April 05, 2023

**C964: Computer Science Capstone**

By: Austin Kim

Task 2 – Parts A, B, C, D, E

[Part A: Project Proposal for Business Executives 3](#_Toc131625851)

[Letter of Transmittal 3](#_Toc131625852)

[Project Recommendation 3](#_Toc131625853)

[Problem Summary 3](#_Toc131625854)

[Application Benefits 3](#_Toc131625855)

[Application Description 3](#_Toc131625856)

[Data Description 3](#_Toc131625857)

[Objectives and Hypothesis 4](#_Toc131625858)

[Methodology 4](#_Toc131625859)

[Funding Requirements 4](#_Toc131625860)

[Data Precautions 4](#_Toc131625861)

[Developer’s Expertise 4](#_Toc131625862)

[Part B: Project Proposal 5](#_Toc131625863)

[Problem Statement 5](#_Toc131625864)

[Customer Summary 5](#_Toc131625865)

[Existing System Analysis 5](#_Toc131625866)

[Data 5](#_Toc131625867)

[Project Methodology 5](#_Toc131625868)

[Project Outcomes 5](#_Toc131625869)

[Implementation Plan 6](#_Toc131625870)

[Evaluation Plan 6](#_Toc131625871)

[Resources and Costs 6](#_Toc131625872)

[Timeline and Milestones 6](#_Toc131625873)

[Part C: Application 7](#_Toc131625874)

[Part D: Post-implementation Report 8](#_Toc131625875)

[A Business (or Organization) Vision 8](#_Toc131625876)

[Datasets 8](#_Toc131625877)

[Data Product Code 9](#_Toc131625878)

[Objective (or Hypothesis) Verification 11](#_Toc131625879)

[Effective Visualization and Reporting 11](#_Toc131625880)

[Accuracy Analysis 11](#_Toc131625881)

[Application Testing 11](#_Toc131625882)

[Application Files 11](#_Toc131625883)

[User Guide 12](#_Toc131625884)

[Summation of Learning Experience 12](#_Toc131625885)

[Part E: Sources 12](#_Toc131625886)

# Part A: Project Proposal for Business Executives

## Letter of Transmittal

Write a single-page cover letter to the organization’s senior leadership. The letter should be concise and target a non-technical audience. Include the following:

* Problem summary.
* Recommendation for a solution centering around your application (called a *data product* in the task directions).
* Describe how the proposed solution benefits the organization.
* Provide an estimate of the total cost (this should match the total given in *Funding Requirements* of part A).
* Expertise and experience qualify you to develop the solution.
* Include all artifacts typical of a professional (business) letter, e.g., subject line, date, greeting, signature, etc.

## Project Recommendation

Write a follow-up proposal to the letter of transmittal providing more details on how your project meets their organizational need(s). Again, the target audience is the same non-technical senior leadership from the *Letter of Transmittal*. Typically, this section is 2-3 pages; **write everything in the future tense.**

### Problem Summary

* Summarize the project.
* Describe the setting and why the project is needed.
* Briefly describe how the project meets the business’s (or organization’s) needs.
* Describe what will be delivered and achieved.

### Application Benefits

* Describe (in more detail than above) how the project meets the business’s (or organization’s) needs.
* Describe how the business (or organization) will benefit from implementing the proposed solution.

### Application Description

* Provide technical details on how the application will solve the problem.

### Data Description

* Identify the origin of the raw data.
* Describe the type (nominal, quantitative, etc.) and data structure.
* Identify dependent and independent variables.
* Describe any anomalies (e.g., outliers) and limitations.

### Objectives and Hypothesis

* Identify and describe desired outcomes of the project.
* If applicable, state a hypothesis.
* If applicable, state the desired prediction accuracy.

### Methodology

* Identify the methodology, e.g., waterfall, agile, etc., used to develop and implement the project.
* Describe why the chosen methodology is appropriate for the project.
* Provide an outline of the project methodology describing each phase, e.g., Design, Implementaion, etc.

### Funding Requirements

* Describe the project’s funding requirements, including environment, personnel, licensing, and tools.
* The funding amount should match the letter of transmittal.

### Data Precautions

* Identify any sensitive or protected data.
* If applicable, review the general guidelines for working with that data.
* If applicable, describe necessary precautions which will be taken.
* If either of the above is not applicable, explain why (public datasets, such as those from Kaggle.com, have no such restrictions).

### Developer’s Expertise

* Describe the developer’s (you) qualifications, e.g., academic training, professional expertise, experience, etc. Using future qualifications, such as your WGU degree in Computer Science, is acceptable.
* Relate the listed qualifications to the needs of the project.

# Part B: Project Proposal

The project proposal should target your client’s technically savvy IT (Information Technology) professional leadership. Use appropriate industry jargon and sufficient technical details to describe the proposed project and its 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

* Describe the problem.

## Customer Summary

* Describe the client (or customers).
* Describe why your proposed *application* (a *data product* in the task directions) will resolve the problem successfully.

## Existing System Analysis

* Describe (if any) what application(s) or tool(s) the client currently uses.
* Describe the shortcomings of this current technological environment, i.e., why your solution is needed.

## Data

* This section should include (where applicable) descriptions of:
  + The raw data set.
  + How data will be collected, processed, and managed throughout the application development life cycle: design, development, maintenance, or others.
  + How data anomalies, e.g., outliers, incomplete data, etc., will be handled.

## 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.

## Project Outcomes

* Provide descriptions of all deliverables. For example:
  + The finished application.
  + A user guide.

## Implementation Plan

* Provide an outline of how the project will be implemented. This description might include the following:
  + General strategy.
  + Phases of the rollout.
  + Dependencies.
  + Details for testing and distribution.

## Evaluation Plan

* Describe the verification method(s) to be used at each stage of development.
* Describe the validation method to be used upon completion of the project.

## Resources and Costs

* Itemize hardware and software costs.
* Itemize estimated labor time and costs.
* Itemize estimated environment costs of the application, e.g., deployment, hosting, maintenance, etc.

## Timeline and Milestones

* Provide a projected timeline, including start dates and end dates for each milestone (a table is acceptable).

# Part C: Application

Part C is your submitted application. The document only needs to include a list of any submitted files or links.

Your submitted *application* (called a data product in the task directions) must include the following features:

* Three visualizations (images). Static images are permissible.
* A *Descriptive method* = anything that describes the data.
  + Images can double count as your visualization and descriptive method.
  + Ex. mean, median, bar plot, scatterplot, k-means clustering, etc.
* A *Non-descriptive method* = anything that infers from the data, i.e., makes predictions or prescriptions.
  + Ex. classification models, regression, image recognition, etc.
* An application of “machine learning” in the non-descriptive OR descriptive method (most data analysis algorithms are acceptable -including regression).
* An interactive “dashboard.”
  + The application must be usable for solving the proposed problem. Any method enabling the user to interact is acceptable, including the command line. A GUI is *not* required.
* A “user-friendly” interface.
  + Following the “user guide” of part D, the evaluator can successfully run your application as described on their machine.
* Security appropriate to your application’s needs.

# Part D: Post-implementation Report

Create a post-implementation as outlined below. Provide sufficient detail so that a reader knowledgeable in computer science but unfamiliar with your project can understand what you have accomplished. Using examples and visualizations (including screenshots) beyond the three required is highly recommended. **Write everything in the past tense.**

## A Business (or Organization) Vision

Description of Company:

* TastyFish is a company that catches and sells fish for consumption. Due to laws and regulations for commercial fishing, the fishers are required to measure various lengths of each fish they catch. If the dimensions fall below a certain threshold, the fish must be released to ensure the younger population can breed.

The Problem:

* The problem was that due to the unstable motion of fishing boats, fishers were unable to weigh each fish individually. The sea’s waves prevented accurate weight measurements. The company’s previous method of weighing fish was letting it dry and placing it on a scale. This proved to be inefficient because the shelf life of the fish was lowered when waiting for it to dry and also during the weighing process.

How the Application Solved the Problem:

* The application solved the problem by using machine learning to predict the weight of each fish. A linear regression algorithm was applied to the machine learning model and by training the model on existing fish data, the model was able to learn how to predict the weight of a fish given its vertical length, diagonal length, cross length, height, and diagonal width.

Use of Application to Solve the Problem:

* A user can use the application to predict the weight of a fish without weighing it. If the user possesses the five dimensions of the fish, the user can input those dimensions into the trained linear regression model to predict the weight of the fish in grams. An example would be if the user inputted the string “23.2, 25.4, 30, 11.52, 4.02”, then the model would provide a predicted weight of 325.23 grams.

## Datasets

The Raw and Processed Data

* The raw data is a csv file that contains seven columns. The columns consist of the fish species, weight, vertical length, diagonal length, cross length, height, and diagonal width.
* The processed data is a csv file that contains six columns. The columns are the same as the raw data except the first column for fish species is removed.

Processing of Raw Data

* The raw data was processed by removing the first column for fish species. Little processing was needed because the raw data was minimalistic and contained mostly necessary columns. The data in its raw form was already accessible to the algorithm but processing the raw data removed the need for the dataframe object to iterate over an unused index in the array. Thus, a miniscule amount of processing power was saved by removing the first index of the arrays in the csv file. The algorithm would have returned the same results if the raw data was used with the appropriate array indexes changed.

Examples of the Raw and Processed Data

* Below is the first three rows from the raw data file named “fish\_data\_raw.csv”
  + Species,Weight,Length1,Length2,Length3,Height,Width
  + Bream,242,23.2,25.4,30,11.52,4.02
  + Bream,290,24,26.3,31.2,12.48,4.3056
* Below is the first three rows from the processes data file named “fish\_data\_processed.csv”
  + Weight,Length1,Length2,Length3,Height,Width
  + 242,23.2,25.4,30,11.52,4.02
  + 290,24,26.3,31.2,12.48,4.3056

Access to Datasets:

* The original dataset is available for download from this link:
  + <https://www.kaggle.com/datasets/aungpyaeap/fish-market>
* The dataset contained a single csv file.
* The “fish\_data\_raw.csv” file is the exact dataset that was downloaded from the link.
* The “fish\_data\_processed.csv” file is the only dataset that was used throughout the project.

## Data Product Code

Review of Code Functionality:

* In the analysis and development of the application, the code was used to execute a variety of functions.
* Overall, the product does exactly as it is supposed to do. It takes in five numbers that represent various measurements of a fish and predicts the weight of the fish using a trained linear regression model.
* Because the raw data was preprocessed manually, the code never touched the raw data. Only the processed data was touched by the code. This is because the processing only involved the deletion of one column in a csv file. This saved the code from having to iterate over that one column, which is an insignificant saving of processing power. The processing of data allowed the code to use all columns of the data.
* The code provided visualizations of data using imported libraries. These visualizations were localized to the “visuals()” function in the “main.py” file. The project used the “matplotlib” library to allow all three visualizations to appear on the screen in their individual windows. These three visualizations were the histogram, scatter plot, and graph displaying the prediction error of the regression model. The “metrics” module from the “scikit-learn” library was used to create the prediction error graph. The “pandas” library was used to create the histogram and scatter plot.
* The code aided in the development of the descriptive methods because as mentioned earlier, the the histogram, scatter plot, and graph displaying the prediction error of the regression model, were all genereated using Python libraries. A descriptive method is anything that describes data. The histogram described the distribution of each variable in the csv file. The data showed that there were certain numbers more prevelant than others in each variable. The scatter plot showed the correlation of each variable in the csv file. As each variable’s measurement increased, the scatter plot showed a positive correlation with all other measurements. The diagonal across the scatter plot displayed the histogram data. The graph displaying the prediction error of the regression model showed the difference between the predicted weight values and the actual weight values. It showed that the extreme ends of the weight predictions were more prone to error than the predictions of weight values near 200 to 700 grams.
* The code aided in the development of the non-descriptive methods because the “scikit-learn” library provided many machine learning modules to train a linear regression model. The “pandas” library was used to create the data structure for the model. After that, the “scikit-learn” library had modules to create linear regression models, train the models with data, and provide predictions with new data.
* The data analysis done through the visuals and reviewing of the csv file helped created correlations between the independent variables that lead to the prediction of the dependent variable. Seeing that there was a positive correlation between the independent variables and the dependent variable, it helped choose the scatter plot because the data points allowed one to realize that as the independent variables increased, the dependent variable increased as well.
* The analysis of the data lead to the creation and improvement of the histogram. Because of the data reviews, it was seen that the model could predict values more accurately when there was more existing data numerically similar to an input.
* This leads to the creation and improvement of the graph that displays the prediction error of the regression model. Due to the histogram showing trends in the existing data, the prediction error followed the histogram. This means that as the variables increased in quantity around a certain numerical range, the prediction error was decreased. This proves the use of the prediction error in showing that the histogram data was relevant in prediciting the accuracy of the model.

Non-Descriptive Method:

* The non-descriptive method of the project was the machine learning algorithm. The project used a supervised learning algorithm. Instead of a classification algorithm, a regression algorithm was utilized because a linear regression algorithm was used to predict the weights.
* The linear regression model was developed by using a Python library called “scikit-learn”. It was also improved using methods within that Python library.
* This use is justified because the independent variables and dependent variable were all numerical. With the five independent variables and one dependent variable, a numerical prediction called for the use of the linear algorithm. Because it was numerical and not qualitative, a logistitic regression algorithm was not used as the prediction was not a binary choice.
* The linear regression model was trained using the dataset and also tested using the dataset. The “test\_size” variable was set to 0.33 which meant ~33% of the dataset was used to train the model and ~66% of the dataset was used to test the model. Using the existing variable data, the model was trained to predict the weight of the fish given five other measurements.
* The training process was appropriate because the default “test\_size” variable is 0.25 and in any machine learning model, you need data to both train and test the prediction algorithm. Thus, the process of the algorithm development followed a reasonable plan with adequate justification.

Project Source Code:

* Because the project is a standalone app that is run in an IDE, the project only requires the “main.py” file and the “fish\_data\_processed.csv” file in order to provide full functionality. The submission of this project will include these files, which should be run in an IDE (ideally PyCharm).



## Objective (or Hypothesis) Verification

Project Objective:

* The objective of the project was to create a machine learning model that would learn from existing fish data and be able to predict the weight of a fish given its five independent measurement variables.
* The objective was met because the resulting project was a linear regression model that used supervised learning to train from the fish dataset. As of April 5, 2023, the R-squared value of the linear regression model was 0.91. Thus, the linear regression model is able to predict the weight of a fish with some margin of error.
* “R-Squared (R² or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit)” (Taylor).

## Effective Visualization and Reporting

* Describe how the descriptive method(s) and visualizations supported your non-descriptive method(s) development process. Items discussed should include:
  + Data exploration.
  + Data analysis.
  + Data summary.
  + Analysis application of three visualizations (include examples of all three).

Non-Descriptive Development Process:

* The development process behind creating and training the linear regression model was supported by the visualizations and descriptive methods. At the most basic level, the development process by first exploring the data. The dataset was explored using various visualization tools to identify trends. Once it was observed that the weight and all five other measurements of the fish has a positive correlation, a linear regression model was seen as an appropriate approach. During the data exploration, it was clear that the five independent variables positively influenced the dependent variable.
* During the analysis of the data, it was clear that the linear regression model would be better suited when using measurements that were close to the mode (highest frequency) of the data. This helped when developing the non-descriptive method (the linear regression model), because it identified that altough the R-squared value might be 0.91, there was a higher prediction error in the upper and lower ranges of the weight predictions.
* The data summaries included the three visualizations and the R-squared value. These four data summaries provided a better insight into how the linear regression model would perform in certain situations.

## Accuracy Analysis

* Describe the metric used to assess your model.
  + If not applicable, describe how future project developments could measure accuracy.
* Provide a description assessing the accuracy of your non-descriptive method.
* Include an example demonstrating the non-descriptive method and discuss the accuracy.

## Application Testing

* Describe how the application was tested.
* Explain how the testing results were used to improve the application. If no modification was necessary, explain why.

## Application Files

* Provide a hierarchical list of files and libraries required to execute (or access) your application through a Windows 10 machine.
* Describe how the files are organized in the submission.

## User Guide

* Include an enumerated (steps 1, 2, 3, etc.) guide to execute and use your application.
  + Include instructions for downloading and installing any necessary software or libraries.
  + Your application will be considered “user-friendly” if the evaluator successfully executes and uses your application on a Windows 10 machine following your instructions.

## 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.

# Part E: Sources

Taylor, Sebastion. “R-Squared**”.** Corporate Finance Institute, 4 Mar. 2023,

<https://corporatefinanceinstitute.com/resources/data-science/r-squared>.