Detecting working patterns from online workers and predicting task completion

Janani Chitra Rajendran, Pritham Vinay Jujjavarapu, Sahithi Tatineni
jrajend@clemson.edu, pjujjav@g.clemson.edu, sahitht@g.clemson.edu

Background of the problem:

As digital platforms become integral to remote work, they generate extensive behavioral data. Analyzing this data is essential for enhancing productivity and managing resources efficiently. This project aims to address the challenge of predicting how long tasks will take to complete by examining user activities, including platform interactions, browser events, and task categories.

Project Objective:

- To identify patterns in user behavior through data analysis.
- To forecast task completion times accurately using machine learning techniques.
- To deliver practical insights that can aid in boosting productivity and guiding decision-making in remote work settings.

Methodological Paradigm:

Methodology : Regression Analysis
Models : Random Forest Regressor and XGBoost

Supporting rationale for the selected methodological paradigm:

- Regression analysis was selected due to its effectiveness in predicting continuous outcomes such as task completion durations.
- Random Forest was chosen for its capability to process both categorical and numerical data types effectively.
- XGBoost was chosen to address the limitations of Random Forest by offering better management of high-dimensional data, capturing complex patterns, and minimizing overfitting through regularization methods.

Summary of the dataset:

Unit of analysis: Individual user activities within crowd work platforms - each row represents a single action like loading a page, starting a task, or communicating within the platform.

Observations in the dataset: The dataset has 3,496,373 observations in total.

Unique Observations:

- Unique Users: 119
- Unique URLs visited: 255,518
- Unique Platforms: 5
- Unique Task Subtypes: 29
- Unique Browser Events: 18
- Unique Type Values: 10

Time Period Covered: The 'time' column indicates all data was collected in May 2020.

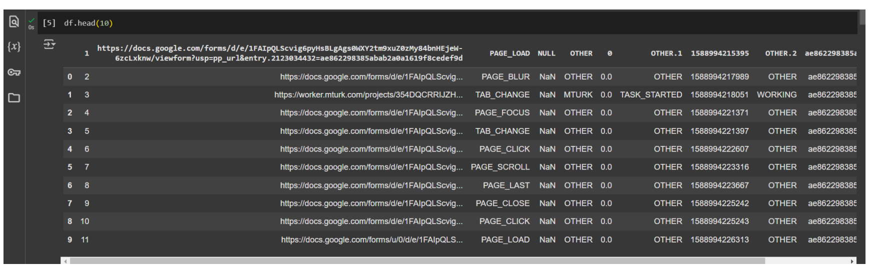
Summary of Data Cleaning:

Percentage of Samples with Missing Data: The dataset has a total of 3,496,373 observations, the percentage of missing data in the user column is approximately: 0.0075 %. This indicates a negligible amount of missing data.

The data cleaning steps performed include:

- **Assigning Column Names:** Initially, the dataset lacked column names, so names were assigned based on the provided data dictionary.
- **Data Type Conversion:** The ***time*** column, originally in milliseconds, was converted to a datetime format using ***pd.to_datetime*** for better temporal analysis.
- **Handling Missing Values:** The ***user*** column had 264 missing entries. Further exploration or consultation was considered to determine their impact.
- **Removing Irrelevant Columns:** Extraneous columns (extra and skip) were removed as per the data dictionary to streamline analysis.
- **Ensuring Consistency:** Checked and resolved inconsistencies in data formats and types across the dataset.

Before Data Cleaning:



```
[5] df.head(10)
```

	1	https://docs.google.com/forms/d/e/1FAIpQLScvig6pyHsBLgAgs0WXY2tm9xuZ0zMy84bnHEjeW-6zcLxknw/viewform?usp=pp_url&entry.2123034432=ae862298385abab2a0a1619f8cedef9d	PAGE_LOAD	NULL	OTHER	0	OTHER.1	1588994215395	OTHER.2	ae862298385a
0	2	https://docs.google.com/forms/d/e/1FAIpQLScvig...	PAGE_BLUR	NaN	OTHER	0.0	OTHER	1588994217989	OTHER	ae862298385
1	3	https://worker.mturk.com/projects/354DQCRRIJZH...	TAB_CHANGE	NaN	MTURK	0.0	TASK_STARTED	1588994218051	WORKING	ae862298385
2	4	https://docs.google.com/forms/d/e/1FAIpQLScvig...	PAGE_FOCUS	NaN	OTHER	0.0	OTHER	1588994221371	OTHER	ae862298385
3	5	https://docs.google.com/forms/d/e/1FAIpQLScvig...	TAB_CHANGE	NaN	OTHER	0.0	OTHER	1588994221397	OTHER	ae862298385
4	6	https://docs.google.com/forms/d/e/1FAIpQLScvig...	PAGE_CLICK	NaN	OTHER	0.0	OTHER	1588994222607	OTHER	ae862298385
5	7	https://docs.google.com/forms/d/e/1FAIpQLScvig...	PAGE_SCROLL	NaN	OTHER	0.0	OTHER	1588994223316	OTHER	ae862298385
6	8	https://docs.google.com/forms/d/e/1FAIpQLScvig...	PAGE_LAST	NaN	OTHER	0.0	OTHER	1588994223667	OTHER	ae862298385
7	9	https://docs.google.com/forms/d/e/1FAIpQLScvig...	PAGE_CLOSE	NaN	OTHER	0.0	OTHER	1588994225242	OTHER	ae862298385
8	10	https://docs.google.com/forms/d/e/1FAIpQLScvig...	PAGE_CLICK	NaN	OTHER	0.0	OTHER	1588994225243	OTHER	ae862298385
9	11	https://docs.google.com/forms/u/0/d/e/1FAIpQLS...	PAGE_LOAD	NaN	OTHER	0.0	OTHER	1588994226313	OTHER	ae862298385

After Data Cleaning:

```
[15] df.sample(10)
```

	ID	URL	browser_event	platform	task_subtype	time	type	user
115809	115811	https://worker.mturk.com/	PAGE_LOAD	MTURK	TASKS_LIST	2020-05-11 05:00:03.257	SEARCHING	ae862298385abab2a0a1619f8cedef9d
480080	480083	https://worker.mturk.com/projects/30B721SJLR5B...	PAGE_LOAD	MTURK	TASK_STARTED	2020-05-12 14:39:43.983	WORKING	4bc91376a0d164c5594c524788723d
2391138	2391141	https://worker.mturk.com/projects/3IE1FJZRAGMW...	PAGE_BLUR	MTURK	TASK_STARTED	2020-05-16 10:05:06.790	WORKING	2e259988387deea78754c4cde4cc8cc
2994509	2994512	https://worker.mturk.com/projects/3V7AFCJQIF5R...	PAGE_LOAD	MTURK	TASK_STARTED	2020-05-18 07:39:40.576	WORKING	bc42df9a3373d27be9780fd6ae128
1865275	1865278	https://worker.mturk.com/projects/3QPQKLUN3VBS...	PAGE_FOCUS	MTURK	TASK_STARTED	2020-05-15 06:42:45.287	WORKING	f0ac4982ad88a5b45e5923a71242145
2735936	2735939	https://www.youtube.com/embed/3ACiAMXT8Lk	PAGE_LOAD	OTHER	OTHER	2020-05-17 07:15:15.222	OTHER	ae1d53c23b4e45aa85e3b668bef64ab
1894097	1894100	https://worker.mturk.com/projects/3YJAOW5N9TAM...	PAGE_LAST	MTURK	TASK_STARTED	2020-05-15 08:05:56.812	WORKING	1164fceb602e16e44dc3f6e49ec51
1734700	1734703	https://www.mturkcontent.com/dynamic/hit?assig...	PAGE_SCROLL	MTURK	TASK_FRAME	2020-05-15 01:26:53.281	WORKING	c45ae0d5903111f7536eff13443b121f
692289	692292	https://worker.mturk.com/projects/3SI493PTTX86...	PAGE_LOAD	MTURK	TASK_SUBMITED	2020-05-12 21:56:21.281	WORKING	c35975cf36e8a25076da28a82ce70be
1952987	1952990	https://eu-u.openx.net/w/1.0/pd?plm=10&ph=8f5e...	PAGE_INACTIVITY	OTHER	OTHER	2020-05-15 11:46:29.835	OTHER	cabad7e3a75eb660956c8bf6cbf3c07

Handling of Missing Data:

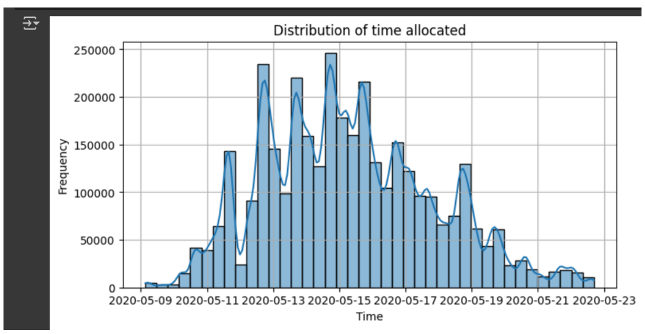
- Missing values in categorical variables were replaced with a placeholder value ('Unknown').

- Numerical features with missing values were imputed using median values to maintain data integrity.

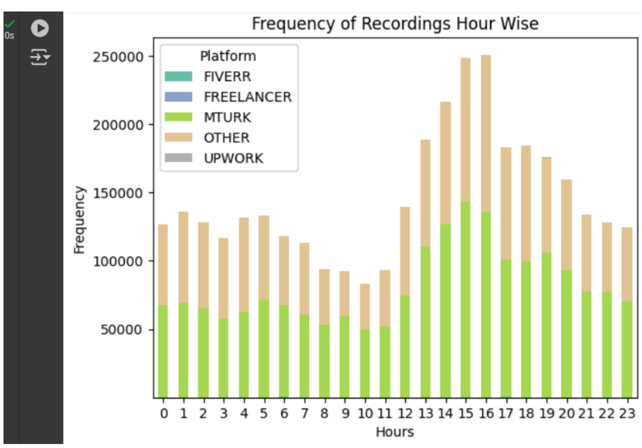
Outcome Variable Summary:

The outcome variable is the task completion time, measured in hours. It represents the duration required to finish tasks based on user behavior data, including platform usage, browser events, and task types. Predicting this variable helps in understanding productivity patterns and optimizing task management.

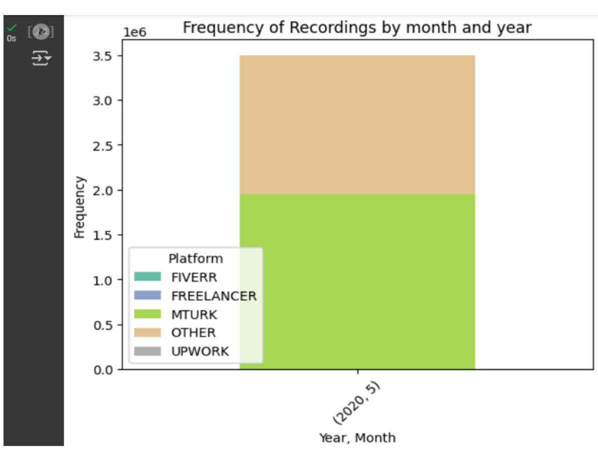
Outcomes of Appropriate Visualization technique:



The visualization highlights May 15, 2020, as a peak in time allocation and user activity. Users engaged more frequently and spent more time on the platform that day than on others.



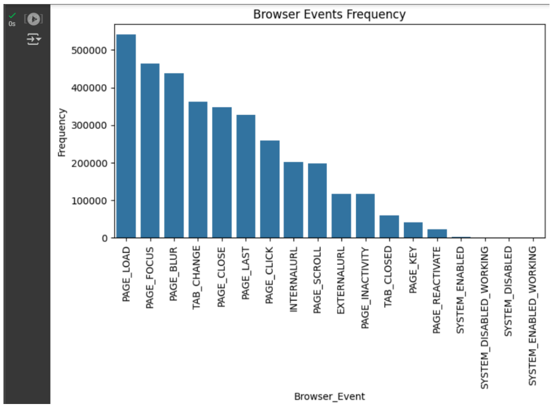
The second visualization represents the hourly distribution of user activity across different platforms, with a strong emphasis on MTurk and similar platforms. Distinct peaks at certain hours indicate periods of increased user engagement. This pattern suggests that users tend to prefer accessing these platforms at specific times, likely influenced by task availability, user schedules, or platform popularity. Analyzing these trends can help platform administrators and researchers determine optimal periods for task releases, improve user engagement strategies, and better understand platform dynamics.



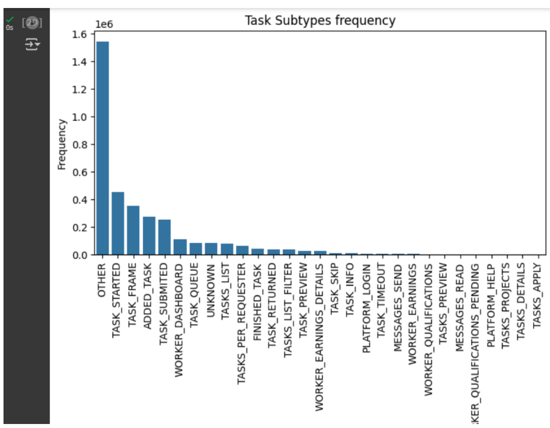
The third visualization showcases a high concentration of recorded activities solely within May 2020, with the majority of interactions occurring on two main platforms: MTurk and other similar platforms.

Key Predictors:

Based on our dataset analysis, we identify browser_event and platform as key predictors and they are structured.

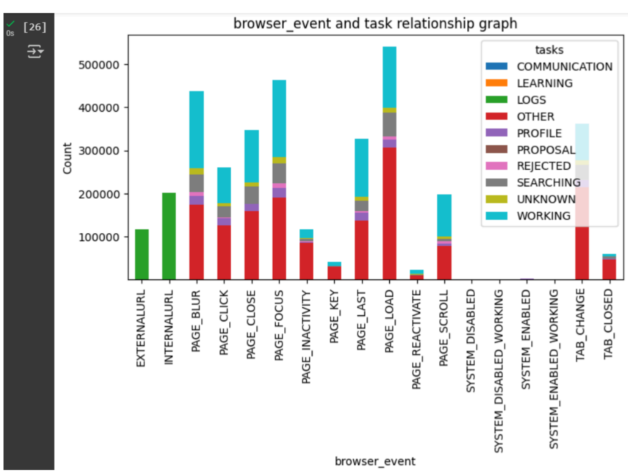


Browser Event: This captures the type of user interaction with the browser, which can influence task outcomes. For instance, a PAGE_LOAD event may signify the initiation of a new task, whereas a PAGE_BLUR event could indicate user distraction or a break. Examining the frequency and sequence of these interactions can help predict user behavior and task completion patterns.



Task Subtype: This refers to the specific type of task a user is engaged in, each requiring varying levels of time, effort, and skill. For instance, a TASK_STARTED event may suggest prolonged user engagement, while a TASK_RETURNED event could indicate difficulty in completing the task. Analyzing the frequency and sequence of these task subtypes can help predict user performance and task outcomes.

Relationship between browser_event and task_subtype:



Relationship between browser_event and task: PAGE_LOAD, PAGE_FOCUS, and PAGE_CLOSE indicate engagement, linking to WORKING and COMMUNICATION, while PAGE_INACTIVITY and PAGE_BLUR suggest disengagement, aligning with REJECTED and SEARCHING. Their sequence helps predict task outcomes.