**Venderly Hackathon Case Study Submission**

**Title:** *“Revenue Forecasting for Student Ventures: Predicting Business Success Across Diverse Communities”*

**Venderly**  
Venderly helps students build real online businesses through a classroom-ready platform integrated with Stripe and Square. Students sell physical or digital products, learn entrepreneurship, and generate real-world revenue, all backed by actual payment processing data.

**The Challenge**

Can we predict the potential total revenue a student-run business will generate based on past transaction data, the type of business it operates, and the surrounding community profile? If so, we would like to make this a consumer-facing tool whereby a school could input its demographic information, select a business type to run, and receive an estimate of its annual revenue.

**Available Data**

We have provided structured datasets in the Google Drive folder below.

<https://drive.google.com/drive/folders/141OSwbcqW14JiJ-DK0bQH2Rcp7NnJ-mR?usp=sharing>

**Sales Transactions**

* Stripe/Square records: spreadsheets will be in separate folders categorized by each provider. - [Data folders linked here](https://drive.google.com/drive/folders/1-CbsC4jBjz6eX0Qg6SJK-WUE3wl9DoI8?usp=drive_link)
* You will also find public school information data, as well as the key columns of interest for each processing provider within this document: [***School - Location ID Mapping***](https://docs.google.com/spreadsheets/d/16DdJDSeIYas64HK8wXUfDmk2xKkydTuA/edit?usp=sharing&ouid=112798907885815522740&rtpof=true&sd=true), separated via sheets

**Store Metadata**

* Product categories / business type (e.g. apparel, digital downloads, event services); (not all sheets contain available information)
  + You define aggregation/bucketing of product category data,
* Launch timing and lifespan of store
  + Note: each spreadsheet contains transaction data for a given store. Stores that have been operating multiple years will have multiple spreadsheets. For Square, the *location* primary key will be the same across all files. Some data files are monthly, others are yearly. For Stripe, all transactions took place between July 1, 2024 and June 30, 2025.
* Product pricing and count

Data we do not currently have – you provide (ideally via AI/ML)

**School & Community Profiles**

* Location (state, city, zip)
* School type (middle/high)
  + We will define for you via the [***School - Location ID Mapping document***](https://docs.google.com/spreadsheets/d/16DdJDSeIYas64HK8wXUfDmk2xKkydTuA/edit?usp=sharing&ouid=112798907885815522740&rtpof=true&sd=true)
* Student population size
* Community size indicator (urban, suburban, rural)
* Regional economic markers (e.g. free/reduced lunch %, per capita income, census data if needed)

**Goals and Expected Outcomes**

Across hundreds of schools, students launch businesses through Venderly with widely varying levels of success. By analyzing payment data alongside school and community characteristics, we aim to build a machine learning model that can:

* Forecast annual revenue potential for a store
* Describe how business type and community context influence success
* Predict the amount of revenue various business models would make based on business type and community context
  + Stretch: output a web-based calculator where potential customers can input their community data, select a business type, and see the predicted annual revenue for their store type in confidence bands.

Storecasters (hackers) are free to choose forecasting methodology.

**Real-World Value**

This project would help:

* **Schools**: Guide schools toward business types with higher success potential based on their community profile
* **Students**: Reduce trial-and-error when launching stores
* **Venderly**: Offer smarter, personalized onboarding tools to each school

The result could be a consumer-facing tool whereby a website visitor could input their demographic information, select a business type to run, and receive an estimate of the annual revenue they could make.

**Suggested AI/ML Approaches**

* **Time series analysis** for sales trajectories
* **Clustering** to group school/store profiles
* **Regression modeling** (linear regression, decision trees, random forest, XGBoost) to predict revenue
* **Recommendation system** (hybrid of content-based + performance-based ranking