Project Requirements Document, v2

Team Name: Excelsior

Project Name: Sibyl

Sponsor: Appfolio

Team Members:

      Sahil Naik (Team Lead)

      Rutvik Jha (Scribe)

      Jake Yim

      Skylar Winig

      Robin Wang

<u>Introduction</u>

Background

Property managers collect an influx of accounting data across their portfolio of properties. Currently, millions of data points regarding their assets, liabilities, expenses, income, and equity are hidden deep in databases with no direct insights or gains being made. Generating insights or projections from this vast accounting data requires a lot of overhead and human labor from the property manager's side. As a result, property managers often face uncertainties with their assets as they are restricted in projecting their income and are forced to look at confusing and incomplete personal spreadsheets of their expenditures when trying to get a better understanding of their properties. When building a portfolio, it is crucial to have a comprehensive understanding of financial assets and to have the foresight to predict net gains or losses in incoming months. Sibyl aims to allow Property Managers to extract insights from their vast accounting databases and analyze them in a meaningful way without extra overhead costs, by leveraging Machine Learning and a UX focused Dashboard to visualize ML insights.

Objective

Our objective is to create a machine learning model to forecast property management expenditures and income. We intend to train our model using the vast plethora of property management data that is hidden or confusing to the managers themselves. Since our model will be pretrained on the exact same format and fields that we expect to create real forecasts on, we

will be able to produce accurate projections of the financial data. Based on our projections, property managers will gain insights into which areas of their business to focus on using our user experience focused dashboard. This is the secondary crux of our project. We plan to create a user experience centered dashboard to present our predictions in a clear and concise manner using dynamic graphs and charts. These graphs will be generated using our machine learning model's predictions and will allow property managers to quickly see their past reports and future forecasts to tackle the financial uncertainties in property management.

A stretch goal of ours is to enable other ML oriented teams at AppFolio to easily integrate the insights from their Machine Learning models onto our dashboard. We aim to accomplish this by supporting a wide range of visualization formats (continuous value insights, discrete insights, and graphing projection functions, etc) and to create a clearly documented API to integrate external model insights into our dashboard.

Current Solutions

Currently, the industry standard for visualizing data is modularization by black-boxing machine learning features and interfacing with the frontend components through API calls. We aim to adopt this design by creating a full-stack web application tasked to visualize the dashboard and python web application to generate ML insights. These two independent applications will interface with each other via API calls. This allows us to build a robust, future-proof solution, where future ML models that AppFolio might take on can easily visualize their insights on our dashboard. Or on the other side, future iterations of a dashboard can easily interface with our ML model through similar API calls.

Implementation Choices

We have chosen to develop our application using the Ruby on Rails framework. Our mentoring company AppFolio uses a Ruby-React stack and we hope that our project can eventually be incorporated into Appfolio's products. Using Rails will allow for seamless integration of our application. We have also decided to use React.js to help with the data visualization aspect of our project. React supports libraries such as Highcharts that allow for interactive and dynamic graphing of data within our web application frontend. We are choosing to use MySQL as our database system. MySQL is fast and reliable and can handle storing the accounting data given to us by our users.

Goals

1. Develop a web application to provide property managers with insights into their accounting data and predictions for their expenditures and income
2. Use machine learning and develop a useful model to make meaningful predictions on property managers accounting data
3. Display these predictions in a clear and concise manner using dynamic graphs and charts

Assumptions

Our project makes a few assumptions. We are assuming that our machine learning predictions will be accurate to provide meaningful insight to our customers or property managers. While we believe that our model is accurate based on the historical training data provided, we are assuming that other variables not provided in the data remain constant. Many things can affect the state of a property manager's business that can not be accounted for in the data provided to us. For example, this can be unforeseen damages or changing or tenets. These may be impossible to account for in our model and thus our predictions will only be helpful if the status quo remains constant.
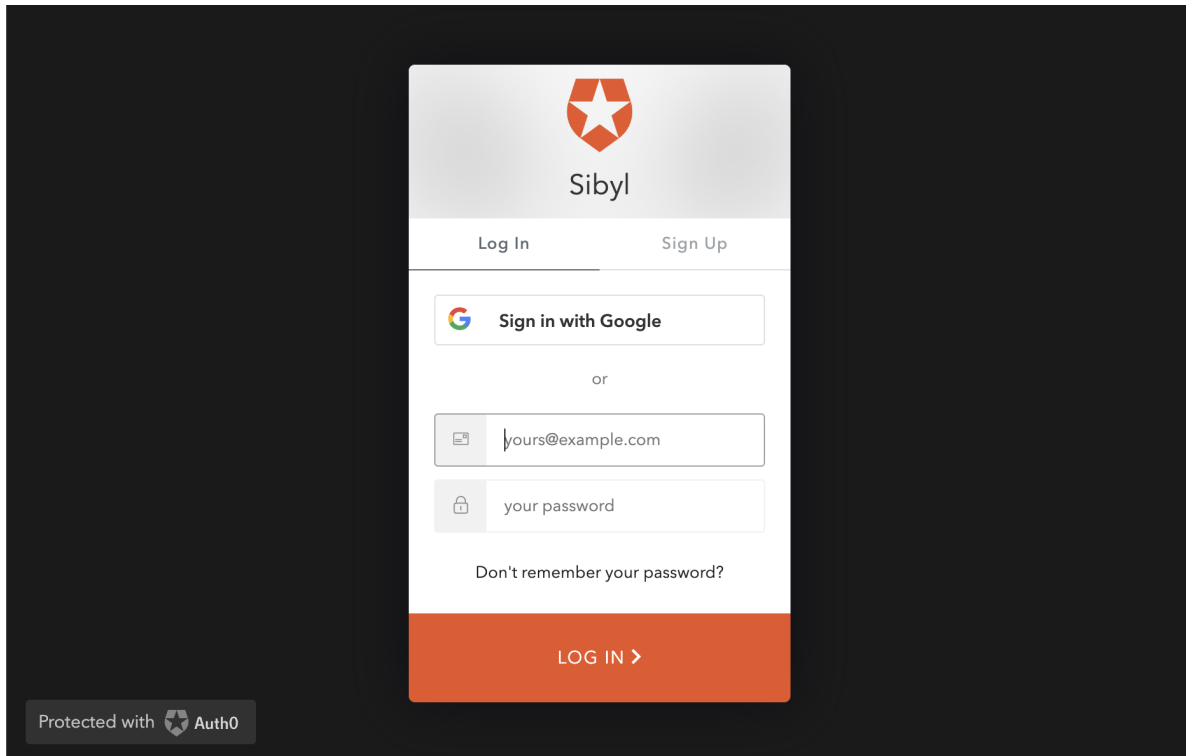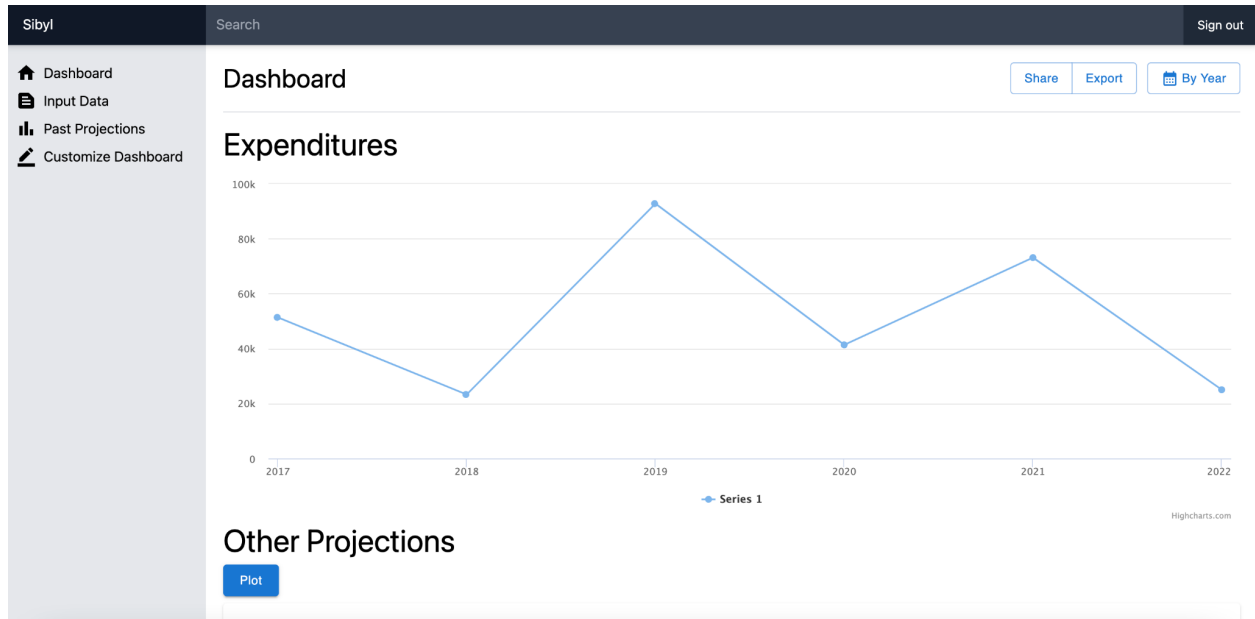
System Architecture Diagram
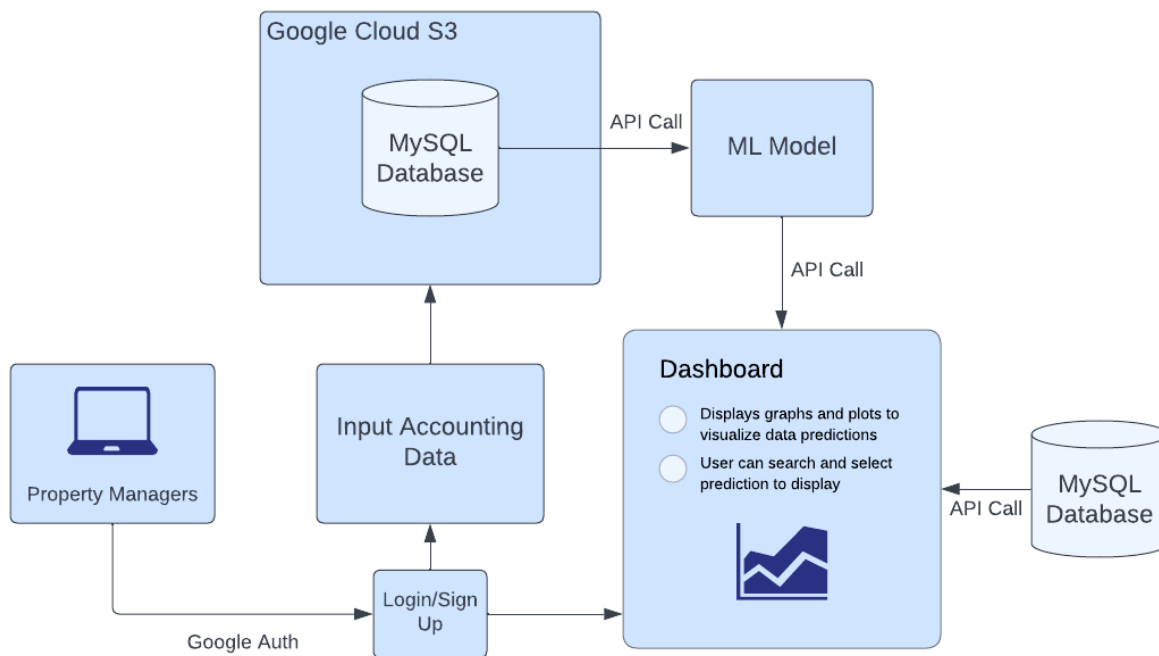
User Interaction and Design
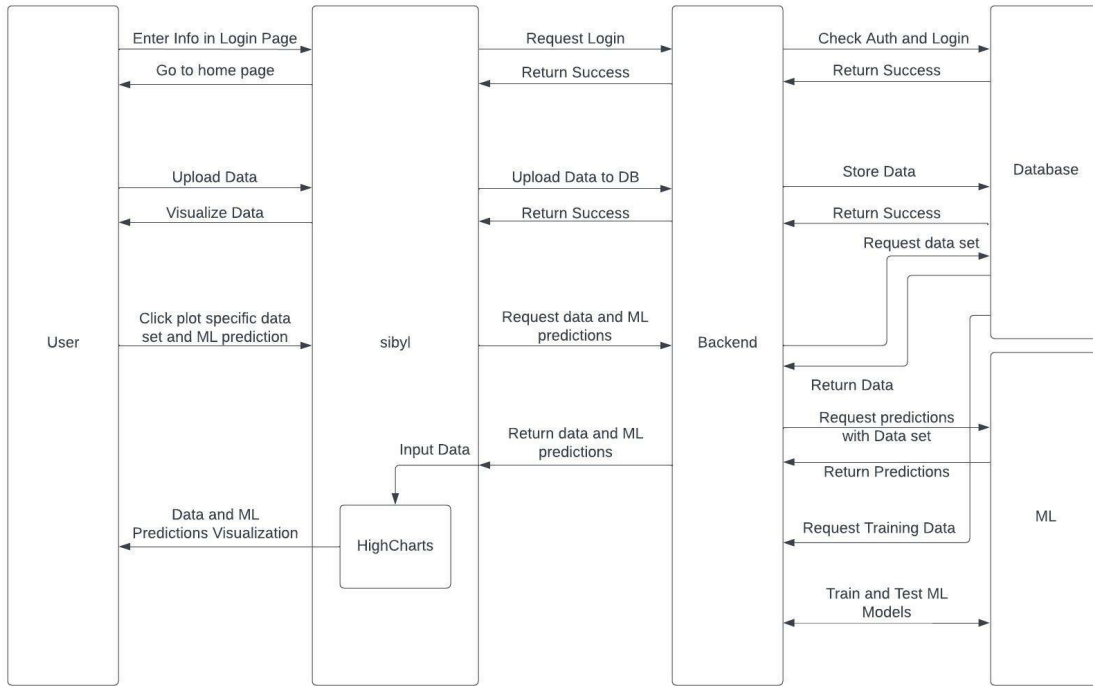
Login Page Mockup

Dashboard Mockup

Use Case/User Stories

1. As a user, I can login or sign up through Auth0 so I can access and view my accounting data and visual projections.

   https://trello.com/c/nFTM8m6h

   a. Scenario 1: The user successfully logs in with their credentials and can access projections specific to their account.

   b. Scenario 2: The user enters incorrect credentials and is presented with a login error.

   c. Scenario 3: The user can click the sign up button to create a new account

   d. Scenario 4: The user can login with newly created credentials.

2. As a user, I can recover my password if I forget it by using my email.

   https://trello.com/c/nFTM8m6h

   a. Scenario 1: The user can click the "Don't remember your password?" button and send a link to their email to reset their login password.

3. As a user, I can import my previous accounting information and data through the provided website interface

   https://trello.com/c/3YJu3izV

    a. Scenario 1: The user imports correctly formatted data (as a csv) and the data is able to be processed through our ML model.

        i. The website will display the data in a table that can be selected by the user to show predictions in graph form

    b. Scenario 2: The user imports incorrectly formatted data and the ML application is unable to process the data.

        i. A pop up window will indicate that the data is incorrect and how the user can correctly import the data.

    c. Scenario 3: The user copies over their database with APM entirely

        i. The website will allow the user to allow a one-time copy of their accounting data with APM to Sibyl

4. As a user, I can select a prediction the app is capable of producing and then import a corresponding dataset that will be run against the model and be displayed as a graph.

    a. Scenario 1: The user selects our expenditures model and then imports their expenditures dataset as a correctly formatted csv.

        i. The model runs successfully and plots the predictions as a graph.

    b. Scenario 2: The user selects our expenditures model but imports the wrong dataset (liabilities, maintenance requests, etc.).

        i. A pop up window will explain the error and tell the user to import the correct dataset for that prediction model.

5. As a user, I can see a visual separation on the graph between the historical data and the predictions from the model.
https://github.com/skylartwinig/sibyl/commit/bce1d5ecad9818ac9693aa790ceaaf56a068b58a

    a. Acceptance Criteria: There is a plotted vertical line at a certain data point that indicates the start of the predictions/end of historical data.

6. As a user, I can select and view past projections/graphs that have already been made.

    a. Acceptance Criteria: User can see a table of past projections, select any row, and then plot it using the plot button.

7. As a user, I can view the confidence in ML predictions, if it pertains to the particular model.

a. Acceptance Criteria: Our models meet an agreed upon accuracy level with its predictions.

8. As a user, I can use the predicted metrics to derive secondary metrics to help me understand the implications of the predicted metrics
   a. Scenario 1: I can see how the predicted occurrence of maintenance requests impact my cash flow this month
   b. Scenario 2: Based on the expected time of delivery for rent, I can view my daily account balance projections for the next 2 weeks.

9. As a user, I would like a clear, non-complicated means to efficiently understand the insights extracted from my data.
   a. Scenario 1: A UX focused design for the front end and the output of ML models should allow the user to easily understand the insights given to them without needing to be familiar with the complexities of ML models.

10. As a developer, I would like a comprehensive set of visualization formats that support a wide variety of ML models.
    a. Scenario 1: An ML developer looking to integrate their models onto our visualization dashboard can select from various graphs or projections that display continuous value insights, discrete insights, insights in the form of projection functions, insights from sequential data, etc.

11. As a developer, I would like a clear usage guide for our APIs that allow ML developers to integrate insights from their models onto our visualization dashboard.
    a. Scenario 1: An ML Developer can refer to a well documented usage guide maintained by the team behind Sibyl to clearly leverage our API to integrate the insights of their ML model with our visualization Dashboard.

12. As a user, I can customize my dashboard for better visualization.
    a. Scenario 1: User can plot multiple datasets to view multiple projections side by side
    b. Scenario 2: User can resize their graphs and move them around by dragging and dropping.

13. As a user, I want a friendly UI that allows for easy manipulation.
    https://trello.com/c/lphZGArB

    a. Scenario 1: Components on the dashboard such as graphs are draggable.

14. As a developer, I would like a clearly defined format for new data to be imported into the ML system.

    a. Scenario 1: An ML Developer can reference Sibly documentation to know how imported data will be formatted so that the new data can properly be applied to an ML model.

15. As a user, I can filter my projections based on date range (week, month, year, etc.)
https://trello.com/c/tGK5UCR3

    a. Scenario 1: User filters data and all data points can fit on the graph.

    b. Scenario 2: User filters data and graph only shows projection for small subset of dataset (one week out of years worth of data). User can click arrow buttons below to scroll through different weeks.

16. As a user, I can export my predictions in different formats. (png, jpeg, pdf)
https://trello.com/c/id6pWT97

    a. Scenario 1: User clicks on the export button and can choose the destination to save the file and what file type to store the chart as. (png, jpeg, pdf)

17. As a user, I can search for different GL accounts/past projections already uploaded to the web application.
https://trello.com/c/LUmGanyV

    a. Scenario 1: User searches a GL account and is able to select it and display the chart for that account.

    b. Scenario 2: User searches for a GL account not already uploaded. A prompt tell the user the account is not found and gives the option to upload another account.

18. As a user, I would like to have access to multiple ML prototypes to get predictions on my data

    a. Scenario 1: User uses a linear regression model to gain insights on their accounting data.

    b. Scenario 2: User decides that the linear regression might not be the most accurate based on the error so they decide to use random forest regression to get better predictions.

19. As a user, if I would like to have additional hyperparameter optimizations, I can run cross validation my data for the best results
    a. Scenario 1: User runs cross validation to run a grid search in log space over combinations of hyperparameters and selects hyperparameters with the best results.
20. As a user, I would like make sure my personally identifiable data is secure end-to-end.
    a. Scenario 1: User decides that they want to make predictions on their data. We will ensure that if there is any movement between clouds, the data will be encrypted non-identifiable (in adherence to Appfolio's privacy standards).

Technologies Employed
1. Django
2. MySQL
3. ReactJS
4. TypeScript
5. Github
6. Python
7. PyTorch
8. Scikit
9. Pandas / Matplotlib