Data Management and Data Analytics Capstone Topic Approval Form

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of the following areas so you will have a complete and realistic overview of your project. Your course instructor cannot approve your project topic without this information.

Student Name: [redacted]

Student ID: [redacted]

Capstone Project Name: "Income Prediction for Non-profit Donations."

Project Topic: Viability of using a supervised learning algorithm to predict income levels

Research Question: Is a supervised learning algorithm capable of significantly outperforming random chance in accurately predicting an individual's income level?

> **Hypothesis:** A supervised learning algorithm is capable of significantly outperforming random chance in accurately predicting an individual's income level.

Context: Situated in Encinitas, CA, Sunshine Benevolence Fund (SBF) is a non-profit organization dedicated to providing educational opportunities and resources to underserved communities. After sending nearly 32k 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 target residents of California who are most likely to donate to the charity. Given the sizeable working population of nearly 15 million individuals in California, the charity is looking for a method that can effectively identify potential donors while minimizing the cost of mailing.

Data: The data collected is from an existing dataset that contains approximately 32k data points with 13 features and 1 target variable per data point. Following is the dataset's featureset exploration (UDACITY, 2021):

Features

- **age**: Age
- **workclass**: Working Class (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)
- **education_level**: Level of Education (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool)
- **education-num**: Number of educational years completed
- **marital-status**: Marital status (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse)

- **occupation**: Work Occupation (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces)
- **relationship**: Relationship Status (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
- **race**: Race (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)
- **sex**: Sex (Female, Male)
- **capital-gain**: Monetary Capital Gains
- **capital-loss**: Monetary Capital Losses
- **hours-per-week**: Average Hours Per Week Worked
- **native-country**: Native Country (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands)

Target Variable

- **income**: Income Class (<=50K, >50K)

The dataset is pre-existing and is available online at github.com

The 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.

Data Gathering: 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 capstone 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.*

Data Analytics Tools and Techniques: To analyze the data, the following data analytics tools and techniques will be used:

- Python programming language
- Jupyter Notebooks
- Python libraries Pandas, NumPy, Matplotlib, and scikit-learn
- Supervised learning algorithms GaussianNB, AdaBoostClassifier, and KNeighborsClassifier
- GridSearchCV
- fbeta_score
- DummyClassifier

- Descriptive statistics
- Data visualization
- Data preprocessing and normalization
- Feature engineering
- Statistical hypothesis testing

Justification of Tools/Techniques: I will be using the Python programming language because of its versatility and ability to efficiently aid in data analysis techniques. It has a vast collection of libraries such as Pandas, NumPy, Matplotlib, and scikit-learn. Using Jupyter Notebooks will help me stay better organized and quickly implement the Python code necessary to complete the project. Descriptive statistics will allow me to get a better look at the dataset by providing insights into its characteristics. Data preprocessing techniques will help normalize numerical features. Supervised learning algorithms will train and evaluate predictive models for identifying individuals who make over $50,000. Applying multiple algorithms along with data visualizations will help me choose the most appropriate model. GridSearchCV will help optimize this model's performance. By using performance metrics such as accuracy and F-score, I will be able to perform quantitative evaluation of the chosen model. When combined, the aforementioned tools and techniques along with a one-sample t-test will show whether a supervised learning algorithm is capable of significantly outperforming random chance in accurately predicting an individual's income level. This will enable effective decision-making by the charity.

Application Type, if applicable (select one):

- ☐ Mobile
- ☐ Web
- ☒ Stand-alone

Programming/Development Language(s), if applicable: Python

Operating System(s)/Platform(s), if applicable: Microsoft Windows 11

Database Management System, if applicable: N/A

Project Outcomes: The goal of this project is to evaluate and optimize several supervised learning algorithm [redacted] yields the highest charitable donation while reducing the total number of solicitation letters sent by the charity SBF. The charity could continue to randomly mail out thousands of letters to the community, but they have commissioned me to see if machine learning algorithms would better serve them in their search for potential donors. I will evaluate several prediction models by visualizing their training and prediction times, accuracy, and F-scores. I will then choose the best model to analyze the dataset. By comparing the model's performance metrics against a baseline or random chance level, I will be able to conclude whether the algorithm's predictions are statistically significant. This will provide me with enough evidence to reject or support my hypothesis that a supervised learning algorithm is capable of significantly outperforming random chance in accurately predicting an individual's income level. I will report my findings and recommend an appropriate prediction model for the charity's mailing service.

Projected Project End Date: 6/16/2023

Sources:

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

Kohavi, R. (1996). Census Income. UCI Machine Learning Repository. Retrieved from: https://doi.org/10.24432/C5GP7S

Human Subjects or Proprietary Information
Does your project involve the potential use of human subjects? (Y/N): N
Does your project involve the potential use of proprietary company information? (Y/N): N

STUDENT SIGNATURE

[redacted]

By signing and submitting this form, you acknowledge that any cost associated with the development and execution of your data analytics solution will be your (the student) responsibility.

TO BE FILLED BY A COURSE INSTRUCTOR

The capstone topic is approved by a course instructor.

COURSE INSTRUCTOR'S NAME AND SIGNATURE:

William L. Dean Jr., Ph.D.

COURSE INSTRUCTOR APPROVAL DATE: 11 June 2023

Project Compliance with IRB (Y/N):Y