Thank you for accepting this task. I need you to train a classification model on the attached dataset in csv format, which contains about 30,000 rows and two classes.

I will need you to send the following items as a final deliverable:

- The final, processed and cleaned dataset that you use to train and evaluate the model (in a csv file formatted the same as the original csv file)
- Predictions generated by that model for that final, cleaned dataset in a csv file with two columns: ID, prediction
- A piece of R or Python code (NOT an iPython notebook or RMarkdown file) that can easily be run on any csv file with the same columns and format (either by providing the new file name as a command line argument, or by changing a filename string in the R/python code)
- All of the code you produce over the course of this project (including code used for data cleaning, model selection, etc.). Included should be a text file that clearly indicates which piece of code does what, and which file(s) are used to generate predictions on new data (and how to do so)
    - Code must include in a comment header "Internal Microsoft Use Only."
- If sending compressed files, please send them in the ".zip" format, not other compression formats (e.g., ".tar," ".rar").

The target variable is the "over_50k" variable in the dataset.

A data dictionary for the provided training data is found below:

- id: an unique identifier for each observation in the dataset
- age: the age (in years) of the person
- workclass: The type of employment held by the person. Possible values:
    - Private
    - State-gov
    - ?
    - Local-gov
    - Self-emp-not-inc
    - Federal-gov
    - Self-emp-inc
    - Without-pay
    - Never-worked
- education: The highest education level attained by the person. Possible values:
    - HS-grad
    - Some-college
    - 10^{th}
    - Bachelors
    - Masters
    - Assoc-acdm
    - 12^{th}
    - 5^{th}-6^{th}
    - 11^{th}
    - Assoc-voc

    - 9th
    - 1st-4th
    - 7th-8th
    - Prof-school
    - Doctorate
    - Preschool
- education_num: The number of years of education attained by the person. Value should be between 1 and 16
- marital_status: The marital status of the person. Possible values:
    - Divorced
    - Married-civ-spouse
    - Never-married
    - Widowed
    - Separated
    - Married-spouse-absent
    - Married-AF-spouse
- occupation: The occupation of the person. Possible values:
    - Adm-clerical
    - Exec-managerial
    - Prof-specialty
    - Other-service
    - Tech-support
    - Machine-op-inspect
    - Craft-repair
    - ?
    - Sales
    - Transport-moving
    - Handlers-cleaners
    - Protective-serv
    - Priv-house-serv
    - Farming-fishing
    - Armed forces
- relationship: The role of the person within their household. Possible values:
    - Other-relative
    - Unmarried
    - Wife
    - Not-in-family
    - Husband
    - Own-child
- race: The race of the person. Possible values:
    - White
    - Black
    - Asian-Pac-Islander
    - Other
    - Amer-Indian-Eskimo
- sex: The gender of the person. Possible values:

    - Female
    - Male
- capital_gain: The amount of money the person gained in the year from stocks, bonds, etc.
- capital_loss: The amount of money the person lost in the year from stocks, bonds, etc.
- hours_per_week: The average number of hours per week the person worked.
- native_country: The person's native country. 42 possible values.
- over_50k: Whether or not the person made over $50,000 dollars in the year.

The task will need to be completed within 72 hours of starting the task.

Microsoft Research Project Participation Consent Form

INTRODUCTION

Thank you for taking the time to consider participating in a Microsoft Corporation research project. This form explains what would happen if you are hired for this Upwork task. Please read it carefully and take as much time as you need. Email the study team to ask about anything that is not clear.

Participation in this study is voluntary and you may withdraw at any time.

TITLE OF RESEARCH PROJECT

An Empirical Study of the Impacts of Upwork Task Structure

Principal Investigator: David Holtz

PURPOSE

The purpose of this project is to understand how tasks on Upwork can be structured in such a way that both clients and freelancers are more satisfied with the work process and end deliverables.

PROCEDURES

During this project, you will complete the task as described in the task description. After completing the task, you will be given a brief survey.

Microsoft may document and collect information about your participation by collecting the work deliverable you submit, the number and textual content of messages between you and the client, survey responses, and publicly available non-protected information included in your Upwork profile (e.g., past ratings, hourly rate). Some data processing and transcription may be performed by Upwork freelancers hired by Microsoft.

Approximately 1,000 participants will be involved in this study.

PERSONAL INFORMATION

Aside from your Upwork ID and Upwork display name, no personally identifiable information will be collected during this study. Your Upwork ID and Upwork display name will not be shared outside of Microsoft Research and the confines of this study without your permission and will be promptly deleted after compensation has been successfully provided (30 days or less). De-identified data may be used for future research or given to another investigator for future use without additional consent.

- Microsoft Research is ultimately responsible for determining the purposes and uses of your personal information.

- **Personal information we collect.** During the project we may collect personal information about you such as your name (as displayed on Upwork), city/town (as displayed on Upwork), employment history (as displayed on Upwork), and education (as displayed on Upwork).

- **How we use personal information.** The personal information and other data collected during this project will be used primarily to perform research for purposes described in the introduction above. Such information and data, or the results of the research may eventually be used to develop and improve our commercial products, services or technologies.

- **How we store and share your personal information.** Your Upwork ID and Upwork display name will be identified by a code. This code is what will be included in the research dataset, and the key to the code will be kept separate from data related to this study, which will be stored in a secured limited access location internal to Microsoft.

  Aside from the researchers of this study, your personal and study information may be shared with study team members outside of Microsoft, applicable individuals within Microsoft, and Upwork freelancers hired by Microsoft, but confidentiality will be maintained, as allowed by law.

- Your personal and study information will be stored for a period of up to 5 years.

- **How you can access and control your information.** If you wish to review or copy any personal information you provided during the study, or if you want us to delete or correct any such data, email your request to the research team at: suri@microsoft.com. Once your Upwork ID and Upwork display name are disassociated from your data we may not be able to remove your data from the study without re-identifying you.

For additional information or concerns about how Microsoft handles your personal information, please see the Microsoft Privacy Statement (https://privacy.microsoft.com/en-us/privacystatement).

BENEFITS AND RISKS

Benefits: There are no direct benefits to you that might reasonably be expected as a result of being in this study. The research team expects to learn how to best structure Upwork tasks from the results of this research. All clients at Microsoft will be able to learn from the conclusions of our study and create future tasks accordingly, which may benefit participants if they work for Microsoft clients in the future.

Risks: There are no anticipated, foreseeable risks or discomforts to you as a result of being in this study.

FUTURE USE OF YOUR IDENTIFIABLE INFORMATION

Identifiers might be removed from your identifiable private information, and after such removal, the information could be used for future research studies or distributed to another investigator for future research studies without your (or your legally authorized representative's) additional informed consent.

PAYMENT FOR PARTICIPATION

You will be paid to complete the task according to the payment structure indicated through Upwork. No other payment will be provided to take part in this study.

Your data may be used to make new products, tests or findings. These may have value and may be developed and owned by Microsoft and/or others. If this happens, there are no plans to pay you.

CONTACT INFORMATION

Should you have any questions concerning this project, please contact; David Holtz or Siddharth Suri, at t-daholt@microsoft.com or suri@microsoft.com.
Should you have any questions about your rights as a research subject, please contact Microsoft Research Ethics Program Feedback at MSRStudyfeedback@microsoft.com.

CONSENT

By accepting the task offer on Upwork.com, you confirm that the study was explained to you, you had a chance to ask questions before beginning the study, and all your questions were answered satisfactorily. By accepting the task offer on Upwork.com, you voluntarily consent to participate, and you do not give up any legal rights you have as a study participant.
This form will be attached to the job posting on Upwork.com. On behalf of Microsoft, we thank you for your contribution and look forward to your research session.