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 2](#_Toc184180828)

[Letter of Transmittal Requirements 2](#_Toc184180829)

[Letter Template 2](#_Toc184180830)

[Part B: Project Proposal Plan 4](#_Toc184180831)

[Project Summary 4](#_Toc184180832)

[Implementation 5](#_Toc184180833)

[Timeline 5](#_Toc184180834)

[Evaluation Plan 6](#_Toc184180835)

[Resources and Costs 6](#_Toc184180836)

[Part C: Application 7](#_Toc184180837)

[Part D: Post-implementation Report 8](#_Toc184180838)

[Solution Summary 8](#_Toc184180839)

[Data Summary 8](#_Toc184180840)

[Machine Learning 8](#_Toc184180841)

[Validation 8](#_Toc184180842)

[Visualizations 8](#_Toc184180843)

[User Guide 8](#_Toc184180844)

[Reference Page 10](#_Toc184180845)

# 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.

### Problem Summary

Every year, Tax Cutters must pay to send advertisements to thousands of properties, hoping that the property owners will read the advertisement 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 is a waste of money and can potentially damage the company's reputation if we try to protest values on already fairly 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.

### Application Benefits

The machine learning model will estimate a large 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.

### Data Description

The data will be downloaded from the Harris County Appraisal District (HCAD) website. They maintain a database of the properties and must make it 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. The downloaded data are text files exported from the county’s Microsoft SQL server; the text files will be read with pandas and stored in 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 model can inject.

* + [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 following data will be used:

* Improvement in square footage (Integer)
* Date erected (Integer)
* Land square footage (Integer)
* Perimeter in feet (Integer)
* Bedroom/full 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)

### Objectives and Hypothesis

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 does not need to be the most accurate since a wide variety of house types, sizes, and locations exist. The model will predict the prices, and then a simple comparison of the predicted values to the current assessed value from the county will be made. If the property's value is favorable to the model, it will be a good candidate for protest; if the assessed value is favorable, then it will not be sent an advertisement.

My hypothesis is that using the Extra Random Trees Regression algorithm, a model can be created to estimate the property value using the features above. This will eliminate more than 40% of the properties from the overall list to send advertisements, reducing advertising costs by about 40% in most areas.

### Project Methodology

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 the users can open with Excel or Google Sheets and will not need to be updated often. It will be generated and used to send advertisements and will not be updated again until next year. Since the timeline to send advertisements is fixed and the protest dates are fixed, all the planning will need to 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.

### Budget Requirements

All of the development will be done with free open-source software, and the data is publically available at no cost. The costs will come from labor and hardware requirements

## 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 solves 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.