# **ML1000 FINANCE GROUP**

**Assignment 1**

Finding Customers for a Robo Advisor

17th February 2019

**ML1000 Finance Group**

Daniyal Shamim

Juan Calvillo

Mark Lewis

Salman Amin

Table of Contents

[Business Background & Context 2](#_Toc1303346)

[Background 3](#_Toc1303347)

[Marketing Strategy 3](#_Toc1303348)

[Our Task 3](#_Toc1303349)

[Analytical Problem Framing 4](#_Toc1303350)

[Business Problem: 4](#_Toc1303351)

[Additional Background 4](#_Toc1303352)

[Analytical Problem 4](#_Toc1303353)

[Data Understanding 5](#_Toc1303354)

[The Dataset 5](#_Toc1303355)

[Data Dictionary 5](#_Toc1303356)

[Data Exploration 7](#_Toc1303357)

[Data Perpetration 8](#_Toc1303358)

[Model Development 8](#_Toc1303359)

[Model Deployment 8](#_Toc1303360)

# Business Background & Context

## Background

Our client is a Startup Robo-advisor who wants to target retail investors that are most likely to sign up for our client’s Robo Advisory service. This service provides customers with diversified portfolios and charge them a fee of 0.5% of Assets Under Management (AUM).

Through previous market research, our client has determined that they should only target prospects with at least $50,000 in annual household income. Households with more than $50,000 in annual income are more likely to generate savings every year to invest in our client’s service and will benefit from the lower fees charged by our client compared to more traditional financial advisors.

## Marketing Strategy

Our clients want to develop a marketing plan and go-to market strategy that minimizes their customer acquisition costs. They want to target the segment of prospects that is most likely to earn more than $50,000 in income a year. The best marketing plan would provide the highest revenue per dollar spent on acquiring the customer.

## Our Task

Our client is looking for our expertise to identify the key drivers of annual household income so they can identify their target audience, and the best way to advertise our client’s services to them. In other words, our client is looking for exactly what demographic segments they should target to have the highest success rate for their marketing efforts.

# Analytical Problem Framing

## Business Problem:

Find the best prospects with an income of at least $50,000 to identify the target audience of the Robo Advisor’s marketing efforts.

## Additional Background

Census Income data from the UCI Machine Learning Repository[[1]](#footnote-1) has been used to conduct the analysis. Our client currently selects prospects using the following attributes:

* Customer socio-demographic and geolocation data.
* Customer education level
* Customer potential growth ranking

This prospect selecting criteria will be used a base line in our data exploration phase.

## Analytical Problem

We want to identify attributes (or features) of those individuals who make more than $50,000 per year. Then we will use these attributes to predict the demographic segment which should be targeted by our client.

According to our client’s policies and our ethical guidelines, we need to ensure that our analysis and prediction does not contain any biases.

# Data Understanding

## The Dataset

We will be using a dataset with almost 49K rows from the census bureau that identifies individuals making more than and less than $50,000 per year. The data is collected from 51 states in the U.S. and contains features such as Age.

The attributes in our dataset could be divided into three main categories: demographic, work related, and financial history. Specific attributes related to each category are listed below.

|  |  |  |
| --- | --- | --- |
| **Demographic** | **Work Related** | **Financial History** |
| Age | Education | Capital-gain |
| Race | Occupation | Capital-loss |
| Sex | Work class |  |
| Marital-status | Hours-per-week |  |
| Relationship |  |  |
| Native-country |  |  |

Race and sex will not be used in our analysis to exclude any potential bias.

## Data Dictionary

The following table describes the various columns available in our dataset and the values allowed in those columns. from the Census Bureau and a description of what they mean

| Column Name | Definition | Values |
| --- | --- | --- |
| Age | Customer age | 1..120 |
| Workclass | Type of customer workclass | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. |
| Education | Customer 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. |
| Marital-status | Customer marital-status type | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. |
| Occupation | Customer occupation segment | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces. |
| Relationship | Customer relationship type within the family | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. |
| Race | Race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. |
| Sex | Customer gender | Female, male |
| Capital-gain | Customer capital gain in a period of time | Continuous |
| Capital-loss | Customer capital lose in a period of time | Continuous |
| Hours-per-week | Working hours for the customer | Continuous |
| Native-country | Place of birth | 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, Holland-Netherlands. |

# Data Exploration

In our data exploration phase, all the features from the dataset were analyzed in detail. Our goal was to understand the data and come up with a right approach to prepare the data for our analysis and developing a model.

Our client is an equal opportunity company and has policies against any type of discrimination. To avoid any possibility of discrimination in our analysis, we decided to remove race and sex as features.

We also looked at correlations among the features. In this process, the goal was to find interdependencies which could provide us quick insights and help in the selection of a good machine learning model. For example, having a high education level (Masters) and belonging to a work class (Private sector) might give us a good start in finding our prospects.

Our data exploration gave us a reduced feature list with Occupation, Native Country, Work Class, Age and Education. We also found issues in our data, like incorrect or missing values, as explained in the next section.

# Data Perpetration

There was not a large amount of incorrect or missing data in the dataset, which represented less than 0.02% which were deleted. Some values were misspelled and were corrected manually in the dataset.

# Model Development

Please refer to the R Markdown for the assignment.

The code can be found in this repository.  
https://github.com/markglewis/IncomePrediction

# Model Deployment

Please refer to the Shiny app for the assignment found in the following link.

https://markglewis.shinyapps.io/IncomePrediction/

1. https://archive.ics.uci.edu/ml/datasets/Census+Income [↑](#footnote-ref-1)