Xin Tang
6/14/2024
Predictive Analytics
DSC630-T302 (2247-1)

Income level prediction

Project Introduction:

In United States, Income provides economic resources that shape choices about housing, education, childcare, food, medical care, and more. For a person or a family, it plays a crucial role in meeting basic needs such as education, shelter, water and sanitation. It is also imperative for companies to know this information to target the most suitable groups and mitigate the financial risk in situations like credit card or loan application, even housing aid application.

For various reasons, consumers often unable or unwilling to provide their true income information, which makes it difficult to trust the self-reported income data.

However, in the big data era, there are other ways to reveal the truth through other more indirect objective information, like education level and marriage status. With this objective in mind, I would like to develop a predictive model to group the people into high- and low-income levels.

The result from this model will provide a cross-check point on self-reported data, as well as an indication of what consumer background will yield high- or low-income level. From there, Interested companies will target correct consumer group and ensure a better yield of investment.

Dataset selection and approach:

Dataset:

The data resource for this project is from Kaggle, which is an open-source data science community with many data suitable for data analysis. This dataset has 11 independent variables like age, work class, education and marital status etc., which are all objective information. It contains almost 32 thousand individual's record, so is large and complex enough to create a predictive model.

As a backup, I picked up a world population dataset, which is also interesting and impacts us on a great scale. As a second backup, I have an insurance dataset I am interested in exploring, if the first 2 do not work out.

For the income dataset, I would like to:

1. Build a model can predict the level of income (higher or lower than >$50K, which is the criteria used in dataset).
2. Find out which factor is the most important to influence the income level.

All datasets can be found in Kaggle. Refer to citation sections for details:

Approach:

What types of model or models do you plan to use and why?

I peeked into the dataset, they have mixed numerical and categorical data. But not a text-based dataset, so At least I do not plan to do sentiment analysis or tf-idf matrix.

This is a classification problem; the outcome will be with a binary classification label. For a classification problem, First I like to try random forest classifier. I will also try to do hyper parameter tuning to see how much it could improve the performance. Next, I will try XGboost. XGBoost, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems (Nvidia).

How do you plan to evaluate your results?

(Wohlwend, 2023), Evaluation plays an indispensable role in the process of building robust and effective classification models. It will help to understand model performance, make a objective model selection, as well as measuring business impact.

I will run classification report to get accuracy, precision, recall and F1 score for the 2 models. I will also run confusion matrix to visualize the result.

What do you hope to learn?

Developing a model and using it to predict is a systematic project. It involves many steps. Like those outlined in CRISP-DM method. I am interested in learning how to improve the performance of the model. How to assess the evaluation result from end user aspect, since sometime the results from one index may not equally important, or not even relevant.

Assess any risks or ethical concerns with your proposal.

Risk1: imbalanced data. → in income data, it is common to have majority of data belongs to low income, high income is a relatively small portion. This can be addressed using proper packages specific to handle this issue.

Risk2: The factors impact income level changes over time. For example, A college graduate majored in art will have dramatic income difference from one majored in computer science. Most census data do not provide college major details.

Risk 3: Ethical concerns: income will play a critical role in determining eligibility of many things, like student loan, social benefit, tax, credit limit for consumption and many more. An unvalidated data resource or incorrect model will bring serious consequences. To

develop a model ready for deployment, the dataset it depends on needs to be carefully validated. The final model selected also needs to be tested extensively.

Identify a contingency plan if your original project plan does not work out.

To avoid the last-minute start over and working on an alternative dataset, I would spend a lot of time on EDA, visualize the data distribution, check correlations, decide weight etc.

To avoid run into technical roadblock, I plan to use experience from others, like professor and some classmate who working on the data science field already.

Last, Since I will travel a lot and not everywhere I can access database in US. I will keep a local copy of everything in my computer, if needed, I will email professor some evidence of work and ask for an extension.

Include anything else you believe is important.

Following the timeline is the key to the on time finish the project. Also prepare in advance for answers to the non-technical questions (like model's application, ROI etc) will also help to show real value of the project.

Resources:

Income dataset:

ANAGHA K, P (Oct. 2023). Adult Income Census, Predict whether income exceeds $50K/yr based on census data, Retrieved June 10, 2024 from https://www.kaggle.com/datasets/anaghakp/adult-income-census/data

World population dataset:

SOURAV BANERJEE (June. 2022). World Population Dataset, Global Headcount: World Population Dataset by Country/Territory, Retrieved June 10, from

https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset

Insurance dataset:

KUSH SHAH (June. 2020). The Insurance Company (TIC) Benchmark. Retrieved June 10, from https://www.kaggle.com/datasets/kushshah95/the-insurance-company-tic-benchmark

Nvidia *(date unknown)*: XGBoost – What Is It and Why Does It Matter?

https://www.nvidia.com/en-us/glossary/xgboost/

Brandon Wohlwend *(Ju16,2023):* Evaluating classification models

https://medium.com/@brandon93.w/machine-learning-evaluating-classification-models-18713af3d764