Hello guys,

I believe we should state the problem in following way:

Business problem – Fraud detection and risk management.

In many cases, when people apply for a loan, they don't provide any proof of their income (self-reported income). To estimate the amount of risk, we need some cost-effective tools that could estimate how reasonable the self-reported income is for an individual. Depending on the probability of the class predicted by our classification model, a bank could make a decision of asking for additional documents or refuse a loan if the proof is not provided.

The dataset we work with provides only two classes of a target value – an income above or below $50k. Nevertheless, it gives an opportunity to explore potential drivers of prediction for explainable 'glass-box' models and make some conceptual decisions about the model that could be trained on data with more classes in 'one-vs-rest' or 'one-vs-one' architecture.