October 15, 2022

C964: Computer Science Capstone Template

Task 2 parts A, B, C, and D

[Part A: Project Proposal for Business Executives 2](#_Toc1515453419)

[Letter of Transmittal 3](#_Toc1044087116)

[Project Recommendation 3](#_Toc1721750458)

[Problem Summary 3](#_Toc205640471)

[Application Benefits 3](#_Toc739883765)

[Application Description 3](#_Toc1967641075)

[Data Description 3](#_Toc1580263866)

[Objectives and Hypothesis 4](#_Toc1084915425)

[Methodology 4](#_Toc1455093807)

[Funding Requirements 4](#_Toc1716359054)

[Data Precautions 4](#_Toc821607173)

[Developer’s Expertise 4](#_Toc742297278)

[Part B: Project Proposal 4](#_Toc1516064891)

[Problem Statement 5](#_Toc602847551)

[Customer Summary 5](#_Toc411678477)

[Existing System Analysis 5](#_Toc1668713332)

[Data 5](#_Toc1911399030)

[Project Methodology 5](#_Toc1307313575)

[Project Outcomes 5](#_Toc1967172832)

[Implementation Plan 5](#_Toc306674583)

[Evaluation Plan 6](#_Toc849271423)

[Resources and Costs 6](#_Toc1695791095)

[Timeline and Milestones 6](#_Toc787992355)

[Part C: Application 6](#_Toc1767968421)

[Part D: Post-implementation Report 7](#_Toc22579049)

[A Business (or Organization) Vision 8](#_Toc1496854501)

[Datasets 8](#_Toc645781746)

[Data Product Code 8](#_Toc157285634)

[Objective (or Hypothesis) Verification 8](#_Toc1315546441)

[Effective Visualization and Reporting 8](#_Toc656892821)

[Accuracy Analysis 9](#_Toc1956352798)

[Application Testing 9](#_Toc871280264)

[Application Files 9](#_Toc1639588139)

[User Guide 9](#_Toc1125023771)

[Summation of Learning Experience 9](#_Toc1702287338)

# 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

## A Business (or Organization) Vision

Our problem is that we need a way to accurately predict the price of real estate so that we can not only accurately list our properties on the market but so that we can also give more accurate offers on real estate that we are looking to aquire. Being able to accurately do these things can potentially save us a lot of money and also make us a lot of money. My project is the perfect solution to this. My project takes data like the price and different attributes of thousands of properties and using machine learning can give us a way of predicting the appropriate price for properties based on a few attributes of the property. Users can easily use my project to enter some basic information on a property and quickly receive an estimate on what that property is worth(screenshot of example below).A screenshot of a computer code

Description automatically generated with low confidence

## Datasets(REMEMBER LINK TO DATASET IN REFERENCES)

My program uses a dataset from Kaggle.com. The raw data was messy and had many variables and numbers that were unnecessary for our use(example data from raw set below).



## Because of how the raw data was I went through quite a lot to process it. First I removed some of the columns that I deemed unecesarry for our use. I then made scatterplots of the data to better visualize all of the data and used that information to go through the data and remove all outliers in all columns of the set(see picture below of cleaned up data).A picture containing text, screenshot, number, font Description automatically generated

## Then with the cleaned data I made a data frame using the pandas python library using the cleaned data file. I then used the command .dropna() to drop any na values from the file and converted all of the values to ints so that the values would work properly with my algorithm. After all of that my data was split into test and training sets and was used in the algorithm.

## Data Product Code

My descriptive methods were mainly used for visualization so that I could optimize the selection of my variables and also make sure that I was able to properly get rid of outliers. I used a correlation matrix, scatter plots and a histogram to visualize the data in different ways to make sure I was making the correct decisions in manipulating the data. I then split the data into the price variable(what I am trying to predict) and the rest of the variables which will be analyzied in the prediction of the price. The data was then split into test and training sets and sent off to my non discriptive method.

My non discriptive method was then used after I used my descriptive methods. My non discriptive method was a random forest regressor. The program feeds the training data to the algorithm and allows it to learn how to properly make price predictions based on the variables given. I tried many different algorithms throughout development but consistently got the best accuracty with the random forest regressor.

## Objective (or Hypothesis) Verification

The hypotheis was that using machine learning we could accuratly predict the price of a proprty based on information about the property. The hypothesis wasn’t confirmed during this project but also can not be rejected. We were able to reach an accuracy of greater than 60% which I can not count as accurate, but it gives hope that maybe with the use of another algorithm and process and much more accurate model could be found.

## Effective Visualization and Reporting

The visualizations we used gave us a great understanding of our data and allowed us to make much better decisions with our data that we would have without the visualizations. The correlation matrix was able to give us valuable information on the relationship of every variable in our data set which allowed us to ajdust the data in good ways. We also made scatter plots that visualized the relationship of price vs every other variable which gave us great information. That information allowed us not only to recognize all of the outliers in our data but also allowed us to manipulate our data individually to best fit our needs. The histogram of price vs frequency gave us a good overview of the distribution of our data. It allowed us to see that the majority of the houses fell between 0 and about 500k in price.

## Accuracy Analysis

The metric I used to asses my model is called r-squared. R-squared is a way to measure the accuracy of a regression model. It is scored between 0 and 1, 1 being the best. R-squared is a measurement that tells how much the change in one variable explains the change in the other variable. The random forest regressor that I used scored 61% in the r-squared metric.

## Application Testing

My application was tested in a unit test style of testing. I would implement a module of code then either use tools to visualize what the code was doing to ensure the right things were happening or I would use inputs where I knew what the output should be and tested the modules using those values. An example of the visual test was when the data was being pulled into the data frame the line of code display(df.dtypes) was used to test that chunk of code. What that line of code does is it displays every category and the type that is associated with every category. This was used to make sure that the previous lines of code correctly created the dataframe and modified the types in the frame to align with our needs which was correct as every category was displayed to be an int which is what we needed to fit into our model.

## Application Files

## My program file heirarchy is quite simple and easy to understand/use. The heirarchy looks like this.

Capstone

Bin

cleanedData.csv

main.ipynb

capstoneDocument.docx

etc

lib

pyencv.cfg

share

Most of the files in the Capstone parent file are files associated with jupyter notebook which I used to code the program or my virtual environment that I created the project in.

The files that are needed for the program to run are the cleanedData file and the main file. The files are all stored under one main folder called Capstone.

## User Guide

1. Install the correct libraries on your computer

* Python 3
* Pip
* Matplotlib
* Sklearn
* Pandas
* Seaborn
* Numpy
* Jupyter notebook

1. Using command prompt go to the correct directory “Capstone”.
2. Run the command “jupyter notebook” to launch the notebook.
3. Starting in the top module where all of the import statements are, click run. Continue clicking run for every portion of code through the notebook.
4. When run the second to last portion of code you will be asked to enter the information on the property you are prediction the value of . Enter the appropriate information using no dollar signs or symbols only integers.
5. Then run the last section of code and you will be given the price estimate for the information you put in.

## Summation of Learning Experience

My previous experiences very much prepared me for this project. The classes at my previous college taught me all I needed to know about programming in python to complete this course. My own personal learning journy on machine learning and the projects I have completed in my free time on the subject helped me so much in putting together this project. Also the previous classes here at WGU helped me with this project because they all taught me great computer science fundamentals which I was able to implement in this project. Because of all of my previous experience there was no need for me to use any outside learning resources. This project contibuted to my life long journy of learning by allowing me to expand my mind and it showed me that the things I’m learning actually matter and are truly benefitting me which motivates me to continue to get better and learn more every day.