October 14, 2023

C964: Computer Science Capstone Template

Note: This is the latest version of the Task 2 template. Following this template meets all the documentation requirements for C964 version SIM2 and SIM3. As it’s more succinct and clear, 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](#_gjdgxs)

[Letter of Transmittal Requirements 2](#_30j0zll)

[Letter Template 2](#_1fob9te)

[Part B: Project Proposal Plan 3](#_3znysh7)

[Project Summary 4](#_2et92p0)

[Data Summary 4](#_tyjcwt)

[Implementation 4](#_3dy6vkm)

[Timeline 4](#_1t3h5sf)

[Evaluation Plan 5](#_4d34og8)

[Resources and Costs 5](#_2s8eyo1)

[Part C: Application 5](#_17dp8vu)

[Part D: Post-implementation Report 6](#_3rdcrjn)

[Solution Summary 7](#_26in1rg)

[Data Summary 7](#_lnxbz9)

[Machine Learning 7](#_35nkun2)

[Validation 7](#_1ksv4uv)

[Visualizations 7](#_44sinio)

[User Guide 7](#_2jxsxqh)

[Reference Page 8](#_z337ya)

# 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

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

* Every year, Tax Cutters pays to send advertisements to thousands of properties, but not all properties are good candidates for protesting their taxes. This leads to many wasted advertisements.
* There are two different ways to increase EBITA: one is to grow, and the other is to reduce costs. I am offering a way to minimize advertisement costs.
* I will deliver a Jupyter notebook that will take in the housing data from HCAD, create a model based on that data using a sample, and apply that model to the rest of the properties. This list will then be exported into a CSV file with the data, estimated assessed values, and even the mailing address of the property to which you can send an advertisement.
* Instead of sending advertisements to everyone in a neighborhood, you can use this model and CSV output to know the best houses to send an ad to, reducing advertising costs.

## Data Summary

* Using a Python script during the setup will automate Chrome to download the data in zip files from HCAD, unzip it, and load it to an SQLite database on the local machine.
  + [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 raw data will be downloaded from HCAD and put into a folder where only the required tables will be extracted from the zip file. Once extracted, each file represents a table from a database, so it will be loaded into a local SQLite database where a single query can be created to quickly combine and load the data into a Pandas data frame. During the design, the data is reviewed for meaning from the HCAD documentation and cleaned to be processed in my program. During the development, a sample of data will be pulled randomly because there will be over one million properties with dozens of attributes. These attributes will be studied to look for realizations, and feature engineering will be done to make them additional meaningful attributes.
* This is the data that the county uses to assess the property values each year. It is only fit to use the same data to see if I can make a different conclusion about the home prices that will benefit the property owner. There will be missing data with over one million rows in multiple tables, so I will remove rows with null values and use the inner quartile range to remove the assessed values of properties to remove outliers and shrink the dataset. There is more than enough data to remove several rows and still have a robust model.
* All of the data is publicly accessible on HCAD’s website. So, no ethical concerns exist about using this data or displaying/sharing it with anyone, but the data does contain property owner's names, addresses, and home values.

## Implementation

* This project will use CRISP-DM to achieve this goal.
* 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.
  + I will start with the voice of the business/customer so I can understand their requirements and problems, what success looks like, and how I can translate how project success will be measured.
  + With the goals in mind, data preparation starts with the data source's who, what, and where. I must find a way to take data from a reliable source and organize it so the machine learning models can read it. Then, I must cleanse the data by removing missing values and using interquartile ranges to identify and remove outliers. With the remaining values, feature engineering will be used. I create and organize additional features from the existing features, such as the distance from the city's center, using latitude and longitude coordinates.
  + Modeling will be done using sci-kit-learn libraries, train test split to split the data into random training and testing sets, and extra tree regressors and gradient boosting regressors. The data will first be sampled with pandas, so from the over 1 million properties, 10,000 will be picked and used to create a prediction model of the home values from the 10,000 home features and assessed value from the county. Grid Search CV will be used for hyperparameter tuning to increase the model's accuracy, measured by the R-squared value of the estimated values against the predicted values. Both models will be used to identify the expected values of the fill data set.
  + The model will be evaluated using the R-squared value of the assessed values vs. the predicted values. Then, the residuals will be analyzed to see if there are any anomalies or trends and to ensure that there is no overfitting. Overfitting is not typical for ensemble models like extra trees, but bias can be present and will be analyzed.
  + The model will be deployed with a terminal application or Jupyter Notebook. The overall output will be a CSV file that users can open in Excel and filter for properties they are looking for.

## Timeline

* Provide a timeline, including projected start 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 varaibles and the application oututs 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 are able to be modified after submission, then 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

* To reduce the cost of advertisement for Tax Cut, a machine learning algorithm will predict the assessed values based on the features of each property and a list will be generated to show the properties where their assessed value is greater than the estimated value.
* The application will use a sample of the property values to create model using an ensemble method (Extra Random Trees and Gradient Boosting Regression) to predict the assessed value of the properties..

## 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:

* Indentify the method and what it does (the “what”).
* Descirbe how the method was developed (the “how”).
* Justifiy the selection and developement of the method (the “why”).

## Validation

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

* An appropiate validaiton method.An appropiate validaiton method.
* Results of the validation method or a future plan to obtain those results.

## Visualizations

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

## User Guide

Include an enumerated (steps 1, 2, 3, etc.) guide to execute and use your application.

* Include instructions for downloaing and installing any necessary software or libraries.
* Provide 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 requried, and this page can be removed if no references are used. To cite sources used for code, you should include the referfences as code comments within the source code.