April 20, 2023

C964: Computer Science Capstone

Task 2 parts A, B, C, and D

[**Part A: Project Proposal for Business Executives** 3](#_Toc132786090)

[Letter of Transmittal 3](#_Toc132786091)

[Project Recommendation 3](#_Toc132786092)

[Problem Summary 3](#_Toc132786093)

[Application Benefits 3](#_Toc132786094)

[Application Description 3](#_Toc132786095)

[Data Description 4](#_Toc132786096)

[Objectives and Hypothesis 4](#_Toc132786097)

[Methodology 4](#_Toc132786098)

[Funding Requirements 4](#_Toc132786099)

[Data Precautions 4](#_Toc132786100)

[Developer’s Expertise 4](#_Toc132786101)

[**Part B: Project Proposal** 5](#_Toc132786102)

[Problem Statement 5](#_Toc132786103)

[Customer Summary 5](#_Toc132786104)

[Existing System Analysis 5](#_Toc132786105)

[Data 5](#_Toc132786106)

[Project Methodology 5](#_Toc132786107)

[Project Outcomes 5](#_Toc132786108)

[Implementation Plan 6](#_Toc132786109)

[Evaluation Plan 6](#_Toc132786110)

[Resources and Costs 6](#_Toc132786111)

[Timeline and Milestones 6](#_Toc132786112)

[**Part C: Application** 7](#_Toc132786113)

[**Part D: Post-implementation Report** 8](#_Toc132786114)

[A Business Vision 8](#_Toc132786115)

[Datasets 8](#_Toc132786116)

[Data Product Code 8](#_Toc132786117)

[Objective (or Hypothesis) Verification 9](#_Toc132786118)

[Effective Visualization and Reporting 9](#_Toc132786119)

[Accuracy Analysis 9](#_Toc132786120)

[Application Testing 10](#_Toc132786121)

[Application Files 10](#_Toc132786122)

[User Guide 10](#_Toc132786123)

[Summation of Learning Experience 10](#_Toc132786124)

[**References** 11](#_Toc132786125)

**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.

April 20th, 2023

Vivian Harper, CEO

Dryft Bikeshare

101 Nowhere Street

Salt Lake City, UT 84101

Mrs. Harper,

Sincerely,

Jesse Mckinney

## 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.

The project will be conducted in two main phases. First we will explore the data and design our machine learning model following the SEMMA methodology. Finally, we will follow the waterfall methodology in developing a user-friendly web dashboard for stakeholders to make use of the predictive model.

SEMMA stands for sample, explore, modify, model, and assess.

* **Sample**: We will collect the data from Dryft’s rental logs.
* **Explore**: We will preform exploratory data analysis within a Jupyter notebook. Using data visualizations we will identify key features and relationships useful for training a predictive model.
* **Modify**: The data will be processed into a state suitable for training a machine learning model. We will ensure that there are no null entries and that only features relevant to the prediction are included.
* **Model**: We will train a variety of configurations for machine learning regression models to predict demand given relevant data.
* **Assess**: The models will be compared based on their prediction accuracy with a portion of the dataset set aside for testing. Using scoring metrics like mean absolute error, we will come to a decision on the best model to use for the final product.
* **Requirements**: Requirements for the dashboard will be gathered based on feedback from stakeholders and lessons learned in the data exploration stage.
* **Design**: The dashboard’s tools and interface will be designed to meet the gathered requirements.
* Implementation: The application will be implemented based on the created designs.
* **Testing**: The app will be hosted locally tested to ensure it is working correctly and meets the requirements.
* **Deployment**: The web app will be pushed to github and deployed to the streamlit community cloud.

### 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

1. Itemize hardware and software costs.
2. Itemize estimated labor time and costs.
3. 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 Vision

* Describe the problem.
* Describe how your application solves the problem.
* Describe how the user can use your application to solve the problem.  
    
  An example can be helpful.

The purpose of this project was to use data analytics and machine learning on the rental logs of bikesharing company Dryft to enable stakeholders to better predict future demand for rentals at given stations. The created machine learning model predicts the relative demand of a rental station given information about the date, time, and weather. A user-friendly dashboard presents these predictions in an intuitive way. Stakeholders can view an overview of demand across the city at a given time. Station managers can use the “day summary” feature to predict the expected demand for their station throughout the entire day.

## Datasets

* Describe the raw and processed data.
* Describe how the raw data was processed. Explain if little processing was needed, but at a minimum, this will include a description of how the data was made accessible to your algorithm.
* Provide an example (or examples) of the raw and processed data (if applicable).
* Provide access to any datasets used.

The proof of concept dataset used simulates rental logs provided by Dryft stations throughout the period of a year. It should be replaced with real-world logs before final deployment. It emulates trends that can be seen in publically available datasets that focus on the demand for whole cities (). The cleaned and dirty datasets can be found as compressed csv files here:

Raw data example:

**Datetime**: A python datetime object of the hour that this log occurs.  
**Holiday**: Boolean value on whether the current day is considered a holiday.  
**Apparent** **Temperature**: The reported “feels like” temperature in farenheit for the given hour. Factors in the wind chill and heat index.  
**Weather** **Intensity**: Reported severity of weather such as rain or snow. 0 means a clear sky while 1 means heavy rain/snow. Anything more severe than heavy rain is also clamped to 1, as heavy rain itself is enough to drop rental counts to zero.  
**Station** **Mean**: The mean of previous hourly rental counts for the given station.  
**Rentals**: Count of the rentals that occurred over this hour at the station.

Confirm there are no null values in the data

Extract relevant data from datatime

Feature generation

## Data Product Code

* Review the code’s functionality used to perform the analysis and develop your application.
  + Processing raw data (if not applicable, explain why).
  + Descriptive methods(s) and visualizations.
  + Non-descriptive method(s)
* For the non-descriptive portion, describe:
  + What analytic methods were applied?
  + Why these methods are appropriate for the project
  + How they were trained and tested.
* Where applicable, discuss how the data analysis supported the choosing and improving of your descriptive and non-descriptive methods.
* Through direct submission or shared links, evaluators should be given access to all sources necessary for developing your project.

## Objective (or Hypothesis) Verification

* Describe the project’s objective (or hypothesis).
* Discuss if and why the objective (or hypothesis) was met (or not met).

## Effective Visualization and Reporting

* Describe how the descriptive method(s) and visualizations supported your non-descriptive method(s) development process. Items discussed should include:
  + Data exploration.
  + Data analysis.
  + Data summary.
  + Analysis application of three visualizations (include examples of all three).

## Accuracy Analysis

* Describe the metric used to assess your model.
  + If not applicable, describe how future project developments could measure accuracy.
* Provide a description assessing the accuracy of your non-descriptive method.
* Include an example demonstrating the non-descriptive method and discuss the accuracy.

In the data exporation notebook I tested a variety of the regression models and compared their accuracy to make the best decision in the final application. I set aside 30% of the dataset for testing the accuracy of the trained models, as testing the models on data they were directly trained with would not properly reflect their accuracy.

Chart, bar chart

Description automatically generated

## Application Testing

* Describe how the application was tested.
* Explain how the testing results were used to improve the application. If no modification was necessary, explain why.

## Application Files

The application files can be found on this webpage:

* **Dashboard.py**

The dashboard web app, using the streamlit library to create a web front-end with just python.

* **DataExploration.ipynb**

The jupyter notebook used for data cleaning and exporatory data analysis. Created in google colab with dirty\_data.zip uploaded to the workspace.

* **Dirty\_data.zip**

The uncleaned data in a compressed csv file.

* **Clean\_data.zip**

The cleaned data in a compressed csv file.

* **Requirements.txt**

Lists the required python modules to run Home.py. Required for web hosting.

## User Guide

**To use the cloud hosted web application:**

1. On a WebGL compatible machine and web browser, navigate to:
2. If you see loading bar saying “Running generate\_model()” just wait. It is training the ML model.
   1. This only occurs if the cloud server’s memory has been cleared, afterwards the dashboard opens quickly for all users. It could take a minute.
3. On the “City Overview” tab
4. On the “Day Summary” tab
5. On the “Analytics” tab you can see various visualizations of the training data.

**To host the application on a local machine (Optional):**

1. Download and extract the project folder to a machine with python 3 installed.
2. Create a python or anaconda virtual environment and install the following modules:
   1. Pandas
   2. Numpy
   3. Scikit-learn
   4. Streamlit
   5. Streamlit\_option\_menu
   6. Pydeck
3. In a terminal for the virtual environment, navigate to the project folder and type the command: **streamlit run Home.py**
4. The locally hosted web app should open within your web browser. Follow steps 2+ in the guide above for usage.

## Summation of Learning Experience

* Describe how your previous experience (academic or professional) readied you for this project.
* Describe any additional learning or resources needed to complete this project.
* Describe how this experience contributed to your concept of lifelong learning.

**References**

Fanaee-T, Hadi, and Gama, Joao, ‘Bike Sharing Dataset’