Capstone Project Proposal

Income Prediction for the Mexican context 2022

By Dámaris Flores Albores

Business Understanding

The purpose of this project is to make accurate income predictions based on an existing Mexican government survey made from October to December of 2022, called: Encuesta Nacional de Ocupación y Empleo (ENOE) – *National Survey of Occupation and Employment.* This model intends to predict based on features as location, gender, age, scholarship, number of children, marital status, salary zone, college career, occupation position and sector of economic activity, the approximate income of a person at present in México.

This project can be applied for social studies. It gives us the opportunity to analyze which features are most relevant when determining income in México and by that, highlight the inequalities that society lives and need to be address by future public politics.

My motivation behind this project is to learn a bit more of my country's context and to be able to use the data that is generated everyday for finding new strategies that could help us solve the problems that afflict the country.

Data Understanding

For this purpose, the data set that I will be using is based on the ENOE survey (4th trimester, 2022), which is open data and available for downloading from the official website of INEGI (National Institute of Statistics and Geography. Due to its complexity and extension, I will select maximum 15 features, among them: location, gender, age, scholarship, number of children, marital status, salary zone, college career, occupation position, sector of economic activity and monthly income.

Data Preparation

The dataset is made up of 396,329 records with 114 variables. It is fully encoded, for that I will need to access the data dictionaries to find out what does each column and number code means and filter by economically active people with no nulls. I will also loose almost 100 of the variables for efficiency purposes.

This will be a challenge due to the extension of the database and the granularity of the questions. Due to the nature of the questions, in some cases people answered in terms of income by week or by hour, this means it will be necessary to calculate in some cases the monthly income or normalize it by hour.

Also, most of the variables are categorical, for that will be necessary to first decode and then convert them (one hot encoding).

Modeling

This project will be a regression problem.
As salary is a continuous variable I will use linear regression, I will also try to use Random Forest.
My target variable will be the monthly income (in case of obstacles to get this data for a significant proportion of the records I will convert it to income per hour).

Evaluation

As this will be a regression problem, I will be using the following metrics:
R Square, Mean Square Error and accuracy

Tools/Methodologies

The plan is to use Linear Regression and Random Forest.