- **[Track 1] Kaggle - breast cancer**
    - 数据内容：
        - Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Label: Diagnosis (M = malignant, B = benign).
        - a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) perimeter d) area e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" - 1)
    - 项目目标：
        - Utilize various tools to clean and explore data.
        - Apply different models in this binary classification problems and improve performance by designed metrics.

- **[Track 2] Lending club数据predict default rate**
    - 项目描述：
        - 基于Lending Club数据，创建模型以预测贷款的违约率 (default rate)，从而提高Lending Club的风险把控。与第一个月的项目相比，model中可加入更多feature，如：loan past payment 的数据, 强化独立完成完整e2e的项目经验。另外，同学们可以选择使用在整个bootcamp中学到的model 来进行预测，并且比较不同模型的performance。

- **[Track 3] 神策数据**
    - 公司介绍：国内领先的用户行为分析产品
    - 数据内容: 该公司官网访问约一周的数据，包括用户访问时产生的点击按钮、申请账号、提交验证码、观看视频、离开页面等行为记录。详细的日志描述并配以官方技术文档以及API手册说明。
    - 项目目标:
        - Clean dirty log data and transform it for analytics.
        - Exploratory data analysis, e.g. find user activity levels for different events, and user interaction with web components.
        - Find the conversion rate of users, identify key factors that bottleneck the conversion rate.
        - Propose any hypothesis and set up experiments for testing.

        - Build machine learning models to predict user behaviors, including but not limited to signup, churn, etc.
        - Discover interesting insights in the dataset and suggest how to improve the user signup rate.

- **[Track 4] 某知名音乐播放盒数据挖掘**
    - 公司介绍：某知名音乐播放平台
    - 数据内容：刚出炉的新鲜数据：）每日260K新增用户的3 million+的歌曲播放记录（不断更新中）， 包括用户uid, 用户os, 播放歌曲的rid, 歌曲的类型, 歌曲名称, 歌手名称, 播歌时长, 歌曲时长等信息。
    - 项目目标：
        - 目标一：Churn Prediction
            - Validate dataset, identify missing values and find inconsistencies in the dataset.
            - Perform data cleaning and transformation, feature engineering
            - Exploratory data analysis, e.g. find most popular songs, most active users
            - Build user churn prediction model based on user behavior, implement full cycle of prediction modeling from population selection and sampling, label definition, feature exaction and engineering, model selection, performance evaluation.
            -
        - 目标二：Recommendation
            - Validate dataset, identify missing values and find inconsistencies in the dataset.
            - Perform data cleaning and transformation, and construct utility matrix from user behavior data
            - Define implicit ratings from user behavior data
            - Build music recommendation system based on user listening history, including: popularity-based recommender, item-item based recommender, matrix factorization-based recommender.

Capstone评价标准是什么？

- 项目完整程度：同学们完成项目的流程的完整度（data processing, explorotary，建立模型， performance等）
- 项目复杂程度：同学们完成的项目需要利用起课程中学到的技术和模型
- 项目商业价值：同学们在项目完成以后对business创造多少价值
- 项目新颖程度：同学们的项目能解决多少未被解决的问题

项目代码要放在哪里？

项目的代码统一放在同学们自己的Github Repo上。