**Personal Budget Analysis**

C964 – Computer Science Capstone

Western Governors University

**Part A: Letter of Transmittal**

February 25, 2024

Julia Smith, CTO

YourBudgetApp

123 Budget Dr

Greenville, South Carolina

Dear Ms. Smith,

This company has been growing quickly in the personal budgeting space. However, we are not substatially differentiating ourselves from the competition. While we excel at customer service, and our product is top notch in terms of quality and stability, we need something to separate us from the competiton. Something that will retain the people willing to give it a try but prefer what they are used too, and pull in people who may not have even bothered to try the product. My proposed feature will do just that.

By implementing machine learning into our product we will offer something not commonly seen in the personal budget space. The implementation of machine learning will enable our customers to see the future potential of their budgetting habits and allow a new dimension to their decision making process. The Machine learning can be implemented into the budget such that the customer chooses what data to forecast and how. With the proof of concept you approve production of we can build out a full featured forecasting system with feedback from our community.

The first iteration of the machine learning build will take a week of work from myself and on additional developer, at a cost of $3000. I will provide the data for developing and testing from my personal budget, scrubbed appropriately for my personal security. The final product will not be storing data separately from what is already used in the application, we will not be implementing any cloud based services; Everything will be done on the users local machine and open sourced just as the rest of the product currently is. My educational background is in machine learning implementations of this nature and I’m passionate about instilling strong personal finance skills into people enabling them greating flexibility and autonomy in life; Making me uniquely qualified to lead and develop this new product.

I am available to answer any additional questions and I look forward to hearing from you. thank you for considering my proposal, I hope we can move forward with this idea to provide new and exciting products to our customers.

Sincerely,

David Mitchell, Project Lead

Part A: Project Recommendation

**Problem Summary**

There are an abundance of personal budget apps on the market today, all offering their own spin on this age old product need. In an effort to distinguish ourselves from the rest of the pack we need something unique to offer our customers that will keep them around in the long term. Every budget app, including ours offers historical reporting, few if any attempt to look into the future for the individual. Implementation of machine learning will enable our company to stand apart from the competition and offer the customer something they can’t get elsewhere. In this Project we will implement machine learning forecasting in our apps suite of reporting capabilities. This will give the user access to a new dimension of information to better enable them to make decisions in their budget.

**Application Benefits**

The implemented machine learning will look at the historical data stored locally on the users system, this is the data generated by normal use of the application, income and spending information, as well as where the user allocates their available funds. The user can then select which categories of their budgets they would like a forecast for. Once they have a visual of the potential future of the decisions they are making they will have a better picture of what can be done to reach whatever goals they have in mind for themselves. This creation of value for the customer will result in value for the company and all of its stakeholder, as well as strong user retention and capturing market share from the competition.

**Application Description**

The machine learning implementation will take user generated budget data (income, expenses, transaction information, budget allocation amounts, etc), normalize the data as necessary, analyze the data and finally display through graphical visual aids and text based explanations results of the analysis based on what the user selected to be analyzed. The results will include things such as expected balance of their tracked 401k account at fixed points in the future or the balance of an emergency fund based on historical use of the emergency fund.

**Data Description**

All of the data comes directly from the customers use of our budget app. All of the data is transaction data that includes the date of the transaction, the budget category associated with the transaction, and the dollar amount of the transaction. The date, category, and transaction amount are the independent variables used by the machine learning and the predicted future balance or transaction amounts are the variables provided to the user after the machine learning has finished it work. The major limitations of this project include the accuracy and completeness of the data provided by the user, as well as smaller data sets suffering from worse predictions. The final limitation is the choice of algorithm used by the machine learning implementation can affect the accuracy of the forecasts.

**Objectives**

The objective of this project is to create a product that will create a new avenue of value for our budget customers. This will be accomplished through a new budget reporting feature that uses machine learning to forecast the future state of the user budget in a given category. The User will be able to select what portions of their budget they would like to see forecasts for, as well as how far out to forecast as well as the period of time the forecast is broken into(e.g., monthly vs weekly change). Finally the product will include visual and text based information generated as well as layman instruction on interpreting the provided forecast as well as the potential limitations of it.

**Methodology**

Development will follow the agile methodology. The agile methodology is most appropriate for this project given the need for regular feedback from the end user. Agile is flexible enough to respond to changes throughout the iterative nature of the projects development. Planning will begin with laying out the vision and big picture deliverables of the project, then breaking those down into smaller scale pieces to be built. We’ll then move on to design, laying down the specific framework of each portion of the application and how they will interact with one another. Next we will develop the actual product to be used by the end user. The Product will be sent to the internal testing teams for testing, if all goes well the product will be deployed to users. Follow a short period for user feedback we will review the product, the processes used and feedback from the user, laying out what we can do better and changes needed to the product for the user. Finally starting the whole cycle over again for the next iteration and feature set to be developed.

**Funding Requirements**

This Project will take about a week of work from two developers. The necessary code bases and libraries are Python, Sktime for the machine learning algoritmh, and MatplotLib for the graphical representation of the information. All of these are open source, freely available and do not require additional funding.

**Data Precautions**

The data used for testing the product befors launch will be provided by myself using my personal data and will be properly scrubbed before hand. The final product will use the users existing data that is stored local on their machine. There are no additional data precautions necessary since all data is stored and used locally and will not be leaving the users machine.

**Developer’s Expertise**

My qualifications to lead this project include a bachelors degree in computer science from Western Governors University(WGU) which had a focus on machine learning and its application. Beyond my academic qualifications growing up working class and attending WGU fulltime while also working full time, I have the lived experience of what people whom may not have a strong understanding of personal finance essentials want and need out of budgetting software.

**Part B: Project Proposal**

The project proposal should target your client’s technically savvy IT professional leadership. Use

appropriate industry jargon and sufficient technical details to describe the proposed project and its

resulting application. Remember, you’re establishing the technical context for your project and what it

will accomplish for the client. Typically, this section is 8 – 10 pages. Write everything in the future

tense.

**Problem Statement**

Implement a machine learning algorithm into the personal budget allowing the user to make informed decisions based on more information than is available to them from other budget software. Building up a suite of choices for regression forecasting based on the individuals budget, dynamically adjusting as necessary based on the sort of forecasting sought after by the user. A user looking for a forecast of the change in their utility bills needs a different forecast than one looking to forecast their 401K balance.

**Client (or Customer) Summary**

The user of this product is the user, or potential user, of our budgetting software. The use of regression algorithms will enable the customer to have access to a much wider range of information regarding their budget and spending habits, this additional information we provide through application of machine learning will enable the user to be a more informed budgetter and spender.

**Existing System Analysis**

The user is currently making use of budgeting software that has been operating on the same fundamentals that budgetting has always operated under. The system does not have many flaws if we looking strictly at the fundamentals of budgetting, but when we look at the human element you can find many. Most relevant to the project is a users lack of access to information to use for decision making, whether that be ignorance of personal finance principles or the unwillingness to seek out information outside of whats put infront of them. This lack of information access is at the heart of the problem this project seeks to solve.

**Data**

The raw data used in this project comes directly from budgeting app being improved upon. The data in the final product will be collected processed and managed on the local machine. Collection will happen at the time of user input into the application. Management and storage of the data will happen through a combination of a Sqlite database and Pandas library dataframes. The Sqlite database stores all of the raw transaction and budget data while the dataframe is used to retrieve and store the specific data to be used in the user requested forecasting. Normalization and transformation of the data will be accomplished using pandas. Regression processing will be handled by the sktime library and visual aid display of the data will be handled by matplotlib. Data anomalies will be excluded or normalized where possible, with warning to the final user as necessary to account for how that will affect the forecasts provided. Data used during development will be provided by myself and scrubbed for privacy before hand.

**Project Methodology**

 Describe an industry-standard methodology to be used to develop and (if applicable) deploy your

application.

 Describe the planned development of your application in each phase of the methodology, e.g.,

analysis, design, etc.

Development will follow the Agile Methodology with the following as the intial starting points for each phase:

1. Plan: Develop a python backend program for generating various forecasts based on a users personal budget data. Beginning with a Mortgage balance forecaster, and expanding from there with user feedback.

2. Design: The backend of the app will have several modules, each serving a specific purpose for the full app, to assist with future expansion of the products features. A module to serve as the user interface, here a simple command line interface, later replaced with a graphical interface when integration with the full budget program takes place. A module that serves to pull requested data from the database. Finally a module to transform, process, and display the data to the user

3. Development: here the app is developed beginning with a basic user interface for testing functionality as it’s developed, followed by working out the database connections, then the data transformation and finally the processing and display of the data.

4. Test: Here the program is passed to internal testing for validation of the apps functionality, ensure all end user facing information is presented in a manner that the greatest number of people will understand it with little to no trouble, as well as validating the accuracy of the regression used.

5. Deployment: the program is then pushed out to the budget apps main dev team to be integrated into the full application.

6. Review: working with information from the main dev team, internal lessons learned, as well as customer feedback we will adjust objectives, deliverables and methods as necessary for the next iteration of the project.

**Project Outcomes**

The final deliverables will be the following:

* Machine learning learning application, with a simple command line interface, for demonstration and testing, to be implemented into the full budget software.
* A user guide for installing the necessary components for using the machine learning software as well as a simple guide to the use of the software.
* A post-implementation report discussing the final product, lessons learned and potential future improvements or additions to the software.

**Implementation Plan**

The application will be coded on a module by module basis. The first module to be built will be the regression and plotting module, this being developed first enables the dev team to know how to structure the Dataframes and Database queries being used to limit the amount of data transformations that need to occur along the way, changes to the regression algorithms at a later date if necessary when testing with the final data will be much less code intensive than rewritting database queries given that the sktime library uses a universal syntax for the majority of its regressions.

The next module will be the Dataframe and Database modules used to retrieve and transform the data for use by the regressions. This allows us to already know how the data needs to be structured and simplifying the process of writing database queries and dataframe transformations.

The final module will be the user interface module. Given that it will be a simple interface for iterim use until implementation into the full budget app this will be the easiest to code as it only needs explanatory text and function calls. The interface will be strictly command line with popup windows for the graphics that need displayed.

Finally the application will be tested to ensure all modules communicate appropriately, all displayed graphics are scaled and labelled appropriately, and the explanatory text is sufficiently clear and understandable as well as succinct. Once that has been verified along with strong code commenting, the application will be passed onto the main dev team for implementation in the full budget app.

**Evaluation Plan**

The first iteration of the project was designed to forecast the users mortgage balance based on their payments and principal balance. Given that this does not actually need machine learning the forecasting can be compared against amoritization calculators to judge its effectiveness, as well as the mean absolute precentage error to judge the forecast against existing data. The regression does not need to be perfectly inline with the amoritization calculator as that is not its primary purpose but being reasonably close implies the regression will be useful as rough guide for the end user and the underlying code will be a good jumping off point for further functionality in the future.

Stand alone database management software can be used to test the accuracy of the database queries to ensure they are pulling data correctly when compared to querying the database directly. Finally the user interface, explanatory text and graphics can be validated through round tables with other developer and alpha testers in the full app implementation.

**Resources and Costs**

The Project does not require any additional hardware not already provided to the assinged developers. Software required includes Python and the libraries MatPlotLib, Pandas and SKTime which are freely available and will not incur any costs. Application Development will take two devs one week of work to complete at a cost of $3000.

**Timeline and Milestones**

 Provide a projected timeline, including start dates and end dates for each milestone (a table is

acceptable).

|  |  |  |  |
| --- | --- | --- | --- |
| Activity | Assigned resource | Start | End |
| Project approval | Executive board | 2/4/24 | 2/8/24 |
| Project meeting: layout project architecture and first iteration deliverables | Project team | 2/11/24 | 2/11/24 |
| regression and plotting module development | Project team | 2/12/24 | 2/12/24 |
| Sqlite and Pandas Dataframe Development | Project team | 2/13/24 | 2/13/24 |
| Interface development, testing and code clean up | Project team | 2/14/24 | 2/14/24 |
| Final project delivered, code review and final approval with manager | Project team, manager | 2/15/24 | 2/15/24 |

**Part C: Application**

The files included in the project zip file include:

* 1. \_\_pycache\_\_
* 2. db.sqlite
* 3. DBQuery.py
* 4. main.py
* 5. regression.py

**Part D: Post-implementation Report**

**A Business (or Organization) Vision**

Building a Budgetting application that stands out from the crowded market is difficult. Building an application that attracts new customers, takes market share from the competition, and retains those customers is even more difficult. We’re solving that problem by applying machine learning to the users budgetting experience. Once the included application has been fully integrated into the budgeting app there will be additional reporting along side the tradtional budget reports. These additional reports will be machine learning algorithms used to predict the future state of some portion of the users budget. The first iteration(included here), a proof of concept, predicts the users mortgage balance based on the information they put into the budget themselves.

**Datasets**

The Dataset used in the testing and development of this project is my personal budget data from the budget application this project is intended to be integrated into, The delivered product will make use of the users personal data. The raw data is stored in an sqlite database and then passed to a Pandas dataframe to be processed for regression. When The Data is Processed The program sums up the transactions for each month into a single dollar amount, this is done to ensure the data has a fixed amount of time between data points which is necessary for the machine learning algorithm to function. The data is made into two separate data sets, one with the aformentioned transformations. The second data set takes the transactions and creates a running balance of the mortgage. These two datasets are passed to the machine learning algorithm for a multivariate regression. The algorithms predictions are then passed to another dataframe to be used in the reporting and graphics shown to the user. The Data is stored locally on the users machine using with password protection and optional encryption. There are no additonal security concerns as the data never leaves the users machine nor is it moved out of the application to a separate program.

**Data Product Code**

 Review the functionality of the code used to perform the analysis and develop your application.

o Processing raw data (if not applicable, explain why).

o Descriptive methods(s) and visualizations.

o Non-descriptive method(s)

 For the non-descriptive portion, describe the following:

o The analytic methods

o Why these methods are appropriate for the project

o How they were developed, e.g., trained and tested.

 Where applicable, discuss how the data analysis supported the choosing and improving your

descriptive and non-descriptive methods.

 Through direct submission or shared links, evaluators should be given access to all sources

necessary for the development of your project.

To ensure future expansion and maintanence of the application the code has been broken into modules. The modules are the user interface, a module for querying the sqlite database and a module for the regression analysis and display of the data. The data is pulled from the database and stored into two dataframes. The first dataframe takes the transactions and sums them into per month totals, the second database does the same transformation then one additional transformation turning the transactions into running totals of the balance of the mortgage. Those dataframes are then passed to the machine learning algorithm for training and prediction. The algorithm uses a 60/40 split on the data for training and testing and, performs predictions 50 years into the future from the most recent month available in the data. The visualizations provided to the user include a scatter plot showing the data provided to the algorithm, the plots are also color coded based on which data points are used to train the algorithm and which points are used testing. This same graphic includes a line showing the algorithms predictions compared to the provided data. The second provided visual is a stem plot showing the change in mortgage balance each month, this plot excludes the transaction for the starting balance as this would make the visual unreadable. The third provided visual is a bar chart showing the balance of the users mortgage at selected periods of time in the future. The algorithm uses a multivariate first degree polynomial regression provided by the SKTime library, This method was chosen as it gave an acceptable mean absolute percentage error in testing. Future iterations of the product will include some user choice as to what kind of regression to apply and information on how to select a regression based on each regressions results.

**Objective (or Hypothesis) Verification**

The Objective of this project was to develop a product that would allow the budget applications user to make more informed choices in their budgetting and spending habits. While there is a long way for the project to go to be truly effective at this purpose, The deliverables provided take a large step in the right direction. The Product is a strong proof of concept and a good foundation for further iterating into a full fledged product for the end user.

**Effective Visualization and Reporting**

 Describe how the descriptive method(s) and visualizations supported your non-descriptive

method(s) development process. Items discussed should include:

o Data exploration.

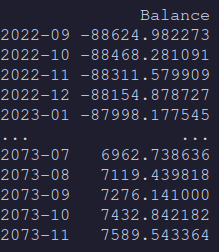
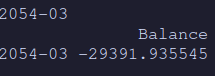
o Data analysis.

o Data summary.

o Analysis application of three visualizations (include examples of all three).

The data visualizations were most helpful in Exploring and testing the regression algorithms alongside the use of the mean absolute percentage error. Once the data was appropriately transformed for use in the regression, testing involved running the regression, graphing the regression against the provided data and checking the mean absolute percentage error. A handful of regressions were tested both as a single variable regression as well as with a second variable passed to the algoritmh. The polynomial regression was tested as both a first and second degree polynomial. Of the tested regressions the first degree polynomial gave the best visual fit as well the best mean absolute percentage error with the limited data available.

**Accuracy Analysis**

The objective metric used to assess the accuracy of the regression metric was the mean absolute percentage error. The methods for measuring the accuracy of the regression could be improved upon in the future in a number of ways depending on what forecasting gets implemented in the future. One way to test the accuracy of the regression on the used data would be to compare the results to an amoritization calculator given that the data is predicting a mortgage balance. The Image to the left shows the head and tail of the data predicted by the regression used in the project and the image below shows the predicted value at 30 years in the future. The reported mean absolute percentage error is 0.16301931145627177%, a smaller percentage is generally indicative of a better prediction, however the predicted values are low in the given regression as we know Mortgages longer than 30 years are quite uncommon the algorithm is predicting a balance beyond the thirty year point.

**Application Testing**

Continually throughout the development process testing was performed to validate each function was working appropriately. Database queries were compared along side data pulled from the database using a Database management program to ensure the validity of both the query and the data pulled when that query is passed to the database by python. The dataframes used by the regression were checked and rechecked against the database and the math checked manually to ensure the transformations were happening as expected. The regresssion was displayed graphically repeatedly to visually check it against the given data as well as running the mean absolute percentage error report after each change to check the regressions mathematical accuracy. The visual data was tested by passing the graphs back and forth between developers to check for readability and adjust as necessary.

**Application Files**

In order to run this program you will need the latest versions of the following:

* Python programming language
* Pandas Data analysis library
* SkTime machine learning library
* Matplotlib visualization library

The application files needed, and included in the zip file, are:

* 1. \_\_pycache\_\_
* 2. db.sqlite
* 3. DBQuery.py
* 4. main.py
* 5. regression.py

The db.sqlite file contains the raw budget data, and used only for reference by the python program. All other files are a part of the program directly.

**User Guide**

 Include an enumerated (steps 1, 2, 3, etc.) guide to execute and use your application.

o Include instructions for downloading and installing any necessary software or libraries.

o Your application will be considered “user-friendly” if the evaluator successfully executes

and uses your application on a Windows 10 machine following your instructions.

 Though not required, an example demonstrating how a user will apply the product is helpful for

evaluators.

Before the program can be run ensure, Python is installed along with the libraries necessary, Pandas, SkTime, and Matplotlib. To install Python and the necessary libraries follow theses steps:

**Summation of Learning Experience**

 Describe how your previous experience (academic or professional) readied you for this project.

 Describe any additional learning or resources needed to complete this project.

 Describe how this experience contributed to your concept of lifelong learning.