**Personal Intro**

I'm a data scientist with an unyielding enthusiasm for interdisciplinary research. With exceptional sensitivity to numbers, I have the instinct to record the world around me onto spreadsheets, trying to summarize everything I see with manipulatable data. But such fondness for numbers won't stop me from being an absolute music lover, and an oddly sentimental naked-eye stargazer.

Currently an undergraduate at UC Berkeley, I am actively looking for internship opportunities in data science industries. I am searching for somewhere that I can emerge myself into the realm of data and explore the infinite possibilities that numbers can give us. I am especially enthusiastic to work as a team because I always cherish the lesson I could learn when two different ideas collide.

**个人介绍**

我是一名对于数据探索以及机器学习有着持续性热情的数据分析师。 由于对数字异常敏感，我会本能地将我周围的世界记录在表格上，并尝试用可操作的数据工具描绘我所看到的一切。

目前作为一位加州大学伯克利分校统计学和计算机科学在读大三学生，我正在积极寻找专注于机器学习的数据科学行业的实习机会。通过我过往在机器学习方向的经历和学术研究，我希望让自己能更深入数据领域，探索到更多数字带来的无限可能。

对数字的喜爱及敏感，不仅促使着我成为一位优秀的数据掌控者，同时，也陪伴着我成为一个绝对的音乐爱好者和一个多愁善感的裸眼观星者。

**HOYO Lab**

Co-Founder

March 2021 – Aug 2021

Full Description:

Two self-initiated, interest-driven data-centered projects on the RPG game Genshin Impact (by MiHoYo), released in September 2020. Under the umbrella name “HOYO Lab”, our analysis was focused on two aspects of the game.

In the first subproject “Damage Calculator”, I used linear regression with stochastic regressors and feature mapping to model the game’s damage mechanism. I aim to estimate the coefficients that was hidden from the players and predict the final damage dealt. I collected data through hundreds of in-game simulations, altering the value of our features through modifying the builds of our in-game characters. With 16 features in total, I constructed many naive feature maps on our variables, then performed cross validation to select the feature map with the best predictive power. I have achieved an prediction error very close to 0. Our result model was published on the website in the form of a web app. The web app was built using the shinyapp package of programming language R. The final model was utilized to help me publish tutorial articles on gamers platform which received more than 400,000 views and 50,000 likes.

In the second subproject “Celestial Database”, I lead a group of 3 with the aim to build a game character power rank based on the character’s usage rate in the game’s most challenging combat called the “spiral abyss”. The game has an associate mobile app that showcased user’s statistics regarding the spiral abyss. By using python requests, our group crawled through thousands of user’s profiles and collected data on character usage. To overcome the mobile app’s user profile protection barrier, we used virtual headers in our requests to fetch the information. We decoded the JSON files to extract the data of interest. The data was stored in a MySQL database that was updated every 3 weeks. The power rank was showcased on the website.

Resume Version:

* Self-initiated, interest-driven, data-centered project on the RPG game Genshin Impact:
* Subproject I “Damage Calculator”:
  + Predicted the game’s underlying damage mechanism using linear regression with stochastic regressors and feature mapping (16 features); collected data through 500+ in-game simulations, and initialized the model with naïve feature map functions and finalized the feature map function through cross validation; achieved prediction error close to 0.
  + Displayed the final model through web app using the shinyapp package in R [1], then utilized the model and published tutorial articles that received 400,000 views and 50,000+ likes.
* Subproject II “Celestial Database”:
  + Aimed to build an in-game character power rank based on character usage in combat challenge.
  + With a group of 3, crawled 1000+ user profile in game associated mobile app using Python requests, and overcame the protection barrier using virtual headers.
  + Decoded the fetched JSON files to extract data of interest, then stored and updated data in MySQL database for every 3 weeks; showcased final power rank on website [2].

HOYO实验室

联合创办人

2021年3月 – 2021年8月

* 该项目是以兴趣驱动、数据为中心，围绕RPG 游戏 Genshin Impact的自发性项目：
* 项目一“伤害计算器”：
  + 使用带有随机回归量和特征映射（feature mapping）的线性回归（16 个特征）预测游戏的伤害计算机制；基于从游戏内模拟收集的500+行数据，用原始的特征映射函数初始化模型，并通过交叉验证最终确定特征映射函数，达到了接近于0的误差；
  + 使用了R 中的 Shinyapp 包制作了网络应用程序以展示最终模型，该模型制作的教程在游戏社区平台获得了400,000+浏览量及 50,000 +点赞量
* 项目二“天空岛数据库”：
  + 该项目旨在根据战斗挑战中角色使用率建立角色强度排行榜；
  + 带领3人小组，使用 Python requests爬虫及虚拟标题克服了应用程序保护屏障，在与游戏相关的移动应用程序中爬取了 1000 +用户个人资料；
  + 解码 JSON 文件以提取相关数据，在 MySQL 数据库中每三周存储并更新一次数据，最终在网站上展示了最终的角色强度排名

**Beijing PM2.5/PM10 Prediction**

Project Leader

Nov 2021 – Dec 2021

Full Description:

Lead and coordinated a team of 3 on a data analysis project. The aim of the project is to predict the real-time measurement of fine particles (PM2.5/PM10) in Beijing through building a linear model. The data are air pollutant measures from Beijing Municipal Environmental Monitoring Center from 2013 to 2017. The data have around 30,000 instances and 12 features. We performed a series of robust exploratory data analysis through visualizing data distributions, bivariate relations, and the effect of interaction terms. We used the Alkaline Information Criteria (AIC) with backward stepwise selection method to engineer our features, arriving with two models of different numbers of features. Then, we carefully diagnosed our selected models by checking for linear model assumptions and highly influential datapoints. Lastly, we tested our models on the validation dataset, constructed prediction intervals, and discussed the tradeoffs between the two models. The final RMSE of our selected model is 0.6. The result is presented in a report with detailed description and diagrams.

Resume Version:

* Lead and coordinated group of 3, aimed to predict the real-time measurement of PM2.5/PM10in Beijing using imported data (30,000 rows, 12 features) from Beijing Environmental Monitoring Center between 2013 and 2017;
* We performed robust EDA through visualizing univariate distribution, bivariate relation, and interaction significance; employed AIC with backward stepwise selection for feature engineering; arrived with two models that was carefully diagnosed for model assumption and outliers.
* We tested our models, which achieved final RMSE of 0.6, and constructed prediction intervals. We presented the result in a report with detailed description and diagrams.

基于空气指标对北京PM2.5/PM10进行实时预测

数据分析项目组长

2021年11月 – 2021年12月

* 该项目旨在利用北京环境监测中心2013年至2017年的数据（30,000项，12个特征）对北京的PM2.5/PM10指数进行实时预测
* 首先，通过观察单变量分布、双变量关系和分类变量的交互影响来执行深入探索数据分析（EDA），再使用向后逐步AIC信息准则进行变量选择，对变量选择后的模型进行了前提条件和异常值诊断
* 不断优化模型，最终使其均方根误差达到 0.6 ，并为其构建预测区间；最终，将模型及所得数据可视化，主导并撰写小组最终研究报告

**DataCVG Shanghai**

Intern

May 2021 – Aug 2021

Under the guidance of a tutor, I performed Extract-Transform-Load (ETL) on different data sources using the ETL tool Kettle and the query language MySQL. The objective of our ETL project is to combine data from two outdated data sources onto a single newly constructed data source. There are over 100 spreadsheets in total, and we need to design the structure of each spreadsheet in the new destination data source. During the construction process, we considered the primary keys that were needed to be included, along with necessary metadata columns. During the extract and transformation process, we discovered that many spreadsheets from the outdated data sources are incompatible with each other. To overcome the issue, I either reshaped the spreadsheets using groupings and unpivots, or modifoed the datatypes of incompatible columns. Then the spreadsheets are joined through matching primary keys conditionally, and was finally loaded onto the new data source. After loading the data, I performed a series of rigorous diagnostics to detect any faulty value, such as an error in the grouping algorithm. This ETL program helped our client (Fosun Pharma) upgrading to their newest, higher performance database system.

Resume Version:

* Performed extract-transform-load (ETL) using Kettle and MySQL; combine outdated spreadsheets from two data sources (300+ spreadsheets) onto an updated data source.
* Constructed the architecture of destination spreadsheets by handpicking 3~5 primary keys along with necessary metadata columns; reengineered incompatible columns through grouping rows, pivoting columns, and modifying column datatypes; loaded the final results through merging transformed spreadsheets by matching primary keys; debugged the final data through a series of rigorous diagnostics for any architectural failures.
* The ETL project had helped the client, Fosun Pharma, to upgrade onto the newest, higher performing database system.

上海数聚软件系统股份有限公司

实习生

2021年5月 – 2021年8月

* 采用Kettle 和 MySQL 执行数据提取-转换-加载 （ETL）技术， 与小组三人将来自两个过时数据源的300+张表格合并到最新数据源上
* 在SQL中筛选并审定多个主键以及必要的元数据列来构建目标表格的架构，独立完成数据清洗，并通过数据分组、unpivot、及修改数据类型来重新调整不兼容列，对最终合并后的表格进行诊断检测
* 该ETL 项目已帮助客户复星医药升级到最新的、性能更高的数据库系统

**Sky&X Aerospace**

Dec 2019 – Jan 2021

Full Description: