Income Prediction for Nonprofit Donations

Student's Name

Western Governors University

Table of Contents

A. Project Overview

A1. Research Question or Organizational Need

This project will compare supervised machine learning algorithms to create a precise model for predicting whether a potential donor to a nonprofit organization earns more than $50,000. This model will need to be significantly more accurate than random chance.

A2. Context and Background

Headquartered in Encinitas, CA, Sunshine Benevolence Fund (SBF) is a nonprofit organization dedicated to providing educational opportunities and resources to underserved communities. After randomly sending nearly 32,000 letters to community members, SBF discovered that every donation received was contributed by individuals earning an annual income exceeding $50,000. Looking to expand their donor base, SBF has decided to use this discovery to specifically target residents of California who are most likely to donate to the charity. Given the sizeable working population of California, the charity is seeking a recommendation for a method that can effectively identify potential donors while minimizing the cost of mailing.

A3. Summary of Published Works

"Nonprofit Organization (NPO): Definition and Example"

This article from *Investopedia* provides a basic overview of nonprofit organizations. Economist Will Kenton details these organizations by discussing topics such as rules of operation, funding sources, and qualifications for NPO status. Kenton included a summary of the article as follows:

> Nonprofit organizations are entities that are tax-exempt and operate to better the community. By receiving funds from individuals, corporations, and governments, nonprofits deploy programs and strategies for public good. Nonprofits must meet a number of regulatory compliances to operate, and they often have the most passionate, loyal volunteers helping operate their mission. (2023)

“Machine Learning: Algorithms, Real-World Applications, and Research Directions”

In this 2021 research article published in the journal *Machine Learning and Applications*, Sarker provides an overview of machine learning algorithms, including supervised, unsupervised, semi-supervised, reinforcement learning, and deep learning. The article discusses the principles behind these techniques and their practical applications in areas such as cybersecurity, smart cities, healthcare, e-commerce, and agriculture. It also highlights challenges and research directions, serving as a resource for researchers, industry professionals, and decision-makers.

“Comparative Analysis of Classification Models on Income Prediction”

“Predictive Analytics is the underlying technology that can simply be described as an approach to scientifically utilize the past to predict the future to help coveted results” (Patel, Kakulapati, and Balaram, 2017). This conference paper, first published in the International Journal of Research in Information Technology and Computer Communications (IJRITCC), focuses on using predictive analytics to segment customers based on income levels. The authors propose an innovative approach that combines different classification techniques to predict income levels more accurately while minimizing risks and costs involved (2017). They also evaluate the performance of their algorithm using metrics such as true positives, false negatives, scored labels, and scored probabilities.

A3a. Relation of Published Works to Project

"Nonprofit Organization (NPO): Definition and Example"

I will be referring to this article to increase my business understanding as it relates to Sunshine Benevolence Fund. Having a general understanding of how nonprofit organizations operate could help with the project.

"Machine Learning: Algorithms, Real-World Applications, and Research Directions"

I will be referring to this research article during the modeling phase to determine which type of machine learning algorithm would be most appropriate for the scope of this project.

"Comparative Analysis of Classification Models on Income Prediction"

I will be consulting this paper and its findings during the modeling phase to choose classifiers to predict the income levels of potential donors.

A4. Summary of Data Analytics Solution

My solution to help SBF predict whether a potential donor earns more than $50,000 is to evaluate and compare a selection of supervised machine learning algorithms to deliver a single precise model to the charity. To do this, I will use Python and its libraries to preprocess the data, train and evaluate the different algorithms, optimize hyperparameters, select important features, and conduct hypothesis testing. This should result in a reliable prediction model for predicting individuals' income. I will deliver this model to the key stakeholders at project completion. They can then implement it to analyze relevant and up-to-date demographic data to help them identify potential donors. I will also deliver a detailed written report containing informative graphs. This report will summarize the analysis performed, including the evaluation results of the different supervised machine learning algorithms. It will show the stakeholders key metrics like accuracy and F-score with comparative bar charts so they can review each model's performance and feel confident with the solution. Other bar charts will visualize the most important features that influence potential donors' income and demonstrate to stakeholders a significant improvement over using random chance to target donation requests.

A5. Benefit to Organization and Decision-Making Process

The outcome of this project will affect multiple stakeholders, starting organizationally with SBF itself. The implementation of the resulting prediction model will help the charity determine the income levels of potential donors. The information can be used to decide how much of a donation to request from an

individual, or whether to approach them at all. This targeted approach will reduce the amount of money wasted on unsolicited donations, such as the 32,000 letters that the organization previously mailed. By knowing who to ask for donations, the charity can use its resources more efficiently and give its decision-makers the ability to plan strategically. The charity will be able to use the prediction model to increase its fundraising efforts, customize its strategies, and efficiently manage resources to make a positive impact in its mission. Existing and potential donors are also important stakeholders. The project's outcomes will help the organization find the least and most likely donors and adapt how they reach out to them. Some individuals will be targeted less as a result, but others will be targeted more. Other stakeholders include those directly working for and with SBF. Executives will gain valuable time to make other important decisions. Fundraising teams will save time and resources previously wasted on targeting unlikely donors. The data analytics team will have a powerful tool that will allow them to quickly and easily make recommendations on which individuals to target and how much should be requested, saving time and resources.

B. Data Analytics Plan

B1. Goals, Objectives, and Deliverables

The goal of this project is to construct a supervised machine learning model that accurately predicts whether an individual makes more than $50,000.

Project Objectives	**Project Deliverables**
Apply various supervised machine learning algorithms to the dataset, selecting then optimizing the most suitable one	Identification and optimization of the most suitable supervised algorithm
Construct a predictive model	Constructed predictive model
Assess model performance using evaluation metrics	Performance evaluation results
Create meaningful insights and analysis from the model's predictions	Meaningful insights and analysis
Inform decision-makers of results	Decision-making recommendations

B2. Scope of Project

In this project, data preprocessing, exploratory data analysis (EDA), feature engineering, model training, and evaluation will be conducted using the Python programming language within a Jupyter Notebook environment. The dataset that will be used is a previously cleaned CSV file that was extracted from the 1994 U.S. Census data. The solution will include transforming features and splitting the data into training and testing sets. Machine learning models, which are yet to be determined, will be trained and evaluated using training and prediction times as well as accuracy and F-score metrics. The best performing model will then be optimized using grid search and cross-validation. Feature importance analysis will be conducted to assess the significance of different features. A baseline dummy classifier will be compared with the supervised learning algorithm, and a hypothesis test will be performed to determine if the algorithm significantly outperforms random chance. The project scope does not include data collection or model deployment.

B3. Standard Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology will be used to manage this project. CRISP-DM provides a structured approach to planning a machine learning project. Specific sections of Python code will correspond to the different phases of the CRISP-DM methodology.

1. **Business understanding**
    - The specific needs of the charity as it relates to this project will be assessed through direct interaction.
    - Tools and technologies will be assessed and gathered.
2. **Data understanding**
    - The necessary data will be acquired and loaded.
    - EDA will be conducted to understand the feature set and uncover patterns or trends, providing insights into the dataset's characteristics to help identify key factors that could affect the project’s success.

3. **Data preparation**
    - The data preprocessing section will handle any missing values and transform features to ensure the data is clean and ready for analysis.
    - Feature engineering will be performed on the data to obtain or modify additional features.
4. **Modeling**
    - The model training section will involve identifying which machine learning algorithms to use and developing models from these algorithms.
    - Technical assessment of these models will occur during this phase.
5. **Evaluation**
    - The model evaluation section will involve evaluating machine learning models using training and prediction times as well as accuracy and F-score performance metrics to determine the most appropriate model.
    - Training and prediction times will be factored into the overall evaluation but are less important than other performance metrics within the scope of this project.
    - Accuracy and F-scores will help to perform quantitative evaluation of the chosen model.
    - Hypothesis testing will show whether the chosen model is capable of significantly outperforming random chance in accurately predicting an individual's income level. This will enable effective decision-making by the charity.
6. **Deployment**
    - This phase will not be applicable as it is outside of the project scope. However, it could be used in the future if the charity's decision-makers chose to proceed with the implementation of the chosen model, at which point a plan for deploying the model would be developed and documented. Any maintenance or enhancements to the chosen model as well as subsequent project reviews and reports would also happen in this phase.

B4. Timeline and Milestones

The following table lists each milestone, its projected start and end dates, and its projected duration:

Milestone	Projected Start Date	Projected End Date	Duration (days/hours)
Understanding the charity's goals and assessing available data	6/14/2023	6/14/2023	1 day
Data collection & exploration	6/15/2023	6/17/2023	3 days
Preparing the data for modeling	6/18/2023	6/19/2023	2 days
Developing and training a machine learning model	6/20/2023	6/22/2023	3 days
Model performance evaluation	6/23/2023	6/23/2023	1 day

B5. Resources and Costs

Resources	Costs
Personnel	
- Me	$0/hour Personal: work hours completed
Technology	
- CSV File	Free
- Python	Free
- Jupyter Notebooks	Free
- Libraries	Free
Infrastructure	
- Personal Computer	N/A
- Internet Connection	N/A

Since I will be the only person tasked with this project, I am the only personnel directly necessary for its completion. As I am not getting paid for the project, the personnel cost will be $0/hour. The only cost to me is the work required which can be objectively measured in hours. The estimated project duration is between 70 and 80 hours. All technology involved in the project will be free. The CSV file that contains the dataset will be downloaded at no cost. Python, its libraries, and the Jupyter Notebook environment are all open source and cost nothing to use. As not much computational power is expected to be necessary, the only infrastructure needed to complete the project will be a personal computer with internet connection. Since this has already been obtained for purposes unrelated to the project, the cost will be zero. The overall project cost, therefore, will be $0.00.

B6. Criteria for Success

To measure the success and effectiveness of the project outcomes, I will consider the following criteria:

- After training different models, I will compare their prediction results with the actual income class labels in the test dataset. This will hopefully output an accuracy score of 80% (0.80) or higher, indicating a successful model.
- I will also compare the predicted income class labels with the actual labels in the test dataset to calculate the F-score, a value that shows the candidate models' ability to handle false positives and false negatives by combining precision and recall. A model with an F-score of 0.75 or higher will be considered successful.
- After choosing and optimizing the best performing model, I will conduct a one-sample t-test to compare its performance with random predictions made by a baseline dummy classifier. The test will evaluate data from the model's predictions and the actual income class labels. The results of this test will output a p-value that will be compared to a significance threshold of 0.05. If the p-value is below this threshold, it can be determined that the final model significantly outperforms random chance.

Criterion/Metric	Required Data	Cut Score for Success
Accuracy of predictive models	Predicted and actual income class labels	>= 80%
F-score of predictive models	Predicted and actual income class labels	>= 0.75
P-value	Predicted and actual income class labels	p-value < 0.05

Note: Model training and prediction times will be considered when selecting the most appropriate model, but they will not be included in the success criteria as I will be subjectively evaluating these results.

C. Design of Data Analytics Solution

C1. Hypothesis

Organizational Need: This project will compare supervised machine learning algorithms to create a precise model for predicting whether a potential donor to a nonprofit organization earns more than $50,000. This model will need to be significantly more accurate than random chance.

Null Hypothesis: There is no significant difference between the performance of supervised machine learning algorithms and random chance in predicting whether a potential donor earns more than $50,000.

Alternative Hypothesis: Supervised machine learning algorithms significantly outperform random chance in accurately predicting whether a potential donor earns more than $50,000.

C2. Analytical Method

The predictive method I will be using in this project is logistic regression.

C2a. Justification of Analytical Method

Logistic regression is a statistical method that was designed for binary classification problems. It can be used to predict the probability of a particular event occurring, in this case, whether a potential donor earns more than $50,000. Since there are only two possible outcomes, this is a binary classification problem, making logistic regression an appropriate analytical method for the project. Using this method will also allow me to determine the features that contribute to higher income levels.

C3. Tools and Environments of Solution

I will be using the Python programming language in a Jupyter Notebooks environment for data extraction. Python has a large set of libraries and packages that simplify data extraction and transformation. Since the dataset I will be working with is a CSV file, I will take advantage of Python's Pandas package to load and analyze the data within. I have extensive familiarity using Python in Jupyter Notebooks, so I am choosing to do all coding and exploratory visualizations within that environment.

C4. Methods and Metrics to Evaluate Statistical Significance

Using the Python programming language within a Jupyter Notebook environment, I will use methods such as data preprocessing, exploratory data analysis, feature engineering, model training, and model evaluation. This evaluation will be based on training and prediction times as well as accuracy and F-scores. Once I choose and optimize the most appropriate model, I will conduct a one-sample t-test to compare the model’s performance with random predictions made by a baseline dummy classifier. If the resulting p-value is below the threshold I set of 0.05, I will be able to reject the null hypothesis and state that the model’s ability to outperform random chance is statistically significant.

C4a. Justification of Methods and Metrics

The methods I will use are necessary for data preparation, analysis, and model building. Data preprocessing makes sure the dataset is of excellent quality and ready for analysis. Exploratory data analysis will help identify any patterns or relationships among the feature set. Feature

engineering will fine tune these features to increase the predictive power of the model. Model training will train the candidate models on the dataset to make accurate predictions. Evaluation metrics such as training and prediction times along with accuracy and F-scores will help determine the best algorithm to use and subsequently optimize it. Training times will measure how long each model takes to train on the dataset, while prediction times will show when each model makes their predictions. Accuracy measures the models' overall correctness, while the F-score shows the models' ability to handle false positives and false negatives by combining precision and recall. The one-sample t-test will determine statistical significance by comparing the final model's performance against a baseline. Collectively, these methods and metrics are appropriate for evaluating the project's success in accurately predicting potential donors.

C5. Practical Significance

The resulting model will provide a high level of practical significance and value for the charity's decision-makers. The model's ability to accurately predict whether an individual makes over $50,000 annually will help the charity target potential donors more effectively via marketing strategies such as fundraising campaigns and customized cold calls. This targeted approach will decrease the money wasted on random solicitation such as the 32,000 letters previously mailed by the organization. Knowing whom to request donations from will allow the charity to use organizational resources more efficiently and increase the decision makers' ability to strategically plan. The results of this project will also provide valuable donor information such as demographics and preferences, increasing the ability to develop lasting relationships with donors to encourage long-term donations.

C6. Visual Communication

I will be using graphical representations throughout the project to communicate my findings as the project progresses. I will use histograms to visualize quantitative data I suspect of being highly skewed. After feature transformation, I will again create histograms to show the new value distributions. Once I have determined which supervised learning algorithms to evaluate, I will visualize their performance metrics

using a series of grouped bar charts. This is because the performance metrics are a mixture of quantitative and categorical values, and grouped bar charts are well-suited for visualizing this type of data. This will provide side-by-side comparisons for each model and make it easier to determine which algorithm performed best. Later in the project, another grouped bar chart will be displayed. It will display the normalized weights of the top five features by order of importance for making predictions using the chosen algorithm. Finally, I will display the statistically significant findings as a simple bar plot showing the accuracy of the supervised learning algorithm compared to the baseline dummy classifier. I will use the above visualizations because they are easy to read and understand and are appropriate for the dataset.

D. Description of Datasets

D1. Source of Data

The lone dataset used for analysis is "census.csv," a comma-separated values file that was originally created using data from the 1994 U.S. Census. It is hosted in the following GitHub repository:

UDACITY. (2021). CharityML Project Repository. GitHub. Retrieved from: github.com/udacity/DSND_Term1/tree/master/projects/p1_charityml

D2. Appropriateness of Dataset

This dataset is appropriate for addressing the project goal due to the following factors:

- It has features such as demographics, occupation, education, and employment that are all relevant to the project's goal of predicting income.
- It is free from data quality issues such as missing values or bad entries. This ensures accurate and reliable data is used when training the machine learning models.
- It has a sufficient number of records to ensure variability and diversity of potential donors.
- It is publicly available.

D3. Data Collection Methods

This dataset has been adapted from a previously published paper titled "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid" authored by Ron Kohavi. The original dataset, extracted from the 1994 U.S. Census database, can be found at https://doi.org/10.24432/C5GP7S. The adapted data specifically used for this project is in the form of a CSV file that includes minor modifications to the original dataset. These modifications involve the removal of the 'fnlwgt' feature and the exclusion of records with missing or incorrectly formatted entries. This cleaned dataset was committed to a git repository by Joshua Bernhard in 2018 and is managed by online educational organization UDACITY. The dataset is publicly available on GitHub, a platform specifically designed for hosting code and data, at github.com/udacity/DSND_Term1/tree/master/projects/p1_charityml. There are advantages and disadvantages of collecting data using this methodology. Advantageously, this free and publicly available dataset allowed me to obtain the data quickly and easily without needing to request permission or deal with restrictions. Publicly available datasets offer data from a wide variety of sources, and a quick web search is all that was needed to find and retrieve this particular dataset. Some disadvantages of this methodology include limited control over the data collection process. Gathering data that has been collected and cleaned by another individual can allow biases to enter the dataset. Additionally, publicly available data often has quality issues. Although not present in the dataset used for this project, issues such as missing values, errors, and inconsistencies can arise. These obstacles can make publicly available data less appealing, but no such obstacles were encountered in this project’s data collection process.

D4. Data Quality

As the CSV file used for this project was thoroughly cleaned and appropriately formatted beforehand, no quality issues need to be addressed.

D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance

This dataset is subject to the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. This license allows the user to utilize the data for any purpose, as long as the user appropriately attributes the original authors and refrains from making modifications or engaging in commercial activities involving the data. The CSV file is publicly available data based on the results of the 1994 U.S. Census. It does not include any personally identifiable information (PII) such as government-issued identification numbers, street addresses, email addresses, dates of birth, or names. There are no data governance issues with the data itself. However, ethical issues could arise from using this dataset in the project, as it includes demographic data such as marital status, education level, race, sex, occupation, and national origin.

D5a. Precautions

In this project, I will ensure that the modeling data is free of personal bias. I will not allow any presumptions about the included demographic data to influence the outcome of the project. I will only use all features or the feature importances attribute for final model training. No other precautions are deemed necessary.

E. Sources

UDACITY. (2021). CharityML Project Repository. GitHub. Retrieved from: github.com/udacity/DSND_Term1/tree/master/projects/p1_charityml

Kenton, W. (2023, June 15). Nonprofit Organization (NPO): Definition and Example. Retrieved from https://www.investopedia.com/terms/n/non-profitorganization.asp

Sarker, I. H. (2021). Machine learning: Algorithms, Real-World Applications, and Research Directions. SN Computer Science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Patel, Kakulapati, and Balaram (2017). Comparative Analysis of Classification Models on Income Prediction. International Journal on Recent and Innovation Trends in Computing and Communication, 5(4), 451–455. https://doi.org/10.17762/ijritcc.v5i4.435