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 clearer, we recommend using this template for both 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 1](#_Toc98598250)

[Letter of Transmittal Requirements 2](#_Toc1738085866)

[Letter Template 2](#_Toc1133266111)

[Part B: Project Proposal Plan 3](#_Toc1370766476)

[Project Summary 4](#_Toc904507251)

[Data Summary 4](#_Toc1736393957)

[Implementation 4](#_Toc1241988654)

[Timeline 4](#_Toc1357178365)

[Evaluation Plan 5](#_Toc623361460)

[Resources and Costs 5](#_Toc1538507987)

[Part C: Application 5](#_Toc1471073175)

[Part D: Post-implementation Report 6](#_Toc651895932)

[Solution Summary 7](#_Toc1134136520)

[Data Summary 7](#_Toc182221765)

[Machine Learning 7](#_Toc1505466430)

[Validation 7](#_Toc391434166)

[Visualizations 7](#_Toc201059345)

[User Guide 7](#_Toc1365484010)

[Reference Page 8](#_Toc1702353417)

# 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 has to pay to send advertisements to thousands of properties in the hope that the property owners read the advertisement and purchase their service. The problem many of the homes are not good candidates for having their property taxes protested and Tax Cutters spends more than they need to advertise. Reducing the advertisement cost will help increase the EBITA (earnings before interest, taxes, depreciation, and amortization) and will benefit the companies profits at the end of the year.

I will deliver a Jupyter notebook that will take prepared data from the Harris County Appraisl District (HCAD) and create a model that will predict the property values with at least an 0.80 R-squared fit. This model will be used to apply to all of the properies and where the counties assessed value is greater than the models predicted value, there will be a property that should be sent an advertisement to have their property taxes protested. To help speed this up even the mailing address will be incorporated with the dataset. The properties that are good candidates will exported as a CSV file that can be viewed in excel or imported into a database for the customer to review easily without the need of programming knowledge. There will be a userguide on how to setup and run the notebook and how to view the exported data.

Data Summary

* Provide the source of the raw data, how the data 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, or others.
* 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 is publically available from HCAD’s website. I have included two CSV files where the data has been prepare for analysis. This process was automated with a python script that would download the zip files from the HCAD website and extract the text files that are exported from the HCAD’s property Microsoft SQL server. I then have the text files loaded into a SQLite database so the can data can be prepared and joined into one Pandas Dataframe.

I did some heavy SQL queries to encode property descriptions from text to an ordinal field dscr\_e; where the descriptions poor, very low, low, average, good, excellent and superior were encoded as 0,1,2,3,4,5,6. This is fitting for the data since each one does have an acutal value to it depending on the order. I found the order in the documentation that was supplied by the county on their website. I also extracted the most influential 11 features from the extra\_features.txt file and pulled their corresponding value for each property and cleaned the data by filling in missing values with a 0. So if the property has a pool, it will have a value that the pool is assessed, but if there is no pool then the value is 0. Another tricky features that was manipulated to work was the number of bedrooms, bathrooms, half baths and total rooms. This was in a table that needed to be summed and then pivoted with a sub query and cleaned to have missing values filled with a 0. The last filtering of the data was a simple filter the improvement type as 1001 (Single family homes), were built after 1900, has more than 50 improvement square footage and has an assessed value more than $0. Once this was pulled in there were still duplicate account numbers if there is more than one building on the account. So I grouped the accounts and summed the improvement square footage, took the minimum date erected (oldest building), maximum building number to account for how many buildings are on the account and averaged everything else since the data was duplicated. After all of this there are still over 1 million properties listed in Harris county, it is the 3rd most populous county in the United States with over 4.7 million residents.

* + [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

All of the processing of the data is scripted and can be done on a computer that has admin rights and it can be prepared in less than 15 minutes depending on the internet connection. This script can also be triggered to run on a schedule so Tax Cutters and have the data ready when the analysis arrive to work.

Duringn the design phase the data resided on the HCAD website and the documentation was being reviewed to see what data needed to be extracted and what the field names mean and where their drscsiptions can be found.

## Implementation

* Describe an industry-standard methodology to be used.
* An outline of the project’s implementation plan. The focus can be the project’s development or the implementation of the machine learning solution.

## Timeline

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

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected start date | Anticipated end date |
|  |  |  |  |
|  |  |  |  |

## 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.

# 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 successfully review your application on a Windows 10 machine.
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 towards solving 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 when including them in the application is not ideal 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, consider providing 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 provides a solution to the problem from parts A and B.

## Data Summary

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

## 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

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.
* Give an example of how the client should use the application.

# Reference Page

Following APA guidelines, include references for any cited works, e.g., (Author, year). References are not required, and 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.