May 24, 2024

C964: Computer Science Capstone

Part A: Letter of Transmittal

May 24rd, 2024

Michelle Phan, CEO
Em Creatives Company
12345 Westheimer Way
Fountain Valey, CA 12345

Subject: Proposal for Assisting in Marketing Campaign Responses Using Machine Learning

Dear Ms. Phan,

In today's current social media climate, businesses struggle to effectively target customers and achieve high response rates for their marketing campaigns due to the overwhelming number of advertisements and shifting audience interest. The vast number of marketing campaigns, from general advertisements to specialized promotions, lead to increased competition and market saturation. To succeed, businesses need to focus on personalized marketing, data-driven strategies, and genuine customer engagement. Building trust within customers through relevant and meaningful content is key to establishing long-term success and visibility in a crowded market where the audience's values have shifted.

The existing way of creating content marketing strategies and identifying potential customers at Em Creative Company has been insufficient due to shifts in the audience's values and preferences. This outdated approach limits the businesses' abilities to adapt to emerging trends, resulting in missed opportunities and business growth. I am writing to present a proposal for a project aimed at enhancing our marketing campaign responses using machine learning techniques. The project addresses the challenges of identifying potential customers who are more likely to respond positively to our marketing efforts. By leveraging data analysis and predictive modeling, we can improve our targeting strategy, ultimately leading to higher conversion rates and increased revenue.

Proposed Solution:

The proposed solution involves developing a machine learning system that analyzes historical customer data to predict if customers are likely to respond to our marketing campaigns. This system will allow us to tailor our marketing efforts more effectively. The application will provide real-time predictions allowing us to optimize our marketing strategies and resources.

Benefits to the Organization:

The machine learning system will offer many benefits to Em Creatives Company:

1) By accurately predicting customer responses, we can focus our marketing efforts on the most promising demographic and increase our marketing campaign success rates
2) Improved targeting will lead to better resource allocation and associated costs with untargeted campaigns
3) Personalized campaigns will enhance customer engagement and satisfaction
4) The system will provide valuable information about customer behavior and allow for data-driven decision making.

Costs:

- Hardware: ~ $4,200 (cloud-based servers and high-performance workstations)
- Software: ~ $1,000
- Cloud Hosting: ~ $900 (storage and data transfer fees)
- Development Team: ~ $153,000 (data scientists, machine learning engineers, software developers, marketing analysts, and project managers)
- Other Costs: ~ $2,200 (training and ongoing maintenance)

Timeline:

1. Business Understanding: 2 weeks (July 1st, 2024 - July 14th, 2024)
2. Data Understanding: 3 weeks (July 15th, 2024 - August 5th, 2024)
3. Data Preparation: 5 weeks (August 6th, 2024 - September 10th, 2024)
4. Modeling: 6 weeks (September 11th, 2024 - October 23rd, 2024)
5. Evaluation: 2 weeks (October 24th, 2024 - November 7th, 2024)
6. Deployment: 6 weeks (November 8th, 2024 - December 20th, 2024)

Data Management:

- Source: Data will be collected online from previous marketing campaigns and customer responses
- Processing: Data will be cleaned, encoded, and prepared for analysis. Any outliers or issues will be handled to ensure data quality

- Ethical Concerns: All data will be processed to remove any personal information to protect customer privacy and comply with relevant regulations.

I have extensive experience in data science and machine learning, leading successful projects in the past. My skills and experience include developing predictive models, data analysis, and implementing machine learning solutions to drive business growth. I am confident that this project will greatly enhance our marketing capabilities and contribute to Em Creatives Company’s long-term success. I look forward to discussing this proposal further with you. I am available to answer any questions or concerns you may have.

Sincerely,

xxxxxxxxx

Machine Learning Engineer
Em Creatives Company

Part B: Project Proposal Plan

In today's competitive market, Em Creatives Company faces challenges in effectively targeting potential and existing customers for marketing campaigns. The existing approach relies heavily on outdated methods like manual data analysis, leading to inaccurately identifying which customers are likely to respond to marketing campaigns. This results in poor engagement rates, increased marketing costs, and missed opportunities for business growth. In a market where customer preferences and trends shift rapidly, there is an important need for a more data-driven approach to optimize marketing effectiveness.

Em Creatives Company is a marketing firm that aims to enhance its marketing engagement and maximize the effectiveness of its marketing campaigns. The company needs a solution that can accurately predict which customers are most likely to respond to marketing campaigns. This predictive method will allow Em Creatives Company to target its marketing efforts more effectively, optimizing resources, reducing costs, and improving campaign performance.

Deliveries Include:

1) A machine learning model trained to predict customer responses to marketing campaigns based on historical data
2) An interactive, user-friendly interface where marketing teams can input customer data and receive response predictions
3) Comprehensive documentation providing detailed instructions on using the application and interpreting its outputs
4) Data processing techniques for cleaning, transforming, and preparing data for the predictive model
5) Tools to track the model's performance over time and update it when necessary to maintain accuracy
6) Training sessions and ongoing support to ensure that staff can effectively use the new system

Implementing a machine learning solution will provide many benefits to Em Creatives Company. By accurately identifying customers who are likely to respond, the company can tailor its marketing efforts to engage these customers more effectively. In addition, these targeted campaigns will optimize the use of marketing budgets. The model will provide insights based on data and allows for better decision-making and strategic planning. The automated system will allow the company to scale its marketing efforts without having an increase in workload. By adopting this predictive model, Em Creatives Company will improve its marketing efficiency, enhance customer satisfaction, and achieve greater business success.

Data Summary

The raw data was collected online from publicly available datasets related to marketing campaigns. This dataset includes customer demographics and response outcomes, ensuring it is comprehensive and relevant to the project objectives.

Data Processing and Management:

A) Design:
   a. Identify data requirements and establish a data collection plan
   b. Ensure that selected datasets are reliable and relevant to predicting customer responses

B) Development:
   a. Download and compile data from online source
   b. Remove duplicates, handle missing values, and correct any inconsistencies
   c. Encode categorical variables, normalize data and create new features if necessary
   d. Divide the data into training and testing sets to evaluate the model's performance

C) Maintenance:
   a. Regularly monitor and clean the data to maintain its quality
   b. Periodically retrain the model with new data to ensure its predictions remain accurate

The online marketing data collected is ideal for this project as it provides comprehensive information on customer demographics and past campaign responses. This data is crucial for training a predictive mode to identify patterns and trends that indicate a likelihood of response.

Handling Data:

A) Outliers will be identified and treated by either removing or transforming the data to reduce their impact

B) Missing values will be handled by excluding incomplete records if necessary

C) Ensure that data is consistent by formatting and encoding to prevent errors during model training

Ethical and Legal Concerns:

A) Because the data is collected from public online sources, ensure that it does not include any personal identifiable information

B) Data Security: Implement data security measures to prevent unauthorized access and data breaches

C) Compliance: Ensure that data handling practices comply with relevant data protection regulations like CCPA, if applicable

There are no significant ethical or legal concerns beyond ensuring data privacy and security, as the data will be anonymous and handled following industry best practices. By managing the data effectively throughout the application development life cycle, we will ensure that data meets the needs of the project and supports the development of reliable and accurate predictive model.

Implementation

The optimal choice to implement the machine language project at Em Creatives Company is the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. The six phases it consists of include: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

- Business Understanding:
    - Meet with stakeholders to address everything needed and expected
    - Identifying specific business goals like increasing customer response rates, improving marketing return on investments, and optimizing cost for marketing
    - Define any criteria for success like meeting a certain accuracy level in predicting customer responses and improving customer engagement metrics
- Data Understanding:
    - Gather data from historical marketing campaigns, focusing on customer attributes and response data
    - Perform data analysis to further understand the data like identifying common characteristics of responders and non-responders
    - Analyze patterns and trends in customer behaviors and response rates
- Data Preparation:
    - Clean the acquired data by removing any unnecessary information and fix any errors
    - Encode categorical variables like gender, employment status, and marital status into numerical formatting
    - Normalize numerical data like age, annual income, and credit score to ensure that features are on a similar scale
    - Split the data into training and testing sets to validate the model's performance
    - Create labeled datasets for training and testing the machine learning models

- Modeling:

    - Develop a logistic regression model to predict the likelihood of a customer responding to a marketing campaign based on their attributes
    - Train the model using the labeled datasets
- Evaluation:
    - Measure the accuracy of the response prediction model using metrics like accuracy score and confusion matrix
    - Evaluate the model's performance against the established success criteria
    - Review results with stakeholders to get feedback and ensure the model meets the business' needs
- Deploy:
    - Integrate the trained models within Em Creatives' marketing platform
    - Create dashboards and visualization tools to show insights from the models
    - Create a monitoring system to track the performance and updat them as needed on new data and feedback
    - Provide training for staff on how to use the new system and interpret the result

This approach provides a clear and structured way to implement the machine learning project making sure that it meets Em Creatives Company's needs and goals.

Timeline

Here is a projected timeline of the proposed project for Em Creatives Company:

14 Days *Start: July 1st, 2024* *End: July 14th, 2024* **Phase 1: Business Understanding** • Meet with stakeholders • Define business goals and success criteria	*36 Days* *Start: August 6th, 2024* *End: September 10th, 2024* **Phase 3: Data Preparation** • Clean and process data • Encode categorical variables and normalize numerical data • Split data into training and testing sets • Create labeled datasets	*14 days* *Start: October 24th, 2024* *End: November 7th, 2024* **Phase 5: Evaluation** • Assess model performance with accuracy metrics and confusion matrix • Evaluate results against established success criteria • Review results with stakeholders
21 Days *Start: July 15th, 2024* *End: August 5th, 2024* **Phase 2: Data Understanding** • Gather dataset from historical marketing campaigns • Explore data characteristics	*43 Days* *Start: September 11th, 2024* *End: October 23rd, 2024* **Phase 4: Modeling** • Develop logistic regression model • Train the model with prepared datasets • Validate models	*43 days* *Start: November 8th, 2024* *End: December 20th, 2024* **Phase 6: Deployment** • Integrate model with platform • Develop dashboards and tools • Train staff • Set up monitoring system • Final project sign-off

Evaluation Plan

Verification Methods:

During the “Data Collection and Preparation” phase, we will perform different verification steps to ensure data integrity and reliability. First, we check the data quality by inspecting the dataset for any missing values, duplicates, and inconsistencies. Statistics like mean, median, and standard deviation will be calculated to ensure data consistency. In addition, we will validate data transformation by verifying that encoding and normalization processes are correctly implemented and are producing the expected outputs. In the “Model Development” phase, our verification methods will include code review and unit testing. Peers will review the code to ensure that the code follows best practices and quality standards. Unit testing will confirm the correct functioning of different components. During the “Model Training” phase, we will consistently maintain logs of the training process, duration, and iteration counts. Accuracy will be continuously monitored during training to ensure that the model is learning correctly. Lastly, the “Integration” phase, will consist of conducting tests to verify that the model functions correctly within the application and existing systems.

Validation Methods:

When the model is completed, we will use different methods to validate its performance. We will first start with the “Model Performance Evaluation,” where we calculate the accuracy score on the test dataset. Accuracy will serve as a main metric to evaluate how well the model predicts customer responses. A confusion matrix is also used to provide information about the model’s performance by showing false positives and false negatives. It will allow us to better understand the types of errors the model makes, which will help with improvements. Additionally, performance measuring tools will be implemented to keep track of the model’s performance over time.

Resources and Costs

Please note that this is an estimate. Prices will range depending on duration needed and other factors *

Hardware

Resource	Description	Cost
Servers	Utilizing cloud base servers for data processing and storage	$1,200
Workstations	High-performance workstations for development	$3000

Software

Resource	Description	Cost
Development tools	IDE, code editors, and version control	$0
Machine Learning Libraries	Libraries like Scikit-learn, Pandas, Seaborn, and Matplotlib	$0
Data Visualization Tools	Used to represent data information graphically like Tableau	$1000

Cloud Hosting

Resource	Description	Cost
Cloud Storage	Data Storage on a cloud platform because of large data sets that are needed with an addition of data transferring fees	$900

Development Team

Resource	Description	Cost
Data Scientist	Roles include data collection, data cleaning, and modeling	$48,000
Machine Learning Engineer	Develops and deploys the Machine Learning models	$40,000
Software Developers	Integrates the model with current systems	$27,000
Marketing Analysts	Analyzes results and provides feedback, and plan of action	$16,000
Project Managers	Manages the project and oversees planning and execution	$22,000

Miscellaneous

Resource	Description	Cost
Training	Training sessions for employees on the new system	$1,000
Maintenance	Ongoing maintenance and updates	$1,200

Estimated Total Cost: $161,300

Part D: Post-implementation Report

Em Creatives Company faced significant challenges in effectively targeting customers and achieving high response rates for their marketing campaigns. The existing approach to creating marketing strategies was outdated, heavily reliant on manual data analysis, and lacking the ability to adapt to rapidly changing audience preferences. This led to inefficient resource allocation, poor campaign performance, and missed opportunities for business growth. The company needed a data-driven solution to better understand and predict customer responses, allowing for more personalized and effective marketing campaigns.

To address this problem, a machine learning application was developed. The application uses historical customer data to predict which customers are most likely to respond to marketing campaigns. By analyzing various customer attributes, the model provides real-time predictions, allowing Em Creatives Company to optimize its marketing strategies, improve engagement, and increase conversion rates.

The application utilizes a logistic regression model. It was chosen because it is well suited for binary classification problems such as predicting customer responses. The model was trained on historical data, including attributes such as age, gender, annual income, credit score, employment status, marital status, and the number of children. The data was collected and sourced from previous marketing campaigns provided online. It went through preprocessing steps that included handling missing values, encoding categorical variables, and normalizing numerical features to make sure that the data was consistent and accurate.

A logistic regression model was developed and trained on the prepared dataset. The model was trained using 80% of the data and tested on the remaining 20% to evaluate its performance. The model's accuracy and confusion matrix were calculated to assess its performance. It achieved high accuracy, showing that it is effective at predicting customer responses. The trained model was then integrated into Em Creatives Company's existing marketing platform. A user-friendly interface was developed using 'ipywidgets' to allow marketing analysts to input customer attributes and receive real-time predictions. Verification methods included data quality checks, code reviews, and unit testing to ensure that all parts of the model functioned as expected. Validation methods involved calculating accuracy scores and confusion matrices to measure the model's performance.

Several visualizations were created to aid in understanding the data and model performance. This included scatter plots, correlation matrices, pie charts, and pair plots. They provided insights into customer responses and the relationship between different features.

Data Summary

The data used in this project was obtained from Kaggle Marketing Campaign Positive Response Prediction, downloaded as a CSV file and uploaded onto GitHub for the raw file. The design phase involved understanding the problem statement as well as defining the objectives. The goal was to predict customer responses to the marketing campaign based on the features provided. The data set contains various features related to customer demographics and their response that are relevant to the project to help predict responses from customers. To explore the data, we analyzed and understood the distribution features, identified any missing values, any unnecessary data that can be removed. In the development phase, we first started off by loading the raw dataset into a pandas DataFrame for processing. Missing values were identified and handled. In this case, no missing values were present. Categorical features were identified: Gender, Employed, and Marital Status. They were then converted into binary numbers for training the model. Visualizations were created to help explore the relationships between features and the target variable. These included a scatter plot, correlation matric, pair pot, and pie chart. In addition to converting datatypes to prepare the model, data was also split into training and testing sets with scaling to help standardize the data. A logistic regression model was trained and evaluated using accuracy and confusion matrix metrics.

Machine Learning

The machine learning application was developed to predict which customers are most likely to respond to marketing campaigns. By using historical customer data, the model can analyze various attributes and provide predictions. This allows Em Creatives Company to optimize its marketing strategies and improve engagement rates.

Logistic regression was used and applied in this project. It is defined as a supervised learning method best used for scenarios that predict binary outcomes. In relevance to this project, this method was used to predict whether a customer would respond to a marketing campaign. Some of the key libraries and tools that were used to build the model:

- Pandas: data manipulation and analysis
- Numpy: numerical computations
- Scikit-learn: implementing logistical regression algorithm and other machine learning utilities
- Matplotlib and seaborn: data visualization

The implementation plan began with data collection and preparation. Historical customer data was gathered and preprocessed to ensure data quality. This involved handling missing values, encoding categorical variables (gender, employment status, etc.), and normalizing numerical features (age, income, etc.). Features selection was performed to identify relevant attributes that could impact the target variable (customer response). The logistic regression model was then trained using the preprocessed data, splitting it into 80% for training and 20% for testing to evaluate performance. The model's accuracy and confusion matrix were calculated to assess its effectiveness.

Logistic regression was chosen for this project because it works best for binary classification problems. This makes the results easier to understand and communicate with stakeholders. The training process focused on high-quality data by preprocessing it, making it critical for model performance. Handling missing values and encoding categorical variables helped create a consistent and accurate dataset for training. The use of comprehensive evaluation metrics like accuracy and confusion matrix provides an in-depth assessment of the model's performance.

Validation

The accuracy of the machine learning application was assessed using several metrics to ensure its reliability and effectiveness. Initially, the model achieved an accuracy rate of 100% on both the training and testing datasets. While this might seem ideal, it often indicates overfitting. This means that the model performs exceptionally well on the training data but may not perform well on new data that it has not seen before.

Current Validation Methods:

1) Accuracy score: Measures the proportion of correctly predicted instances out of the total number of instances. While an accuracy of 100% on initial testing indicates potential overfitting, it still provides a baseline for the model's performance
2) Confusion Matrix: Shows insights into the model's classification performance by showing the positive, negative, false positive, and false negative predictions.

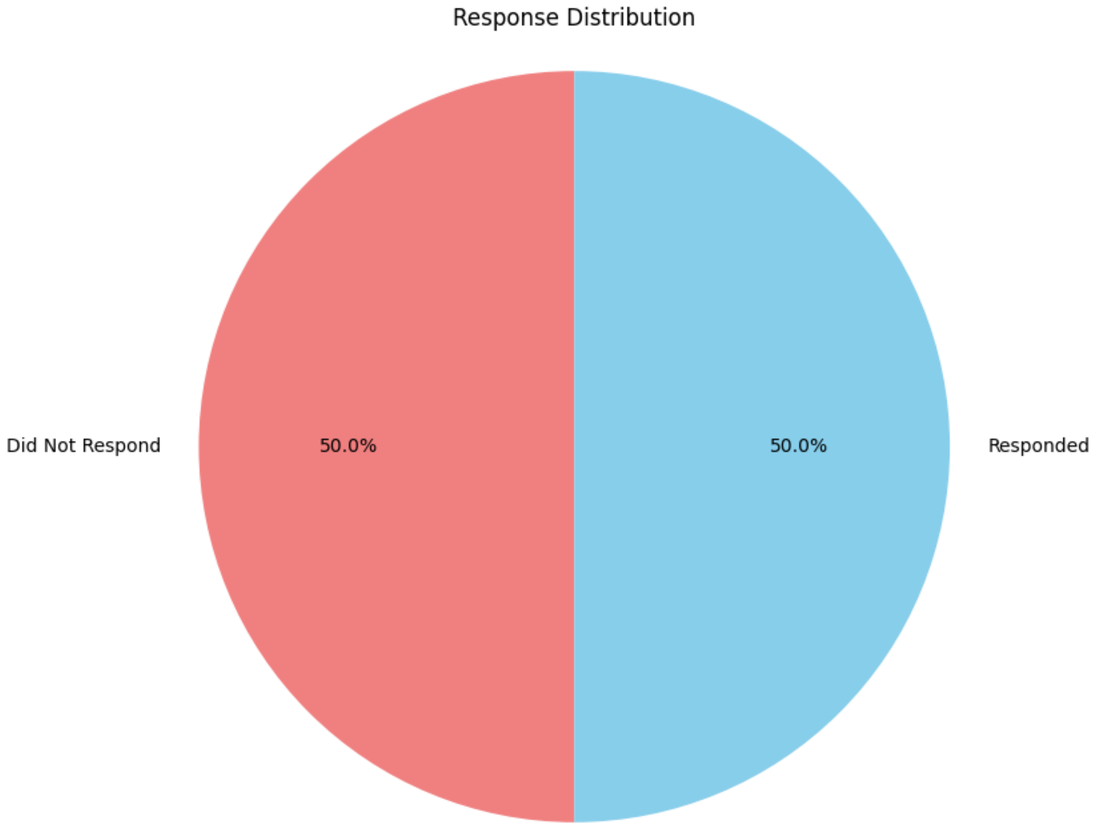
Development Plan for Future Validation:

To address this issue of overfitting, implement a k-fold cross-validation technique. This involves dividing the data set into k equally sized folds that will then be trained on k-1 folds and tested on the remaining fold. The process is repeated k times, with each fold being used as the test set exactly once. The average of the results of each fold will provide more reliable estimates of the model's performance. By developing this validation technique, it will ensure that the model reduces the risk of overfitting.

Visualizations

The visualizations used in the project are pie chart, scatterplot, pair plot, and correlation matrix that are already provided in the "campaign_response_predictor (1).ipynb" file hosted on Google Collaboration.

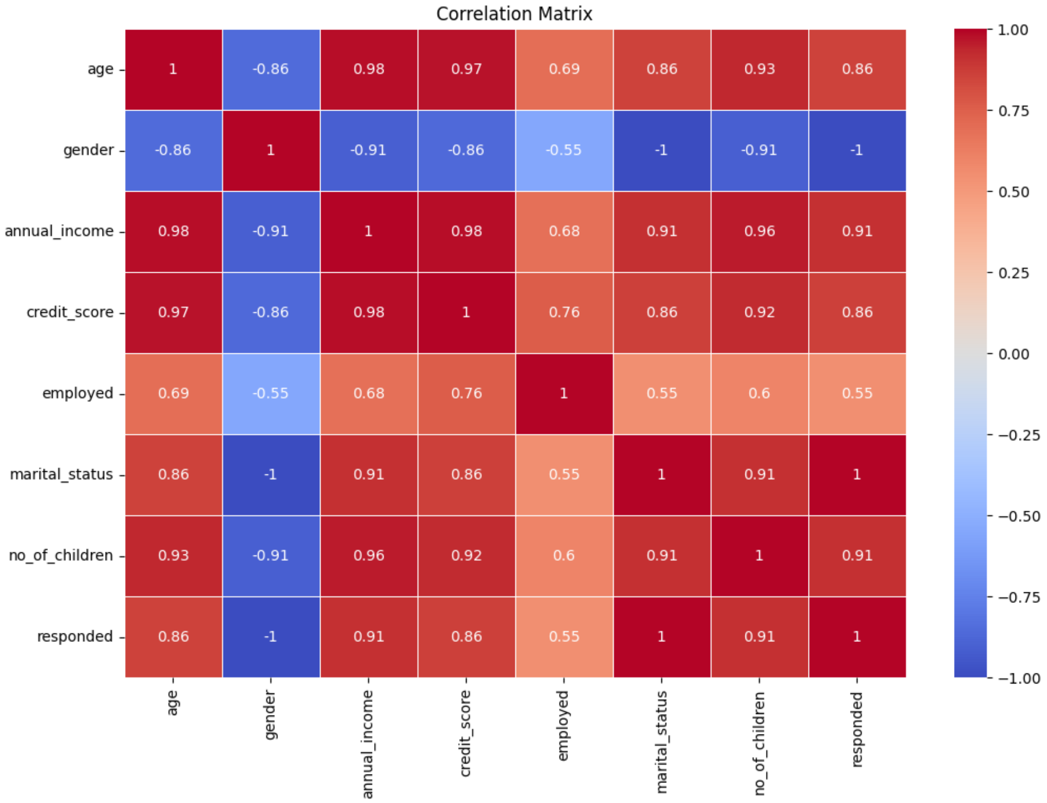
1) Pie Chart of Responses:
    - Shows the proportion of people who responded or did not respond
    - This visualization shows that the dataset is balanced, having an equal number of customers responding and not responding to campaigns
    - Having a balanced response rate makes this optimal for training the model



2) Correlation Matrix Heatmap:

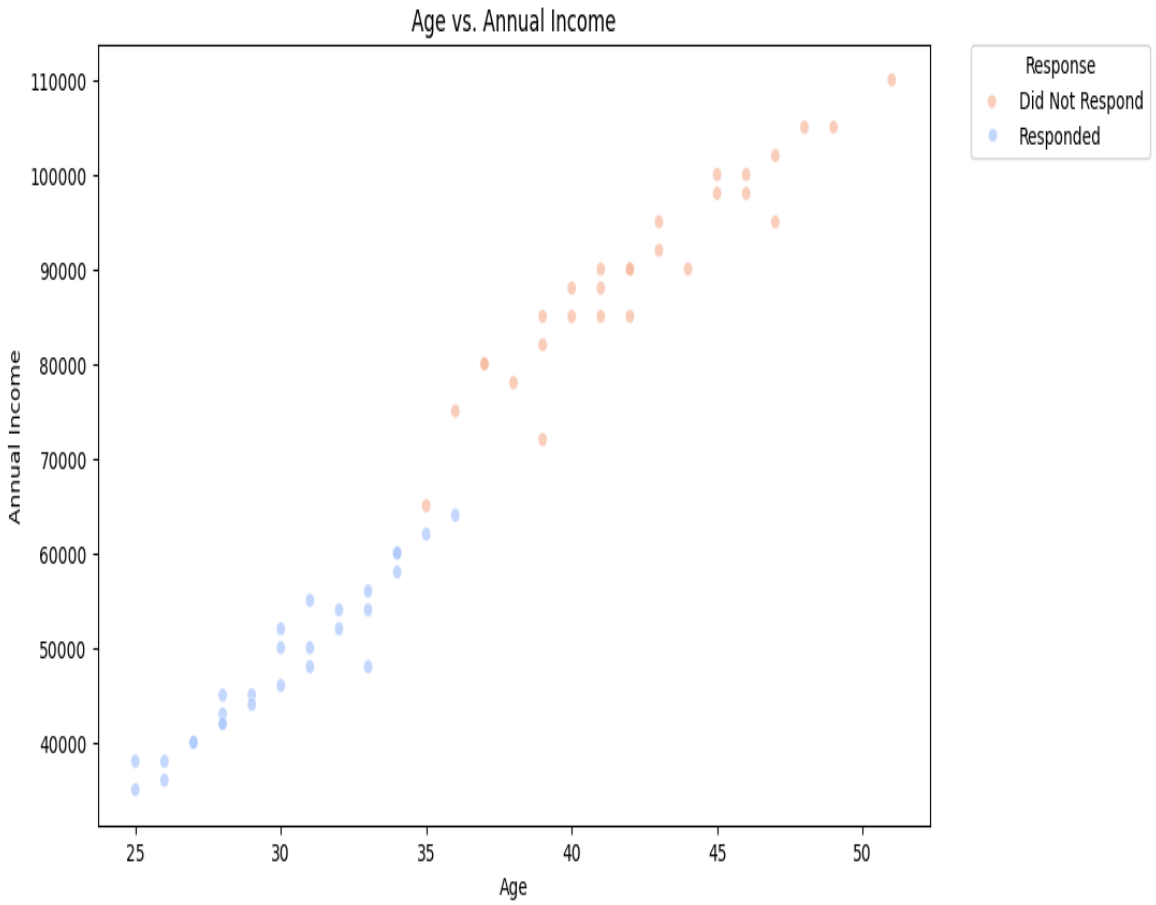
- Shows the different features and how they correlate with each other.
- Correlation Interpretation:
  i. Positive Correlation:
    1. **Age and Annual Income and Credit Score:** Older individuals are usually the ones with the higher income and credit score
    2. **Responded and Annual Income:** Those who responded to the campaigns are likely to have a high income

  ii. Negative Correlations:
    1. **Gender and Marital Status:** Shows that all males are married, and all females are single, this is due to the provided data
    2. **Gender and Response:** Shows that males tend to respond more than females.



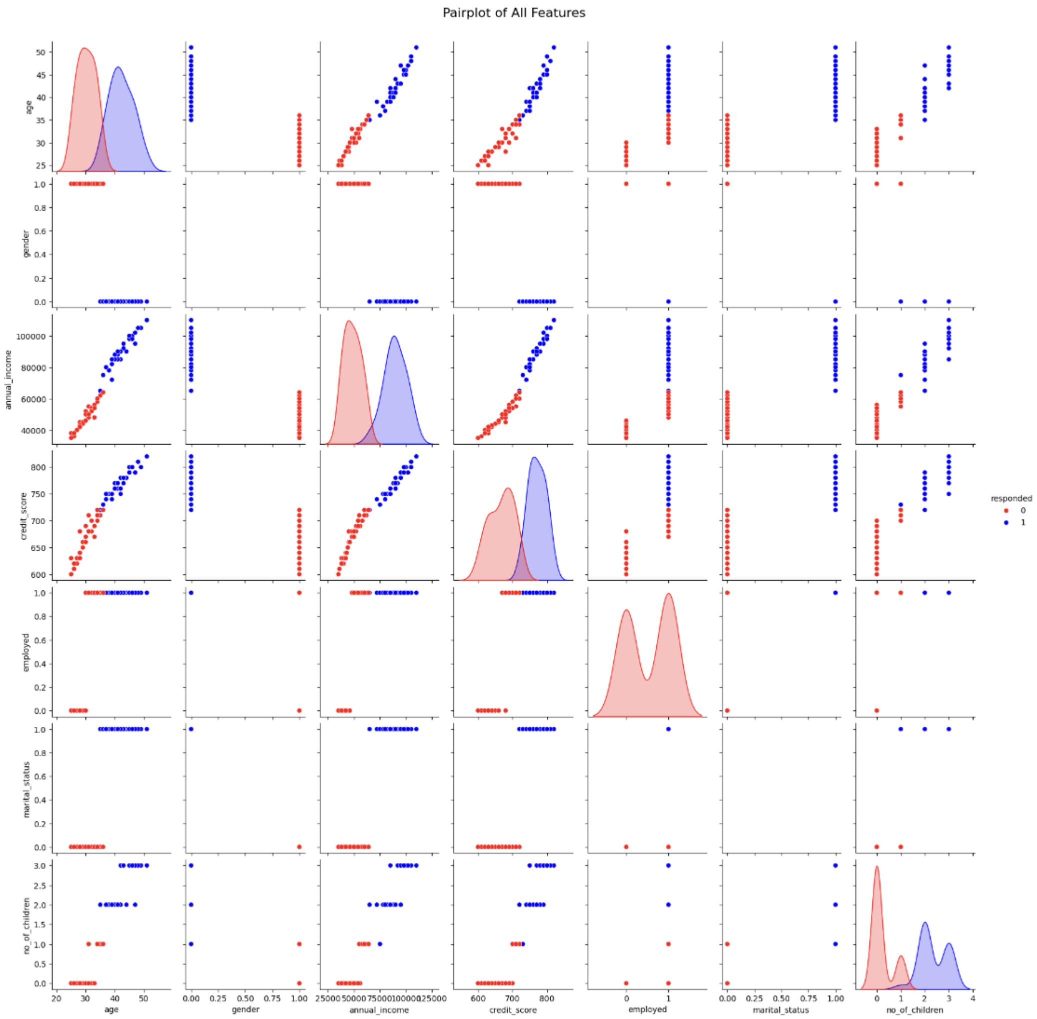
3) Scatter Plot of Age vs. Annual Income with Response as the Hue:

- This visualization helps identify how response is distributed through the different age and income tiers.
- Younger people with lower incomes are more likely to respond
- Older individuals with higher income are more likely to not respond



4) Pair Plot of All Features:

- Like a correlation matrix, it shows how other features relate to each other
- Individuals who are younger tend to respond and have lower income
- High credit scores are more common with responders
- Most responders are employed and have less children



In Conclusion:

- Younger individuals with lower incomes are more likely to respond to marketing campaigns while older individuals with higher income are less likely to respond.
- Higher credit scores and employment are associated with high responses

User Guide

Data Set: https://www.kaggle.com/datasets/sujithmandala/marketing-campaign-positive-response-prediction

GitHub: https://github.com/ivvle/c964_Task_2_Part_C

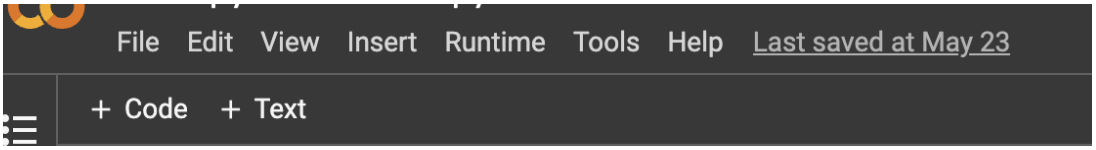
Google Collab:
https://colab.research.google.com/drive/1FQfGmeyrbS7ka7MuA6tGkgKwKUbDCZCp?usp=sharing

Using the provided Google Collab Link:

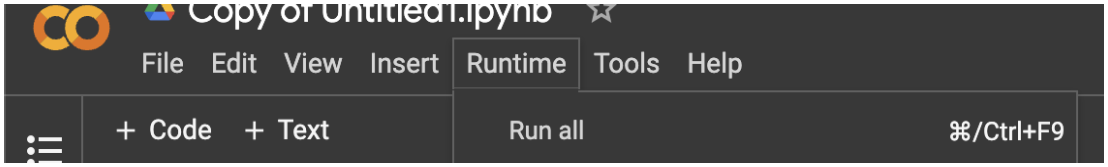
1) Please navigate to the provided Google Collab Link:

   https://colab.research.google.com/drive/1FQfGmeyrbS7ka7MuA6tGkgKwKUbDCZCp?usp=sharing

2) On the tool bar, select runtime and a drop-down menu will appear.

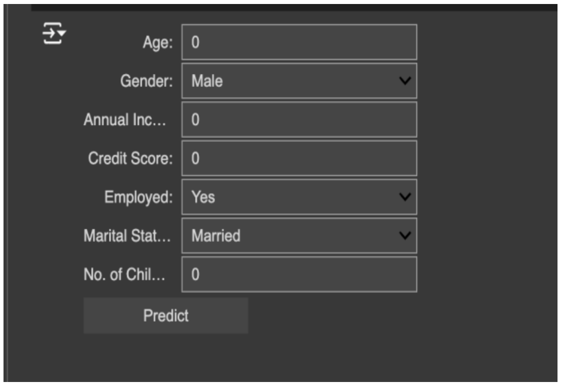


3) Select the first item "Run All"



4) Depending on your machine, it might take some time for all the cells to completely run.

5) Once the cells have completely run, navigate to the bottom of the page. This is where the interactive application is located. Input the required fields and clicked predict to receive a prediction on if the input data will respond to the campaign or not.



Age: 0

Gender: Male

Annual Inc… 0

Credit Score: 0

Employed: Yes

Marital Stat… Married

No. of Chil… 0

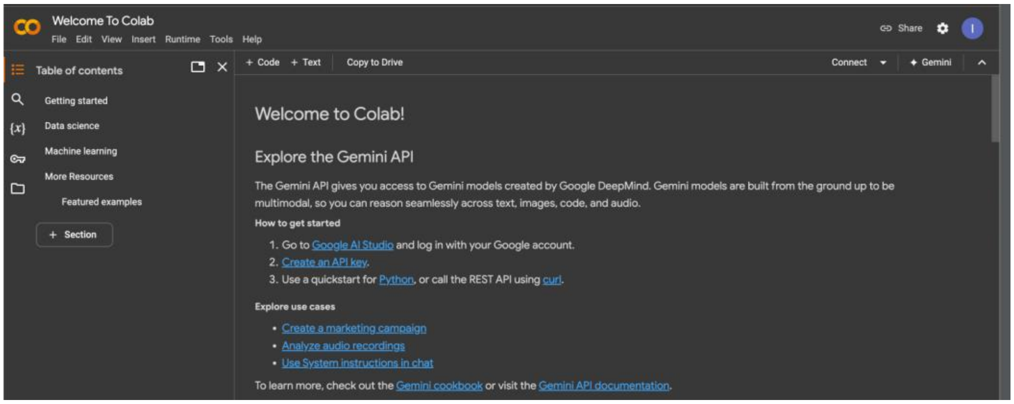
The application should run without any errors. If there are any issues with loading or viewing the file, please refer to the directions provided below as another option to execute and run the application.

Uploading Files on Google Collab:

1) Navigate to https://colab.google/

2) Sign in or create an account

3) Once you are signed in, navigate to the top left side of the window and select ‘File’



Welcome to Colab!

Explore the Gemini API

The Gemini API gives you access to Gemini models created by Google DeepMind. Gemini models are built from the ground up to be multimodal, so you can reason seamlessly across text, images, code, and audio.

How to get started

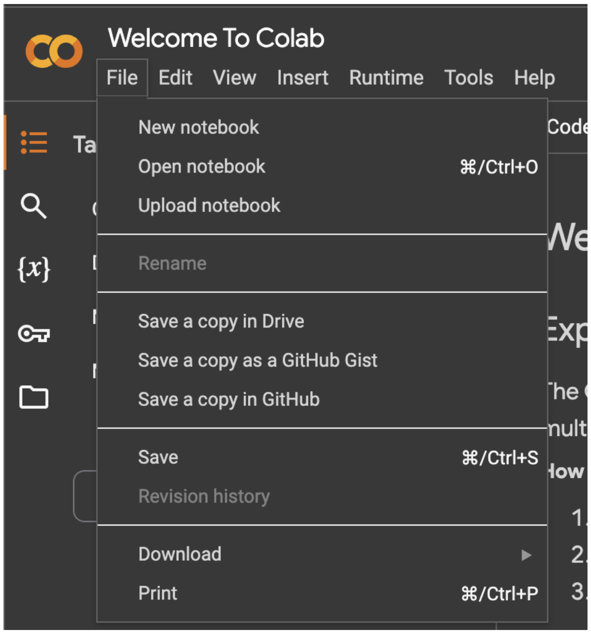
1. Go to Google AI Studio and log in with your Google account.
2. Create an API key.
3. Use a quickstart for Python, or call the REST API using curl.

Explore use cases

- Create a marketing campaign
- Analyze audio recordings
- Use System instructions in chat

To learn more, check out the Gemini cookbook or visit the Gemini API documentation.

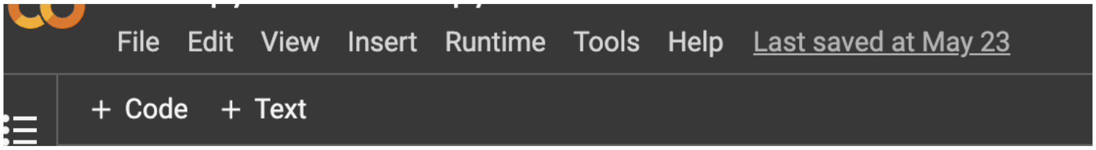
4) A drop-down menu will appear, select ‘Upload notebook’



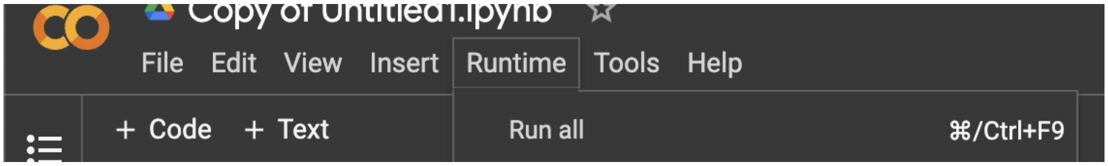
5) Please locate the file ‘campaignresponse_predictor.ipynb’ that is provided in the submission file.

    a. If you cannot find it, the file is also provided on the GitHub Link. Click here, download the file and proceed back to step 5 to continue.

6) After the file has been successfully uploaded, navigate to the tool bar, select runtime and a drop-down menu will appear.

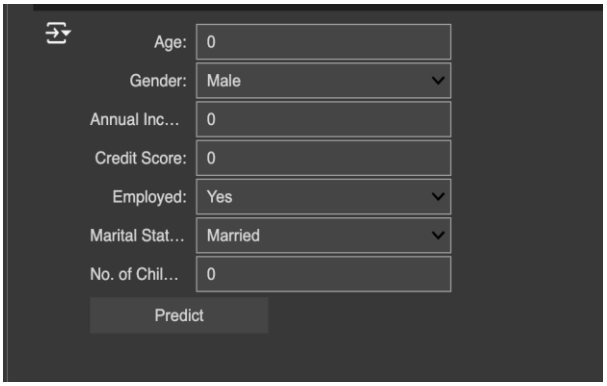


7) Select the first item “Run All”



8) Depending on your machine, it might take some time for all the cells to completely run.

9) Once the cells have completely run, navigate to the bottom of the page. This is where the interactive application is located. Input the required fields and clicked predict to receive a prediction on if the input data will respond to the campaign or not.



Age:	0
Gender:	Male
Annual Inc...	0
Credit Score:	0
Employed:	Yes
Marital Stat...	Married
No. of Chil...	0