March 01st, 2024

Predicting Customer Purchases

C964: Computer Science Capstone

Task 2 parts A, B, C and D

[Part A: Letter of Transmittal 1](#_Toc98598250)

[Letter of Transmittal 2](#_Toc1738085866)

[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

03/01/2024

Annabeth Chigings

Marketing Solutions™️

2309 49th St, Unit 11

San Diego, CA 92930

United States of America

Dear Annabeth Chigings,

I am writing to present a proposal for implementing a predictive analytics solution aimed at optimizing our organization's marketing strategy. As senior leadership, I believe you will find this proposal both timely and beneficial to our overall objectives.

In today's competitive market, understanding customer behavior is paramount to success. However, identifying potential customers who are likely to make a purchase remains a challenge. Traditional marketing approaches often lack precision and may result in inefficient resource allocation and missed opportunities.

Our proposed solution involves developing a predictive analytics application that leverages machine learning techniques to forecast customer purchase likelihood. By analyzing demographic traits such as gender, age, and salary, the application will provide valuable insights into customer behavior, allowing us to target potential customers more effectively.

Implementing this solution will lead to several practical benefits for our organization:

1. Improved Marketing Effectiveness: By accurately predicting customer purchase likelihood, we can tailor our marketing efforts to target individuals who are most likely to make a purchase, thereby increasing conversion rates and maximizing ROI.
2. Resource Optimization: With better insights into customer behavior, we can allocate marketing resources more efficiently, reducing costs and improving overall profitability.
3. Competitive Advantage: By staying ahead of market trends and understanding our customers' needs, we can maintain a competitive edge in our industry.

The implementation plan involves several key steps:

1. Data Collection and Preparation: Gather relevant customer data and preprocess it for analysis.
2. Model Development: Build and train machine learning models to predict purchase likelihood based on demographic traits.
3. Application Development: Develop a user-friendly application interface for inputting customer data and generating predictions.
4. Testing and Validation: Validate the accuracy and reliability of the predictive models through rigorous testing and evaluation.
5. Deployment: Roll out the application for organizational use, ensuring seamless integration and user adoption.

The estimated costs include personnel time, software licensing, and potential hardware expenses. The projected timeline spans several weeks, with each milestone carefully planned and executed. The data used for training and testing the models will be collected from customer surveys, ensuring compliance with data privacy regulations and ethical standards.

As a Systems Integration Engineer, I bring 15 years of experience in machine learning, including expertise in data analytics. Our team is well-equipped to handle the complexities of this project and deliver tangible results that align with our organizational objectives.

In conclusion, I am confident that implementing this predictive analytics solution will position our organization for long-term success and growth. I welcome the opportunity to discuss this proposal further and address any questions or concerns you may have.

Thank you for your consideration.

Sincerely,

Samuel Diaz

Samuel Diaz, Systems Integration Software Engineer

# Part B: Project Proposal Plan

## Project Summary

**Describe the Problem**

We aim to develop a predictive analytics solution to assist our client in targeting potential customers more effectively based on demographic traits. This solution will predict whether a customer will purchase a product, allowing for tailored marketing strategies.

**Client’s Needs**

Our client is a marketing company aiming to optimize their sales strategy by targeting potential customers more effectively. Their primary need is a predictive model that can accurately forecast whether a customer will make a purchase, allowing them to tailor their marketing efforts accordingly.

**Deliverables**

1. Finished application: A machine learning model integrated into an application interface for easy input of customer traits and quick prediction of purchase likelihood.
2. User guide: A comprehensive guide explaining how to use the application effectively, interpret results, and troubleshoot any issues that may arise.

**Justification of Application Benefits**

The application will benefit the client by providing them with actionable insights into customer behavior. By accurately predicting purchase likelihood, the client can allocate resources more efficiently, leading to increased sales and higher ROI on marketing efforts.

## Data Summary

**Source of Raw Data**

The raw data will be collected from customer surveys conducted by the client. It includes demographic information such as gender, age, salary, and whether the customer made the purchase.

**Data Processing and Management**

During the application development life cycle, the data will be processed by cleaning, encoding, categorical variables, and splitting into training and testing sets. During maintenance, the model may be retrained periodically using new data to ensure its accuracy.

**Justification of Data**

The data meets the project's needs as it contains relevant demographic information and purchase behavior. Any anomalies, such as outliers or incomplete data, will be handled through data preprocessing techniques like outlier detection and imputation.

**Ethical and Legal Concerns**

There are no ethical or legal concerns regarding the data, as it is collected with consent and anonymized to ensure privacy.

## Implementation

**Industry Standard Methodology**

CRISP-DM (Cross-Industry Standard Process for Data Mining) is the industry-standard methodology that will be used for the implementation of the project. This methodology is commonly used in data science and machine learning projects as it is a comprehensive framework that guides the entire lifecycle of a data mining project, from understanding the business problems to the deployment and maintenance of the solution.

**Project’s Implementation Plan**

1. Data collection and preprocessing: Gathering raw data and preparing it for analysis.
2. Model development: Building and training the predictive model using machine learning algorithms.
3. Application development: Creating the user interface for the predictive model.
4. Testing and validation: Ensuring the application functions correctly and provides accurate predictions.
5. Deployment: Releasing the application for client use.

## 

## Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration | Start Date | End Date |
| Data collection and preprocessing | 2 weeks | April 15, 2024 | April 29, 2024 |
| Model Development | 3 weeks | May 1, 2024 | May 22, 2024 |
| Application Development | 2 weeks | May 23, 2024 | June 5, 2024 |
| Testing and Validation | 1 week | June 6, 2024 | June 12, 2024 |
| Deployment | 1 week | June 13, 2024 | June 19, 2024 |

## Evaluation Plan

**Verification methods**

Verification will be conducted through code review, unit testing, and validation against benchmarks to ensure each stage of development meets the specified requirements.

**Validation Methods**

Upon completion of the project, the model’s performance will be evaluated using metrics such as accuracy, precision, recall, and F1 score to validate its effectiveness in predicting purchase likelihood.

## 

## Resources and Costs

**Hardware and Software Costs**

* $1,000 - Hardware: Standard laptop or desktop computer for development and testing
* $0 – Python programming language, Juptyer Notebook for development, scikit-learn library for machine learning, Flask for web application development

**Estimated Labor Time and Costs**

* $4,000 - Data Scientist: 100 hours
* $3,690 - Software Engineer: 90 hours
* $2,000 - UI/UX Designer: 40 hours
* $2,200 - Project Manager: 40 hours

**Estimated environment costs**

* $400 - Deployment: Minimal cost for web hosting and maintenance.

# 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.

(see attached pythonProject.zip)

# Part D: Post-implementation Report

## Solution Summary

**Problem/Solution**

The problem addressed was predicting whether a potential customer would purchase a product based on demographic traits. The solution involved developing a predictive analytics application using machine learning techniques. This application assists in targeting potential customers more effectively, thereby optimizing marketing strategies.

How Application Solves the Problem in Parts A & B

The application takes input on customer traits such as gender, age, and salary, and predicts whether the customer will make a purchase based on purchases made by previous customers. By leveraging machine learning models trained on historical data, the application provides valuable insights into customer behavior, enabling the client to allocate resources efficiently and increase sales.

## Data Summary

**Source of Raw Data**

The raw data was collected from customer surveys conducted by the client.

**Data Processing**

Throughout the application development lifecycle, data was processed by cleaning, encoding categorical variables, and splitting into training and testing sets. During maintenance, the model was retrained periodically using new data to ensure accuracy.

## Machine Learning

**Method: Logistic Regression**

* What: Logistic regression is a classification algorithm used to model the probability of a binary outcome based on one or more predictor variables.
* How: The logistic regression model was developed by fitting the training data to the logistic regression algorithm using the scikit-learn library in Python.
* Why: Logistic regression was chosen due to its simplicity, efficiency, and interpretability, making it suitable for binary classification tasks like predicting purchase likelihood.

## Validation

**Method: Cross-Validation**

Validation Method: k-fold cross-validation was employed to validate the logistic regression model.

Results: Cross-validation results indicated a mean accuracy of 85% on the test data, demonstrating the model's effectiveness in predicting purchase likelihood.

## Visualizations

*The following three visualizations were created using matplotlib. They were omitted from the code to prevent cluttering.*

**ROC Curve**

This curve plots the true positive rate against the false positive rate for different thresholds.

A screen shot of a graph

Description automatically generated

**Distribution of Predicted Probabilities**

This plot visualizes the distribution of predicted probabilities for the purchase of the item.

A screen shot of a graph

Description automatically generated

**Feature Importance Plot**

This plot showcases the importance of each feature in predicting purchase likelihood based on the logistic regression coefficients.

A screenshot of a computer

Description automatically generated

User Guide

1. Download and Install Software/Libraries
   1. Download Python 3.11 from <https://www.python.org/downloads/windows/>
      1. Select the package to download based on your computer architecture
   2. Download PyCharm CE (Community Edition) at <https://www.jetbrains.com/pycharm/download/?section=mac>
      1. Note: scroll to the bottom of the page until you see ‘PyCHarm Community Edition’ and click on ‘Download’
   3. Install the application onto your computer and make sure to integrate the Python version you just downloaded when prompted by PyCharm
   4. Under projects, click on ‘open’ and select the pythonProject directory provided.
   5. Make sure to install the pandas and sklearn modules or else you will get an error
      1. Go to File > Project Structure > Platform Settings > SDKs
      2. Go to your Python version listed, choose the packages tab
      3. Click on the + button, search for ‘pandas’, select the option matching the name, and click ‘install package’
      4. Click on the + button, search for ‘scikit-learn’, select the option matching the name, and click ‘install package’
   6. Alternatively: hover over the red, squiggly line shown under the names of the import modules giving the error (first few lines of code). Click on the red light bulb button that appears and select the option of “Install package (name)”
      1. Example 1: Hover over the red, squiggly line shown under the import pandas (first line of code). Click on the red light bulb button that appears and select the option of “Install package pandas”
      2. Example 2: Hover over the red, squiggly line shown under the import sklearn.model\_selection (second line of code). Click on the red light bulb button that appears and select the option of “Install package scikit learn”
      3. Do the above for all import modules.
2. Usage Instructions
   1. Run the Python application script.
   2. In the command line, enter customer traits (gender, age, salary) when prompted.
   3. Press Enter after each prompt.
   4. Receive a message on whether the new customer will either purchase or not purchase the item.
3. Example Usage
   1. Input gender as “Male”, age as 30, and salary as 70000.
   2. Press ‘Enter’ after each prompt.
   3. The application will display the following message: “Prediction for new customer: Not Purchased”

# Reference Page

Not applicable.