July 31, 2024

C964: Computer Science Capstone

## Task 2 parts A, B, C and D

[Part A: Letter of Transmittal 3](#_Toc173329764)

[Part B: Project Proposal Plan 5](#_Toc173329765)

[Project Summary 5](#_Toc173329766)

[TheProblem 5](#_Toc173329767)

[TheClient 5](#_Toc173329768)

[Deliverables 5](#_Toc173329769)

[Justification 6](#_Toc173329770)

[Data Summary 6](#_Toc173329771)

[Implementation 7](#_Toc173329772)

[Timeline 9](#_Toc173329773)

[Evaluation Plan 10](#_Toc173329774)

[Verification Methods: 10](#_Toc173329775)

[Validation Methods: 11](#_Toc173329776)

[Resources and Costs 12](#_Toc173329777)

[Part C: Application 16](#_Toc173329778)

[Application Files 16](#_Toc173329779)

[Prerequisites 16](#_Toc173329780)

[Solution Summary 17](#_Toc173329781)

[Data Summary 18](#_Toc173329782)

[Machine Learning 18](#_Toc173329783)

[Validation 20](#_Toc173329784)

[Visualizations 21](#_Toc173329785)

[Installation Guide 22](#_Toc173329786)

## Part A: Letter of Transmittal

July 31, 2024

Mr. Alexander Bennett

Senior Vice President

Frosty Forecasts Ice Cream Co.

12345 Ice Cream Lane

Chillville, CA 90210

Dear Mr. Bennett,

I am writing to present a proposal for a new machine learning project designed to address a critical challenge faced by Frosty Forecasts Ice Cream Co. Our current inability to accurately predict daily revenues based on temperature fluctuations has hindered our ability to optimize operational decisions. This project aims to develop a solution that will significantly enhance your forecasting capabilities and operational efficiency. Frosty Forecasts Ice Cream Co. currently lacks a reliable method for predicting daily revenue, which affects decision-making regarding truck operations. Without accurate predictions, the company risks either overextending resources on low-revenue days or missing potential sales on high-revenue days.

We propose developing a machine learning application using linear regression to predict daily revenue based on historical temperature data. This solution will leverage Python and its powerful libraries, including “scikit-learn” for implementing and optimizing the linear regression model. By using this approach, we will provide a data-driven tool that forecasts revenue with high accuracy. Implementing this predictive model will offer several key benefits:

1. Optimized Resource Allocation: Make informed decisions about truck operations, reduce wasted resources and improve overall efficiency.
2. Increased Revenue: Identify and capitalize on high-revenue opportunities, enhancing profitability.
3. Data-Driven Insights: Gain valuable insights from historical data, aiding strategic planning and operational adjustments.

**Cost and Timeline:**

The project is designed to be completed within one month, from August 1, 2024, to September 1, 2024. The total estimated cost for the project is $5,635. This includes:

* Hardware and Software Costs: $1,700
* Labor Costs: $3,750
* Initial Environment Costs: $185

The costs cover all necessary components, from hardware and software to labor and initial deployment. There are no significant ethical or legal concerns regarding the data, as it is sourced from a reputable educational dataset and is already cleaned and prepared for use.

As an experienced software developer with a strong background in machine learning and data analysis, I am confident in my ability to deliver a high-quality solution that meets your needs. My expertise in Python and model optimization will ensure the development of a reliable and effective forecasting tool. I am enthusiastic about the potential impact of this project on Frosty Forecasts Ice Cream Co. and am eager to discuss it further. Thank you for considering this proposal. I look forward to the opportunity to contribute to the company’s continued success.

Sincerely,

Robert N. King

# Part B: Project Proposal Plan

# Project Summary

### **The Problem**

Frosty Forecasts Ice Cream Co., an ice cream truck company, currently faces challenges in predicting daily revenue. The company lacks a reliable method to forecast profits based on influential factors such as weather conditions. This uncertainty complicates decision-making regarding truck operation schedules, leading to potential inefficiencies and missed revenue opportunities. Without accurate predictions, the company struggles to optimize truck deployment, which can impact overall profitability and operational efficiency.

### **The Client**

Frosty Forecasts Ice Cream Co. is a company operating ice cream trucks and seeking to enhance its revenue forecasting capabilities. The client needs a solution that enables accurate prediction of daily revenue to make informed decisions about truck schedules. Currently, the company does not have a system in place to integrate weather data, such as temperature, into its revenue planning process. Addressing this gap will allow the client to optimize truck operations, reduce costs, and maximize revenue by aligning truck activity with predicted profitable days.

### **Deliverables**

The primary deliverables for this project include:

1. **Predictive Application**: A machine learning application developed using Python and Jupyter Notebook. This application will employ linear regression to predict daily revenue based on temperature data. The application will include functionality for training the model, making predictions, and displaying results.

2. **User Guide**: A user guide detailing how to operate the application. This guide will cover instructions on how to install the necessary tools to run the program and provide the model with input to with which it will make predictions. It will be designed to ensure that users with minimal technical expertise can effectively utilize the application.

3. **Documentation**: Detailed technical documentation including explanations of the model development process, data preparation, and implementation details. This will also cover how the model was validated, and the performance metrics used to assess its accuracy.

4. **Performance Report**: A summary report outlining the results of the k-fold cross-validation, including metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). This will be included as part of the Jupyter Notebook project.

### **Justification**

The implementation of this machine learning application will significantly benefit Frosty Forecasts Ice Cream Co. by providing a data-driven approach to revenue prediction. With accurate forecasts based on temperature, the company will be able to make strategic decisions about truck deployment, ensuring that trucks are operational on days with higher predicted revenue. This optimization will lead to more efficient resource allocation, reduced operational costs, and increased revenue opportunities. By minimizing uncertainty and aligning truck schedules with predicted profitability, the application will enhance overall business performance and support the company's growth objectives.

# Data Summary

The dataset that will be used for this project has been obtained from Kaggle.com, where it was created specifically for educational purposes and is freely available. The raw data, provided in a well-structured CSV format, was pre-cleaned and prepared, making it directly suitable for use without further preprocessing. To support the development and evaluation of the machine learning model, the dataset was partitioned into two separate subsets: one for training the model and another for testing and validating its performance. This separation ensures that the model is evaluated on data it has not seen during training, providing a clearer picture of its predictive accuracy.

In the development phase, the data will be used as-is from the CSV files to train and test the linear regression model, ensuring that the data's quality and consistency are maintained. During maintenance, periodic checks will be conducted to confirm that the data continues to meet the project’s needs, though no modifications are anticipated unless future updates are required.

This dataset aligns well with the project’s objectives as it is appropriately cleaned and formatted, ensuring that the machine learning model can be developed and tested effectively. Additionally, since the dataset is publicly available and intended for educational use, there are no ethical or legal concerns regarding its use. The transparency and accessibility of the data eliminates potential issues related to data privacy and intellectual property, ensuring compliance with legal and ethical standards.

## Implementation

The implementation of this project will follow the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which is widely recognized in the industry for its structured approach to data mining and machine learning projects. CRISP-DM provides a framework that has six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This methodology ensures a systematic approach to project execution, from understanding the client’s needs to deploying the final solution.

**Machine Learning Solution Implementation Plan:**

**1. Data Preparation:**

* **Objective:** Prepare the dataset for machine learning tasks to ensure it is suitable for model training and evaluation.
* **Activities:** Import the CSV file containing the dataset into the Python environment using Pandas. Split the dataset into two distinct subsets: a training dataset for model training and a test dataset for model evaluation.

**2. Model Development:**

* **Objective:** Develop and train a linear regression model to predict daily revenue based on temperature.
* **Activities:** Utilize the scikit-learn library in Python to implement the linear regression model. Initialize the LinearRegression class and fit the model using the training dataset. Implement gradient descent optimization to fine-tune the model parameters and minimize prediction errors. Conduct k-fold cross-validation to evaluate the model’s performance on different subsets of the training data, ensuring robust performance assessment.

**3. Model Evaluation:**

* **Objective:** Assess the performance of the linear regression model and ensure it meets the project’s objectives.
* **Activities:** Use the test dataset to evaluate the model’s performance. Compute performance metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to gauge the accuracy and reliability of the predictions. Analyze these metrics to determine how well the model generalizes to new, unseen data and whether it satisfies the project's success criteria.

**4. Model Deployment:**

* **Objective:** Deploy the trained model within an application for practical use and integration.
* **Activities:** Develop a Python application using the Jupyter Notebook environment to interface with the trained linear regression model. Implement features to load new temperature data, generate revenue predictions, and present the results in a user-friendly format. Ensure that the application allows for easy interaction and provides clear output to support decision-making based on model predictions.

**5. Performance Monitoring and Maintenance:**

* **Objective:** Ensure the ongoing effectiveness and accuracy of the deployed model.
* **Activities:** Monitor the model’s performance periodically to detect any potential issues or degradation over time. Provide support for any user-reported issues and adjust as needed. Perform routine updates and maintenance to ensure that the model continues to perform well with new data.

## Timeline

The project will be executed over a four-week period from August 1, 2024, to September 1, 2024. In the first week, the focus will be on data preparation, including importing and splitting the dataset. The second week will be dedicated to model development, where the linear regression model will be implemented and trained using the training dataset. During the third week, the model’s performance will be evaluated using the test dataset, with metrics such as R-squared, MAE, and RMSE computed to assess accuracy. The final week will involve deploying the model within a user-friendly application, integrating it with the prediction functionality, and performing final tests. Ongoing monitoring and maintenance will be conducted to address any issues and ensure optimal performance.

|  |  |  |
| --- | --- | --- |
| Week | Dates | Activities |
| 1 | 08/01/2024 - 08/07/2024 | Data preparation: Import, format, and split the dataset. |
| 2 | 08/08/2024 - 08/14/2024 | Model development: Implement and train the linear regression model. |
| 3 | 08/15/2024 - 08/21/2024 | Model evaluation: Assess performance with metrics (R-squared, MAE, RMSE). |
| 4 | 08/22/2024 - 09/01/2024 | Deployment: Develop application, integrate model, and perform final testing. |

# Evaluation Plan

## Verification Methods:

**1. Data Preparation:**

**Verification Method:** During the data preparation phase, verification will involve a series of integrity checks to confirm the accuracy and completeness of the dataset. This includes validating that the data has been correctly imported from the CSV file. Additionally, the data splitting process will be reviewed to ensure that the training and test datasets are correctly partitioned without overlap.

**2. Model Development:**

**Verification Method:** In the model development stage, verification will focus on ensuring the proper implementation and training of the linear regression model. This will be achieved through unit testing of individual components of the model, such as data handling functions and training algorithms. Model training will be monitored to confirm that the gradient descent optimization is functioning correctly and that the model parameters are being adjusted as expected. Cross-checks will be performed against theoretical results and expected outcomes for small test cases.

**3. Model Evaluation:**

**Verification Method:** During model evaluation, the performance metrics (R-squared, MAE, RMSE) will be verified by comparing them to known benchmarks or expected ranges. The calculations for these metrics will be verified against manually computed values or using alternative implementations to ensure correctness. Consistency checks will be performed to validate that the evaluation metrics are accurately computed across different test subsets.

**4. Model Deployment:**

**Verification Method:** Verification in the deployment phase will involve integration testing to ensure that the model functions correctly within the application environment. This includes testing the model’s interaction with the user interface and its ability to process new data inputs accurately. Functional testing will be carried out to confirm that the application performs all intended tasks without errors.

## Validation Methods:

**Validation Method:** Upon completion of the project, the overall effectiveness and accuracy of the machine learning solution will be validated through a final assessment involving multiple steps:

**Performance Validation:** A final round of comprehensive validation will be conducted using k-fold cross-validation to assess the model’s performance. This will involve evaluating the model’s predictive accuracy on different data subsets and confirming that the performance metrics (R-squared, MAE, RMSE) consistently meet the project’s success criteria. This step ensures that the model generalizes well and performs reliably on unseen data.

**User Acceptance Testing (UAT):** The application will undergo User Acceptance Testing to validate that it meets the client's requirements and operates as intended in a real-world environment. Feedback will be collected from end-users to assess the application’s usability, effectiveness, and alignment with user needs and expectations.

**Documentation Review:** The project documentation, including the user guide and technical specifications, will be reviewed for completeness and accuracy. This review will ensure that all instructions are clear, comprehensive, and useful for both end-users and maintenance personnel. The documentation will be validated to confirm it adequately supports the application’s use and ongoing support.

## Resources and Costs

The outline below describes the associated resources and costs for the solution. The project benefits from leveraging open-source software and cloud-based solutions, which significantly reduce both initial and ongoing costs. The efficient use of existing tools and technology allows for a cost-effective solution while ensuring high-quality outcomes.

**Hardware and Software Costs:**

* Hardware Costs:
  + Development Workstation: $1,500 - A high-performance computer equipped with sufficient memory and processing power to handle data processing and model training tasks.
  + Backup Storage: $200 - External or cloud storage solutions for backing up project data and ensuring data integrity.
* Software Costs:
  + Python Development Environment: $0 - Python and other libraries are open-source and free to use.
  + Jupyter Notebook: $0 - Jupyter Notebook is also open-source and free.

**Estimated Labor Time and Costs:**

* Model Development:
  + Time Estimate: 15 hours
  + Labor Cost: $1,125 - Based on an average rate of $75 per hour for a machine learning engineer.
* Model Evaluation:
  + Time Estimate: 12 hours
  + Labor Cost: $900 - Based on an average rate of $75 per hour for a data scientist or machine learning engineer.
* Model Deployment:
  + Time Estimate: 15 hours
  + Labor Cost: $1,125 - Based on an average rate of $75 per hour for a software developer or deployment specialist.
* Documentation and Training:
  + Time Estimate: 8 hours
  + Labor Cost: $600 - Based on an average rate of $75 per hour for technical writers or trainers.
* Total Estimated Labor Time: 50 hours
* Total Labor Cost: $3,750

**Estimated Environment Costs:**

* Deployment:
* Cloud Hosting: $50 - Monthly cost for hosting the application on a cloud platform such as AWS, Google Cloud, or Azure. This includes the cost of virtual machines or managed services for running the application.
* Maintenance:
* Ongoing Maintenance: $100 per month - Includes updates, bug fixes, and minor enhancements post-deployment. This cost may vary based on the level of ongoing support required.
* Additional Environment Costs:
* Data Storage: $20 per month - Cost for storing data used and generated by the application on a cloud storage solution.
* Backup Services: $15 per month - Cost for additional backup services to ensure data safety and integrity.
* Total Estimated Environment Costs:
* Initial Deployment Cost: $50
* Monthly Maintenance and Operational Costs: $135
* Total Cost for Initial Month: $185
* Estimated Monthly Cost for Subsequent Months: $135
* **Total Initial Project Cost: $5,635**

# Part C: Application

# Application Files

.\cd964-capstone

\IceCreamRevenuePrediction.ipynb Jupyter notebook file; source code

\ ice\_cream\_sales\_data.csv Training dataset

\ test\_data.csv Testing dataset

# Prerequisites

Anaconda - <https://www.anaconda.com/download>

**Part D: Post-implementation Report**

## Solution Summary

Frosty Forecasts Ice Cream Co faces a significant challenge in predicting daily revenue, which is essential for making informed decisions about truck operation schedules. Currently, the company lacks an effective system for forecasting profits. This uncertainty complicates planning and may lead to inefficient truck deployment, resulting in missed revenue opportunities.

To address this problem, a machine learning project will be developed that will use linear regression to predict daily revenue based on temperature. Historical data on daily revenue and corresponding temperature has already been collected and prepared for this use. The core of the solution involves developing a linear regression model to capture the relationship between temperature and revenue. Gradient descent will be employed to optimize the model parameters and minimize the error between predicted and actual revenue values.

The development will take place using Python in the Jupyter Notebook environment, which offers an interactive platform for coding, visualization, and documentation. The model will be trained on historical data to understand the relationship between temperature and revenue and validated with a separate dataset to assess its predictive accuracy. The model's performance will be evaluated to ensure reliability.

Once validated, the model will be deployed to forecast daily revenue based on current or forecasted temperatures. This forecasting tool will enable Frosty Forecasts Ice Cream Co. to make strategic decisions about truck operation, optimizing schedules to maximize revenue and improve operational efficiency. By implementing this solution, the company will gain a data-driven approach to revenue prediction, enhancing planning accuracy and resource allocation, ultimately leading to increased profitability.

## Data Summary

Link to dataset: <https://www.kaggle.com/datasets/vinicius150987/ice-cream-revenue>

The data used for the project was sourced from a dataset created for educational purposes and posted on Kaggle.com. This dataset was already cleaned and prepared, provided in the form of a CSV file. To facilitate the development and evaluation of the machine learning model, the data was divided into two distinct sets: the training dataset, used to train the model, and the test dataset, employed to test the model and assess its performance. After splitting the data into these two separate CSV files, the content remained unchanged and was directly loaded into the application for processing by the linear regression model. This approach ensured that the data utilized in the model development was consistent and appropriately handled.

## Machine Learning

For this project, linear regression was chosen as the machine learning method to predict daily revenue based on temperature. Linear regression is a statistical technique that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. In this case, it aims to establish a correlation between temperature (independent variable) and revenue (dependent variable). The method was employed by importing the “scikit-learn” library and utilizing its LinearRegression class, which was fed the independent and dependent variables loaded from the CSV files for training and testing the model.

The following screenshots demonstrate the implementation of Linear Regression:

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A computer screen shot of a code

Description automatically generated

The model was developed by applying gradient descent, an optimization algorithm that adjusts the model parameters to minimize the difference between predicted and actual revenue values. This approach was selected due to its simplicity and effectiveness in capturing linear relationships, making it well-suited for predicting revenue based on a single variable, temperature. Linear regression's straightforward nature also facilitates interpretability and ease of implementation, aligning well with the project's requirements for accuracy and operational efficiency.

The following screenshot demonstrates the implementation of Gradient Descent:

A screenshot of a computer program

Description automatically generated

## Validation

In this project, k-fold cross-validation was employed to validate the performance of the linear regression model. The dataset was divided into five equally sized folds using the KFold method from the scikit-learn library. For each of the five iterations, the model was trained on four of the folds and tested on the remaining fold. This process was repeated such that each fold served as the validation set once. Performance metrics, including R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), were computed for each iteration to evaluate how well the model generalized to unseen data. By averaging these metrics across all folds, a robust estimate of the model’s predictive accuracy and reliability was obtained, minimizing the risk of overfitting and ensuring a more comprehensive assessment of its performance.

The following screenshot demonstrates the implementation of k-fold cross-validation:

A screenshot of a computer program

Description automatically generated

## Visualizations

Visualizations for this project include summary statistic charts, scatterplots, regression lines, and residual plots. Below are some screenshots of each:

**Summary Statistic Charts:**

A screenshot of a computer

Description automatically generated

**Scatterplots:**

A screenshot of a computer screen

Description automatically generated

**Regression Line:**

A graph with a red line

Description automatically generated

**Residual Plot:**

A screen shot of a graph

Description automatically generated

## Installation Guide

1. Extract contents of "cd964-capstone.zip" to desired directory

2. Navigate to <https://www.anaconda.com/download>

3. Enter your email address

4. Anaconda will email you installation instructions

5. Install Anaconda

6. Open Anaconda

7. Open Jupyter Notebook and navigate to directory chosen in step 1

8. Open the file "IceCreamRevenuePrediction.ipynb"

9. Read through the content as desired to learn about the prediction model

10. In the menu bar, select "Run", then "Run all"

11. Enter a temperature with which the model will make a prediction and press enter