Project Milestone 2: Progress Report

Project Title: Factors Driving Developer Influence and Efficiency in the AI Community
Date: October 26, 2023

1. **Progress on Research Questions**

Currently, we have completed the data extraction, cleaning, and preliminary statistical analysis for all three proposed research questions.

RQ 1: Contribution Volume vs. Influence (by Language)

Status: Completed.

Progress: We successfully aggregated user repository counts and correlated them with follower counts, segmented by the top 5 most common programming languages. Spearman's rank correlation was used to handle the non-normal distribution of follower counts.

RQ 2: Predictors of Popularity

Status: Completed.

Progress: A Linear Regression model was constructed and trained. We successfully extracted features including account tenure (account_age), activity level (repo_count), and primary programming language. The model coefficients have been interpreted to understand feature importance.

RQ 3: Factors Affecting Issue Resolution Time

Status: Preliminary Analysis Completed.

Progress: We successfully calculated the Time-to-Resolution (TTR) for closed issues. We overcame data linkage challenges to merge the Issues and Comments tables, allowing us to bin issues by engagement level. A Two-sample T-test was conducted to compare TTR between high and low engagement groups.

2. **Data Wrangling Techniques Applied**

To prepare the dataset for analysis, we applied several key data wrangling and feature engineering techniques using **Python (Pandas):**

Complex Merging: Performed an inner join between the Users and Repositories tables by extracting the username from the repository full_name string (e.g., splitting "owner/repo" to match the login column).

Issue-Comment Linkage: Merged the Issues table with aggregated data from the Comments table using id and pr_id as foreign keys to calculate engagement metrics.

Date & Time Standardization:

Converted all timestamp columns (created_at, closed_at) to UTC to prevent timezone-naive vs. timezone-aware subtraction errors.

Derived a new feature, account_age_days, by calculating the difference between a reference date (current timestamp) and the user's account creation date.

Calculated TTR_hours (Time-to-Resolution) by subtracting issue creation time from close time, filtering out negative or null values.

Handling Missing Data & Encoding:

Imputed missing values for the main_language feature with "Unknown".

Applied One-Hot Encoding to categorical language variables for the RQ2 regression model, grouping less common languages into an "Other" category to reduce dimensionality.

Implemented logic to clean duplicate columns (e.g., _x, _y suffixes) resulting from repeated merge operations.

3. Preliminary Findings/Results

RQ 1: Does creating more repositories lead to more followers?

Finding: No significant correlation. Our analysis using Spearman's rank correlation reveals that the volume of contributions (repository count) has a negligible relationship with user influence (followers) across all major languages.

Python: 0.0252

Java: 0.0340

JavaScript: -0.0111

HTML: -0.0152

Conclusion: Quantity does not equal popularity. Developers cannot simply increase their influence by creating a large volume of empty or low-quality repositories.

RQ 2: What are the strongest predictors of popularity?

Finding: Tech stack choice outweighs tenure and volume. The Linear Regression model provided distinct coefficients for different developer features.

Positive Drivers: The strongest predictors of having a higher follower count were the use of specific languages: Java (+4.27) and TypeScript (+3.21). repo_count had a positive but moderate impact (+3.72).

Negative Drivers: Users primarily associated with PHP (-13.7) and HTML (-6.3) tended to have fewer followers compared to the baseline.

Neutral: account_age_days had a coefficient near zero (+0.014), suggesting that merely having an old account does not guarantee influence.

RQ 3: Does higher engagement affect issue resolution time?

Finding: Engagement level does not significantly impact TTR. We tested the hypothesis that issues with higher engagement (high comment counts) would take longer to resolve due to complexity, or shorter due to attention.

Metric: Median comment count was 5.0.

Test Result: Two-sample T-test yielded a p-value of ~0.30 (t-statistic $\approx$ -1.04).

Conclusion: Since $p > 0.05$, we fail to reject the null hypothesis. There is no statistically significant difference in resolution time between high-engagement and low-engagement issues in this dataset.

4. **Link to GitHub Repository(https://github.com/RoxyLiu66/data-wrangling-group-8.git)**

5. **Remaining Work to be Completed**

With the core analysis code functioning, the remaining work focuses on refinement, visualization, and final reporting:

Refine RQ3 Analysis: Given the negative result for comment counts, we plan to investigate if Repository Stars (popularity of the repo) have a significant effect on TTR, as originally proposed in Milestone 1.

Advanced Visualization: Generate publication-ready visualizations (correlation heatmaps, regression coefficient bar charts, and TTR boxplots) to include in the final report.

Code Optimization: Refactor the Jupyter Notebooks to ensure reproducibility and add comprehensive markdown documentation.

Final Report: Synthesize all findings into the final project paper, focusing on the implications of these results for the AI development community.

In [ ]: