May 1, 2024

C964: Computer Science Capstone Template

**Note:** This is the latest version of the Task 2 template. The following template meets all the documentation requirements for C964 version SIM2 and SIM3. As it’s more succinct and precise, we recommend using this template for SIM2 and SIM3. However, using the [previous template](https://westerngovernorsuniversity-my.sharepoint.com/:w:/g/personal/jim_ashe_wgu_edu/EcklZjLXTB5EpDS4BVYc8SEBhT3VHy3s_9lZSIZ5aH6Q5w?e=5tCTQb) is still acceptable.

Task 2 parts A, B, C and D

[Part A: Letter of Transmittal 2](#_Toc184202999)

[Letter of Transmittal Requirements 2](#_Toc184203000)

[Letter Template 2](#_Toc184203001)

[Part B: Project Proposal Plan 4](#_Toc184203002)

[Project Summary 4](#_Toc184203003)

[Data Summary 4](#_Toc184203004)

[Implementation 6](#_Toc184203005)

[Timeline 6](#_Toc184203006)

[Evaluation Plan 7](#_Toc184203007)

[Resources and Costs 7](#_Toc184203008)

[Part C: Application 8](#_Toc184203009)

[Part D: Post-implementation Report 9](#_Toc184203010)

[Solution Summary 9](#_Toc184203011)

[Data Summary 9](#_Toc184203012)

[Machine Learning 9](#_Toc184203013)

[Validation 9](#_Toc184203014)

[Visualizations 9](#_Toc184203015)

[User Guide 9](#_Toc184203016)

[Reference Page 11](#_Toc184203017)

# Part A: Letter of Transmittal

## Letter of Transmittal Requirements

The *Letter of Transmittal* should convince senior leadership to approve your project. Write a brief cover letter (suggested length 1-2 pages) describing the problem, how the application (part C) applies to the problem, the practical benefits to the organization, and a brief implementation plan. Include all artifacts typical of a professional (business) letter, e.g., subject line, date, greeting, signature, etc.

The letter should be concise and target a non-technical audience. Include the following:

* A summary of the problem.
* A proposed solution centering around your application.
* How the proposed solution benefits the organization.
* A summary of the costs, timeline, data, and any ethical concerns (if relevant).
* Your relevant expertise.

## Letter Template

11/28/2024

Jordan Axwell

Tax Cutters

123 Equity Lane, Suite 400, Fairview, TX 75069

Dear Mr. Axwell,

As a resident of Harris County, a property owner, and a real estate data engineer, I know that property owners are always looking for ways to cut costs, and reducing the taxable amount of the property is a way to do that. Protesting property taxes is an easy way to save several hundred dollars a year, but as a company that does the work for protesting the taxes, it can be challenging to know who a potential customer is. In the past, advertisements were sent to entire neighborhoods, and very few people protested their taxes. I have used the HCAD (Harris County Appraisal Districts) data to create an advanced machine-learning model of the property’s assessed values. This model is made from the attributes of the properties listed in the HCAD database, which my program will automatically download and load into a local database for exploration and model generation. The program will output a list that can be used to quickly determine who would be excellent candidates for having their homes protested. This list can reduce your company's advertising costs by showing the top candidates for home protests. It can also create personalized advertisements to show how much money they can save using your service. This program can also reduce the time and effort of protesting the homes' values because it ranks all homes equally, given their features. It can also reduce the processing time your caseworkers need to complete a protest. Knowing how much the service will cost and the potential savings to the customer, the property owner can be sent a net optimistic estimate on the advertisement. This will increase the chances of the property owner becoming a new customer and the chances that the property owner will use the service next year, expanding the company's revenues and decreasing advertising costs.

Sincerely,

Blake Bowden,

Real Estate Data Engineer

# Part B: Project Proposal Plan

The project proposal should target your client’s middle management. This audience may be IT professionals but have limited computer science expertise. 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 how it will be implemented for the client. **Write everything in the future tense.**

## Project Summary

* Describe the problem.
* Summarize the client and their needs as related to the problem.
* Provide descriptions of all deliverables. For example, the finished application and a user guide.
* Provide a summary justifying how the application will benefit the client.

Every year, Tax Cutters must pay to send advertisements to thousands of properties, hoping that the property owners will read the ad and purchase their service. The problem is that many of the homes are not good candidates for having their property taxes protested. Tax Cutters spends thousands of dollars mailing advertisements to properties that would not benefit from our service and are not likely to purchase our service. This spray-and-pray method wastes money and can damage the company's reputation if we try to protest values on already reasonably priced houses. Identifying properties that are good candidates would help target our efforts and deliver more benefits for our customers.

By creating a model to estimate property values and compare the difference to the county's assessed values, we can determine the best properties to send advertisements for our services.

The machine learning model will estimate a significant subset (10,000) of the property values using the property’s attributes (square footage, bedrooms, bathrooms, rank description, etc.) and apply the predicted values to all single-family homes in the county. This application will output a CSV file that can easily be opened in Excel or imported into a database for the property evaluators to find what properties are overvalued by the county. With a more data-driven, targeted approach to advertising, the cost will be lower than mailing advertisements to every property. When the properties are identified, the mailing information will also be included in the out file, lowering the effort to create advertisements and send them out on time.

The main objective of this project is to deliver a list of good candidates for advertising. To do this, the data from HCAD must be downloaded and organized into a structured format and preprocessed in formats for the algorithm to understand. The model can be less accurate since various house types, sizes, and locations exist, and the value of the properties is subjective. If needed, I will deliver a functioning program that can update the model and apply it to the entire dataset. A list will be delivered along with the documentation on updating and generating the model.

## Data Summary

* Provide the source of the raw data, how it will be collected, or how it will be simulated.
* Describe how data will be processed and managed throughout the application development life cycle: design, development, maintenance, etc.
* Justify why the data meets the needs of the project. If relevant, describe how data anomalies, e.g., outliers, incomplete data, etc., will be handled.
* Address any ethical or legal concerns regarding the data. If there are no concerns, explain why.

The data will be downloaded from the Harris County Appraisal District (HCAD) website. HCAD maintains a database of property attributes and must make them publicly available. The county data is updated about once per month, so a Python script triggered once a month can download the data will need to be created. If the script fails, the data can be downloaded manually using a web browser and copied to the correct folder for the Python script to read. Below are the links to the main pages.

* + [HCAD Property Data](https://hcad.org/pdata/pdata-property-downloads.html): Real Property Data and Building Information
  + [HCAD GIS Information](https://hcad.org/pdata/pdata-gis-downloads.html): Tax Parcels

The downloaded data from HCAD Property data are text files exported from the county’s Microsoft SQL server that have been compressed into zip files. The text files will be extracted and read with pandas and then loaded into a SQLite database for local development. With the data in an SQLite database, a combination of SQL and Python Pandas methods will join and organize the data into a structured format that the machine learning algorithm can use to create the model.

The HCAD GIS information of the tax parcels is the only file that is not a text file. This file requires the entire folder of files to be used and will be read by the Python package Geopandas. The point of the GIS information is to extract the latitude and longitude of each property and use the Haversine formula to calculate its distance in miles from the center of Houston.

The following fields will be used from the downloaded and extracted data:

* Improvement in square footage (Integer)
* Date erected (Integer)
* Land square footage (Integer)
* Perimeter in feet (Integer)
* Bedroom/whole bath/half bath/total number of rooms (Integer)
* Latitude/longitude and the distance from the center of Houston (Real)
* Description of property (poor, very low, low, average, good, excellent, and superior) (Text into ordinal values Real)
* Extra feature values (Integer)
  + Pool/pool heater
  + Brick Garage
  + Frame detached garage
  + Shed
  + Carport
  + Foundation Repair
  + Cracked slab
* Assessed Value of the property (Integer)

Once the model has been created and tuned to have a fit greater than 0.80 R-squared, it will be saved locally on the computer. The model will be applied to the entire data set, and then the data set will be filtered down to only include properties that are good candidates for protest. The entire and filtered dataset will be saved to the local computer, but if there are more than 1.2 million rows, the entire data set cannot be opened with Excel.

## Implementation

* Describe an industry-standard methodology to be used.
* An outline of the project’s implementation plan. This outline can focus on the project’s development as a whole or on only implementing the machine learning solution.

The project will use the waterfall method since the data does not change daily, and the work is well-known. The product that will be delivered is a simple CSV file that users can open with Excel or Google Sheets, and it will only need to be updated yearly. It will be generated and used to send advertisements and will not be updated again until next year. Since the timeline to send advertisements and the protest dates are fixed, all the planning must be done upfront.

1. Requirements – Meet with the users of the CSV file to determine who to send advertisements to and get their requirements for the output file.
2. Design – ensure that the data required from step 1 is available and we can meet the timeline and format.
3. Implementation – the application will be created to ensure it meets the requirements and design.
4. Verification – the model will be validated for outliers and given to the end user for evaluation.
5. Maintenance – feedback from the end user will be taken, and reasonable issues or features will be fixed or included during the off-season.

## Timeline

* Provide a projected timeline, including each milestone's start and end dates (a table is not required but encouraged).

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone or deliverable** | **Duration**  **(hours or days)** | **Projected start date** | **Anticipated end date** |
| Automated data download and extraction | 6 hours | 11/28/24 | 11/28/24 |
| Loading data to SQLite database | 3 hours | 11/29/24 | 11/29/24 |
| Cleaning and organizing data into structured format from multiple files | 24 hours | 12/2/24 | 12/5/24 |
| Building model and hyperparameter tuning | 16 hours | 12/5/24 | 12/6/24 |
| Exporting list and building a complete pipeline | 8 hours | 12/9/24 | 12/10/24 |

## 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.

During the requirements stage, the end user will be crucial in validating what needs to be delivered. The end user must validate much of the work throughout the project. The product's design will also require that the end user sign off on it before development and that IT is ok with the security aspect of the program. During the development, the model will be evaluated by fitting a linear regression of the assessed values with the predicted values of the testing set and achieving a minimum of 0.8 R-squared value on the testing set of 10,000 samples. The model will be saved so the list can be updated quickly without having to retrain the model. The end users will review the final output file to ensure the data is valid.

## 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.

The software used to create this is all free and open-sourced. Python and all of the libraries used are free to use. The development of this program will be done with Visual Studio Code, a free text editor from Microsoft. It handles Python files and Jupyter notebooks well and works with the virtual environments built into Python.

The data the model will be built on is also free and publicly available on the internet.

The hardware costs are minimal; for this to run, a laptop will be used, and it must have a minimum of 32GB of RAM to run the model. The space complexity will require it. I estimate $1000 for a suitable laptop with a 300Mbs internet connection to download the data. The program and data will be run locally on the computer; no hosting fees will be required.

The labor will be charged at $70 per hour. The estimated time is 57 hours to complete the project so that the labor will be $3,990. A retainer fee of $40 per month will ensure that I can support any setup or bugs in the program; if you decide not to keep me on retainer, it will be $90 per hour for support and new features.

# Part C: Application

Part C is your submitted application. This part of the document can be left blank or used to include a list of any submitted files or links.

The minimal requirements of the submitted *application* are as follows:

1. **The application functions as described.** Following the ‘User Guide’ in part D, the evaluator must be able to review your application on a Windows 10 machine successfully.
2. **A mathematical algorithm applied to data,** e.g., supervised, unsupervised, or reinforced machine learning method.
3. **A “user interface.”** Following the ‘User Guide’ in part D, the client must be able to use the application to solve the proposed problem (as described in parts A, B, and D). For example, the client can input variables, and the application outputs a prediction.
4. **Three visualizations.** The visualizations can be included separately in the application, which could be better or possible; e.g., the visualizations describe proprietary data, but the application is customer-facing.
5. **Submitted files and links are static and accessible.** All data, source code, and links must be accessible to evaluators on a Windows 10 machine. If parts of the project can be modified after submission, matching source files must be submitted. For example, if the application is a website or hosted notebook, the `.html` or `.ipynb` files must be submitted directly to assessments.

Ideally, submitted applications should be reviewable using either Windows or Mac OS, e.g., Jupyter notebooks, webpages, Python projects, etc. If the source files exceed the 200 MB limit, provide screenshots or a Panopto video of the functioning application and contact your course instructor.

# 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 recommended (but not required). **Write everything in the past tense.**

## Solution Summary

* Summarize the problem and solution.
* Describe how the application solves the problem from parts A and B.

## Data Summary

* Provide the source of the raw data, including how it was collected or simulated.
* Describe how data was processed and managed throughout the application development life cycle: design, development, maintenance, etc.

## Machine Learning

For each employed method (at least one is required), provide the following:

* Identify the method and what it does (the “what”).
* Describe how the method was developed (the “how”).
* Justify the selection and development of the method (the “why”).

## Validation

For each employed method described in the section above, provide the following:

* A proper validation method (typically a model performance metric).
* Results of the validation method *or* a future to obtain those results.

## Visualizations

Identify the location of at least three unique visualizations. They can additionally be included here.

## User Guide

1. Please install Google Chrome on your computer.
2. Download the latest version of Python 3.12 from python.org/download
   1. [Here](https://www.python.org/ftp/python/3.12.8/python-3.12.8-amd64.exe) is the latest version for an x86\_64 processor, the most common in business laptops
   2. Install it, but ensure that you check the add python to path box and then click next.
3. Download the working files into a folder and then hold shift, right-click in the folder (not on a file), and click on open Powershell window here.
4. Copy this command, paste it into the PowerShell window, and enter. This will create a virtual environment.
   1. python -m venv .venv
5. Now activate the Python environment by entering the following command in the PowerShell window.
   1. .venv/Script/activate.ps1
6. Now, we will install the required Python packages with the following command in the Powershell window. This will require an internet connection to complete and may take several minutes, depending on the internet connection and processor of the computer.
   1. pip install -r .\requirements.txt
7. To download the data from HCAD, enter the following command in the PowerShell window.
   1. Python .\Setup.py
   2. While this is running, a Google Chrome instance will open and may appear to be doing nothing, but it lets the page load before downloading the information. If it doesn’t work the first time, try it again since the caching of the HCAD website sometimes stops.
   3. This will download 1.5GB of data so that it may take some time, depending on your connection. After the download finishes, it will extract the data and create and load it to a database in the same directory. Give it time to complete; the text on the PowerShell window will say “Done!” when finished.

# Reference Page

Following APA guidelines, include references for cited works, e.g., (Author, year). References are not required; this page can be removed if no references are used. To cite sources used for code, you should include the references as code comments within the source code.