Supplementary Material

Appendix A: Data Description

Table A-1. Description of all 84 metrics that we gathered**.**

Metric Name	Description
Project Metadata (8)	
snapshot_date	Date of the first commit that the snapshot considers.
days_from_start	Number of days since the first commit.
proj_created_at	
company	1 if owner_company is not null or owner_type = 'ORG', otherwise 0.
total_file_size	Total size of all files in the project in bytes
bytes	Total size of all files written in the most-used language (excluding html)
num_languages	Number of languages used by the project.
language	One of the top 30 most popular languages. The lower the number, the more popular is the language.
Project Popularity Data (2):	
num_stars	Number of the stars the project has received
num_forks	
Commit Activity Data (7):	
num_commits	Number of commits.
first_commit_date	Timestamp of the first commit made to the repository.
end_commit_date	
months_committed	Number of months elapsed between the first commit timestamp and the last commit timestamp.
months_active	Number of months that the project is "active". I.e. There are commits or pull requests on the project.
commits_per_month	Frequency of commits, i.e. number of commits made per month.
num_pr_commits	Number of commits associated with pull requests.
Issue Activity Data (30):	
num_issues	Number of issues
issues_num_renames	Number of events in which an issue title was changed.
issues_num_labels	Number of events in which a label was added to an issue.
issues_num_add_to_project_board	
issues_num_moves_in_project_board_columns	
issues_num_create_from_project_board	

issues_num_remove_from_project_board	
issues_num_subscribes	
issues_num_unsubscribes	
issues_num_assigns	Number of events in which an issue was assigned to a user.
issues_num_unassigns	Number of events in which a user was unassigned from an issue.
issues_num_user_mentions	
issues_num_review_requests	
issues_num_locks	
issues_num_unlocks	
issues_num_references_in_commits	
issues_num_merges	
issues_num_closes	
issues_num_reopens	Number of events in which an issue was reopened.
issues_num_marks_as_duplicate	
issues_num_unmarks_as_duplicate	
issues_num_total_labels	Number of labels assigned to issues within the repository. Labels allow users to organize/prioritize their work. Users can create their own labels or use the default labels that GitHub provides (bug, duplicate, enhancement, good first issue, help wanted, invalid, question, wontfix). Note: For a description of each type of label, see: https://help.github.com/articles/about-labels/
issues_num_bug_labels	Number of labels on issues indicating bugs.
issues_num_duplicate_labels	Number of labels on issues indicating the existence of similar issues.
issues_num_enhancement_labels	Number of labels on issues indicating a new feature request
issues_num_good_first_issue_labels	
issues_num_help_wanted_labels	Number of labels on issues indicating that the maintainer wants help on the issue.
issues_num_invalid_labels	Number of labels on issues indicating that the issue is not relevant anymore.
issues_num_question_labels	Number of labels on issues indicating that the issue requires more information.
issues_num_wontfix_labels	Number of labels on issues indicating that work will not continue on the issue.
Pull Request Activity Data (19):	
prs_num_user_mentions	Note: For a description of each kind of event, see the following links:

	https://developer.github.com/v3/issues/events/#list-events-for-an-issue
prs_num_review_requests	
prs_num_locks	
prs_num_unlocks	
prs_num_references_in_commits	Number of events in which a pull request was referenced in a commit
num_prs	Number of pull requests that are made to the repository.
prs_num_opens	
prs_num_merges	
prs_num_closes	
prs_num_reopens	Number of events in which a pull request was reopened.
prs_num_synchronizes	Number of events in which commits are added to or removed from the project repository.
pr_num_total_labels	Note: For a description of each type of label, see: https://help.github.com/articles/about-labels/
pr_num_bug_labels	
pr_num_duplicate_labels	
pr_num_enhancement_labels	
pr_num_help_wanted_labels	
pr_num_invalid_labels	
pr_num_question_labels	
pr_num_wontfix_labels	
Team Members Data (9):	
num_committers	
num_pr_mergers	Number of users who can merge pull requests on the repository.
team_size	num_committers + num_pr_mergers
num_commenters	Number of users who commented on the project within discussions or code reviews.
avg_months_committed	Avg number of months that each team member has committed to the project.
avg_months_merged	Average number of months that each team member has merged pull requests on the project.
avg_months_active	Average number of months that each team member has been "active" on the project, i.e. has committed or merged pull requests on the project.
avg_commits_per_month	Average number of commits that each team member has made per month.
avg_prs_merged_per_month	Average number of pull requests that each team member has merged per month.
Experience of Developers on the Project(5)	
std_dev_months_committed	Standard deviation of number of months that each team member has committed

std_dev_months_merged	Standard deviation of the number of months that each team member has merged pull requests on the project.
std_dev_months_active	Standard deviation of the number of months that each team member has been "active" on the project.
std_dev_commits_per_month	Standard deviation of the number of commits that each team member has made per month.
std_dev_prs_merged_per_month	Standard deviation of the number of pull requests that each team member has merged per month.
Discussion Data (4):	
num_issue_disc	Number of discussion comments made on issues.
num_pr_disc	Number of discussion comments made on pull requests.
num_pr_code_rev	Number of code review comments made on commits associated with pull requests.
num_commit_code_rev	Number of code review comments made on commits not associated with pull requests.

Table A-2: The reduced set of 30 metrics used for modeling

Metric name	Description
Project Metadata (2)	
language	One of the top 30 most popular languages. The lower the number, the more popular is the language.
company	1 if owner_company is not null or owner_type = 'ORG', otherwise 0.
Project Popularity Data (1)	
num_stars	Number of stars the project has received
Team Members Data (5)	
avg_commits_per_month	Average number of commits that each team member has made per active month.
avg_months_committed	Average number of months that each team member has committed to the project.
avg_months_merged	Average number of months that each pull request merger has merged pull requests on the project.
avg_prs_merged_per_month	Average number of pull requests that each pull request merger has merged per month.
team_size	Number of committers + number of pull request mergers
Commit Activity Data (3)	
num_commits	Number of commits
months_committed	Number of months elapsed between the first commit timestamp and the last commit timestamp
months_active	Number of months that the project is "active". I.e. There are commits or pull requests on the project.
Issue Activity Data(10)	

num_issues	Number of issues
issues_num_duplicate_labels	Number of labels on issues indicating the existence of similar issues.
issues_num_assigns	Number of events in which an issue was assigned to a user.
issues_num_labels	Number of events in which a label was added to an issue.
issues_num_wontfix_labels	Number of labels on issues indicating that work will not continue on the issue.
issues_num_unassigns	Number of events in which a user was unassigned from an issue
issues_num_enhancement_labels	Number of labels on issues indicating a new feature request
issues_num_help_wanted_labels	Number of labels on issues indicating that the maintainer wants help on the issue.
issues_num_invalid_labels	Number of labels on issues indicating that the issue is not relevant anymore.
issues_num_question_labels	Number of labels on issues indicating that the issue requires more information.
Pull Request Activity Data (3)	
prs_num_synchronizes	Number of events in which commits are added to or removed from the project repository.
prs_num_reopens	Number of events in which a pull request was reopened.
prs_num_references_in_commits	Number of events in which a pull request was referenced in a commit
Discussion Data(2)	
num_commit_code_rev	Number of code review comments made on commits not associated with pull requests
num_pr_code_rev	Number of code review comments made on commits associated with pull requests.

Experience of Developers on the Project(4)	
std_dev_months_merged	Standard deviation of number of months that each team member has merged pull requests on the project
std_dev_commits_per_month	Standard deviation of number of commits that each team member has made per month
std_dev_months_committed	Standard deviation of number of months that each team member has committed
std_dev_prs_merged_per_month	Standard deviation number of pull requests that each team member has merged per months

Table A-3: 16 metrics found to be most important for accurate modeling

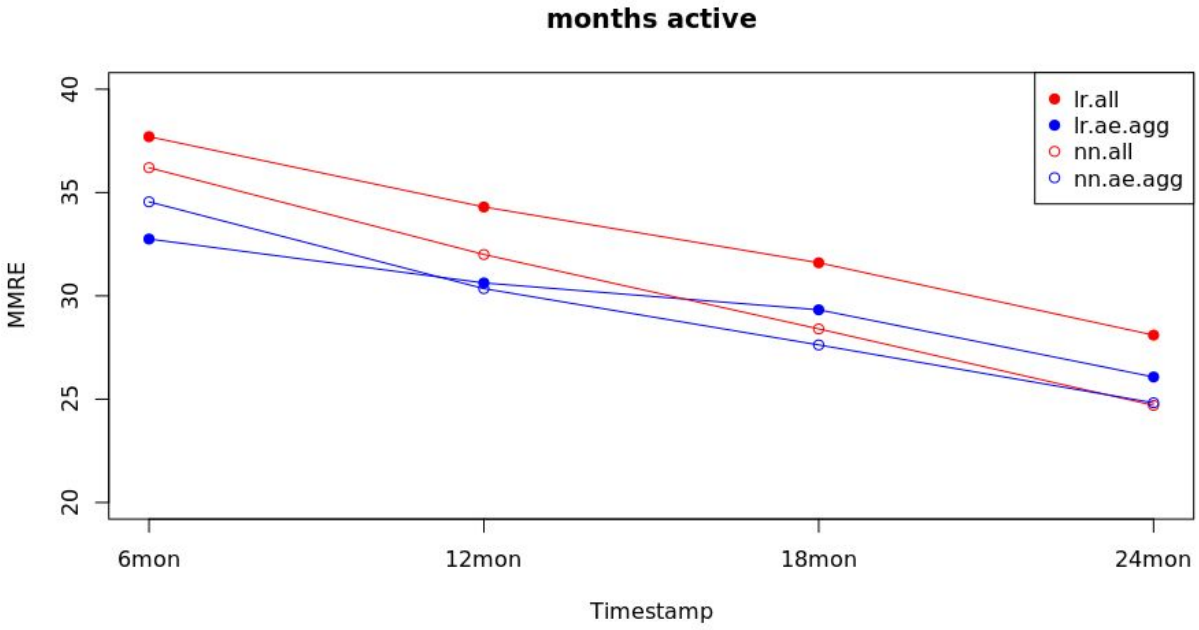
"avg_commits_per_month"
"avg_months_committed"
"avg_months_merged"
"avg_prs_merged_per_month"
"team_size"
"months_active"
"issues_num_duplicate_labels"
“issues_num_assigns"
"issues_num_labels"
"issues_num_wontfix_labels"
"num_commit_code_rev"
"num_issues"
"prs_num_references_in_commits"
"std_dev_months_committed"
"std_dev_commits_per_month"
"std_dev_months_merged"

Appendix B: Full results from our hybrid models (Autoencoder clustering + MLR)

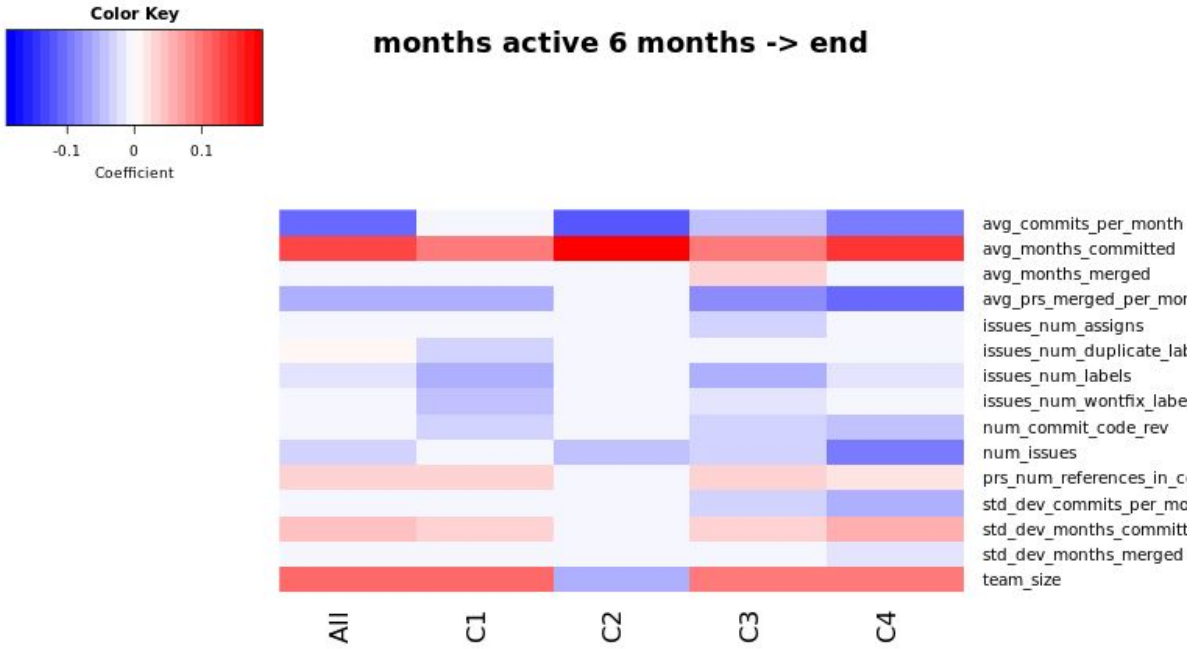
Below,

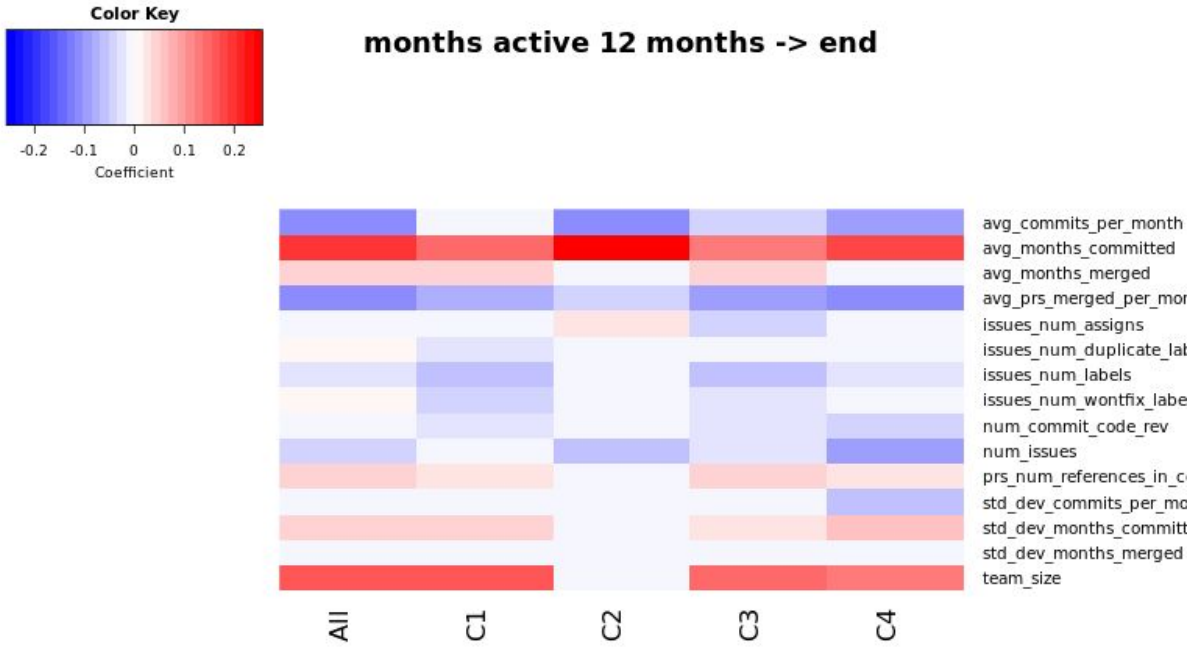
- lr = multiple linear regression,
- nn = neural network model,
- X.all = modeling on all data, using model X (lr or nn)
- X.ae.agg = modeling per cluster, using model X (lr or nn), followed by aggregating (averaging) over all clusters
- X -> end, means predicting the outcome of the final snapshot using snapshot X

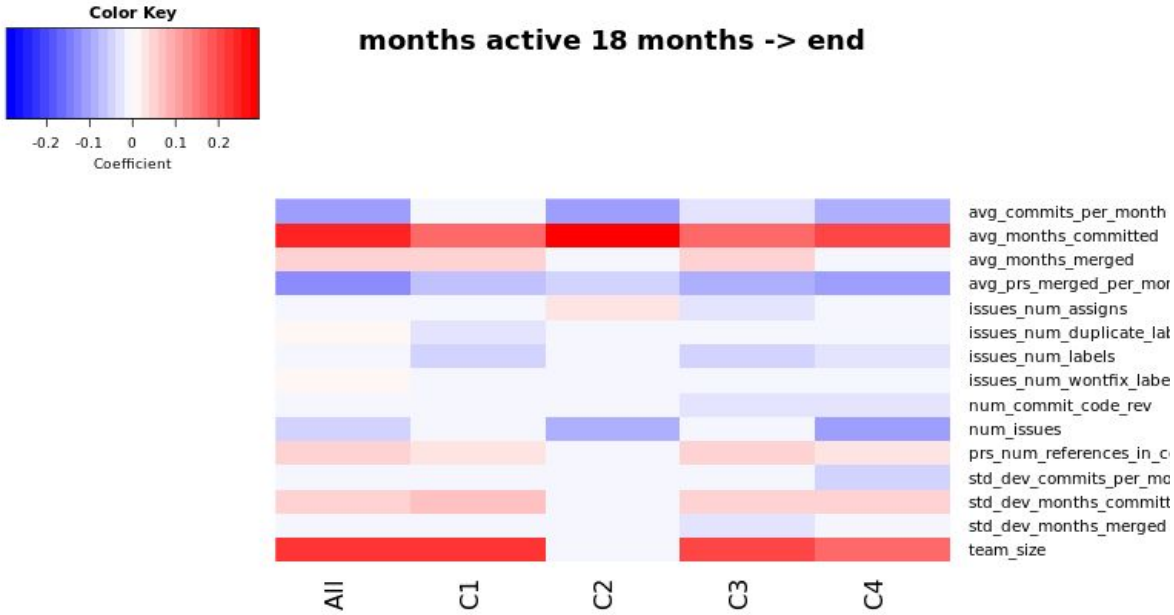
months_active: Forecasting accuracy

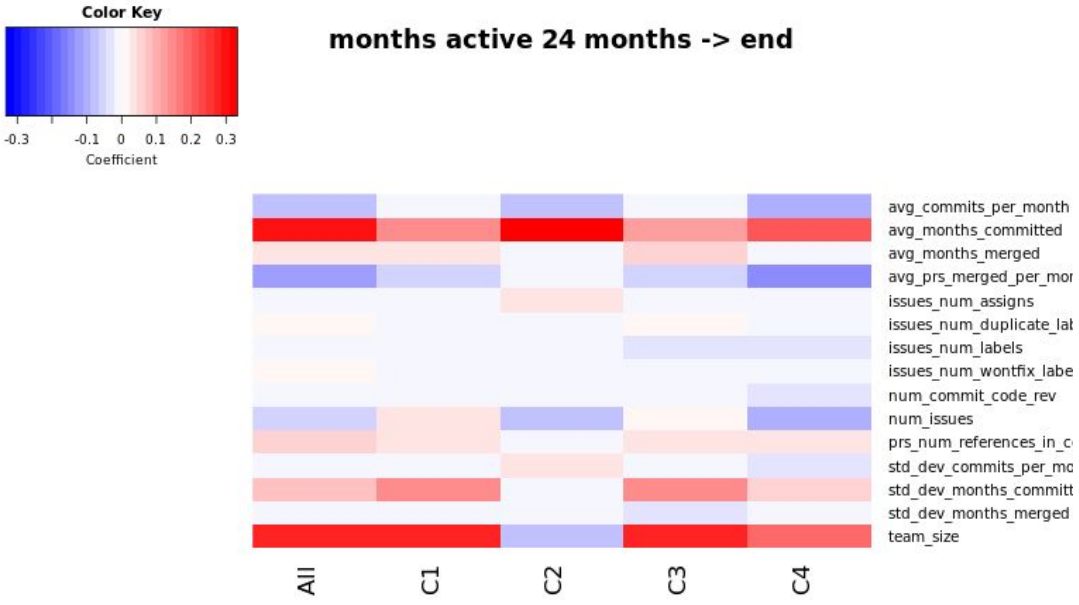


months_active: Feature Importance Maps from the Autoencoder+MLR models

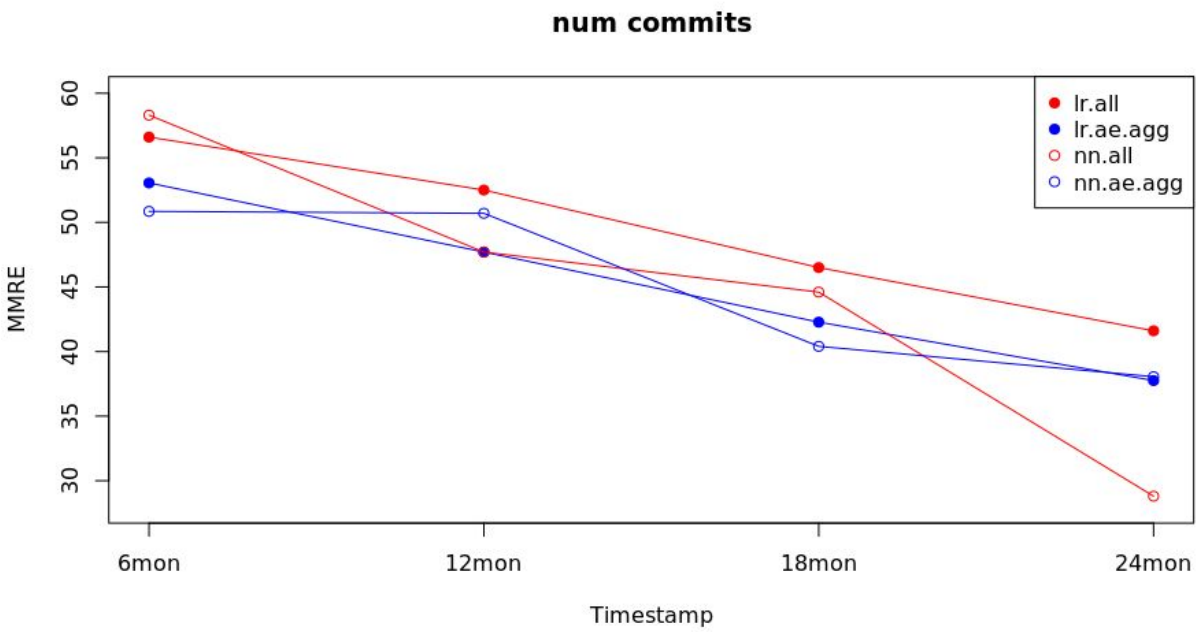




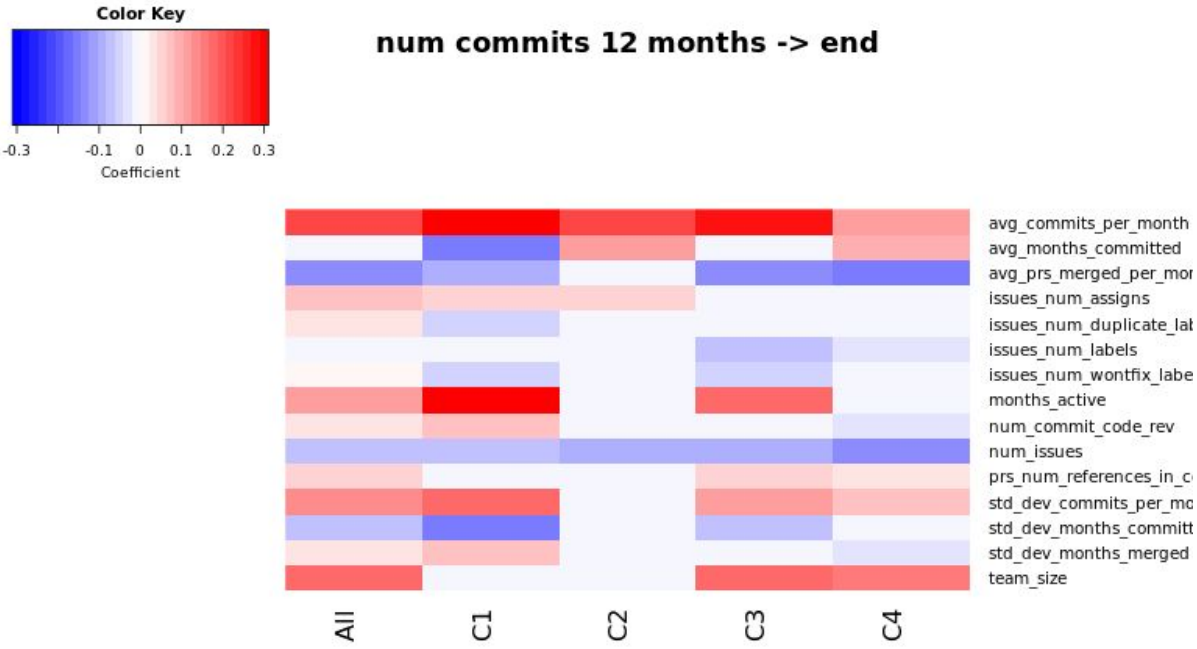

Color Key
-0.2 -0.1 0 0.1 0.2
Coefficient
months active 18 months -> end
avg_commits_per_month
avg_months_committed
avg_months_merged
avg_prs_merged_per_mor
issues_num_assigns
issues_num_duplicate_lal
issues_num_labels
issues_num_wontfix_labe
num_commit_code_rev
num_issues
prs_num_references_in_c
std_dev_commits_per_mo
std_dev_months_committ
std_dev_months_merged
team_size
All
C1
C2
C3
C4

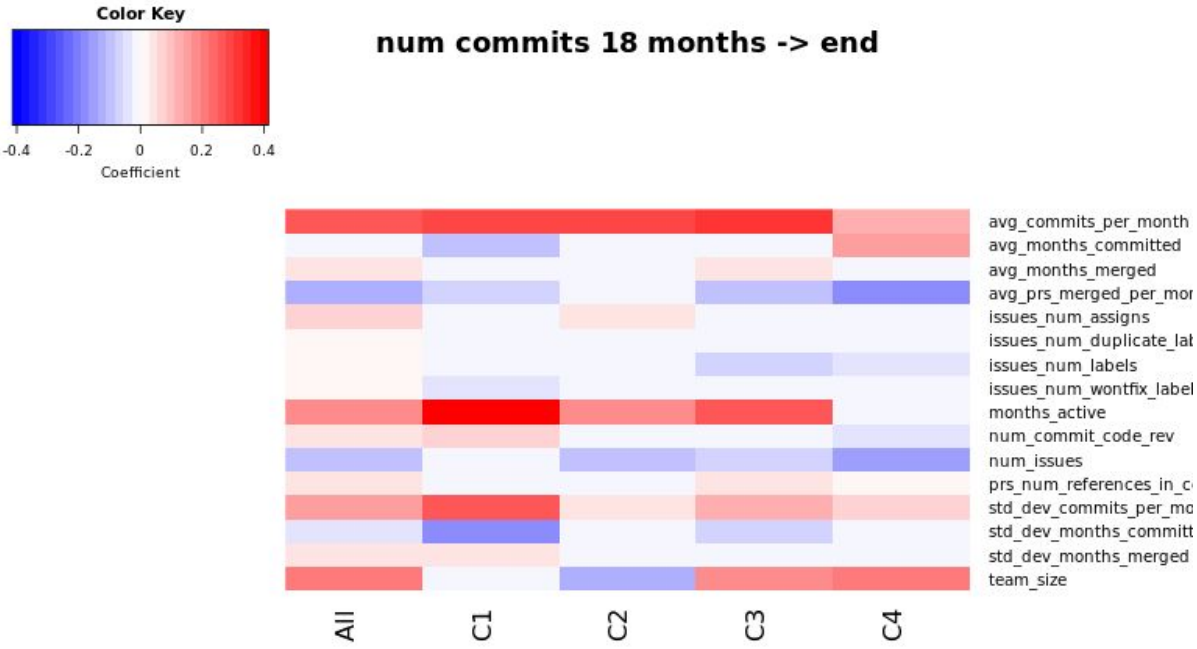

Color Key
-0.3 -0.1 0 0.1 0.2 0.3
Coefficient
months active 24 months -> end
avg_commits_per_month
avg_months_committed
avg_months_merged
avg_prs_merged_per_mor
issues_num_assigns
issues_num_duplicate_lal
issues_num_labels
issues_num_wontfix_labe
num_commit_code_rev
num_issues
prs_num_references_in_c
std_dev_commits_per_mo
std_dev_months_committ
std_dev_months_merged
team_size
All
C1
C2
C3
C4

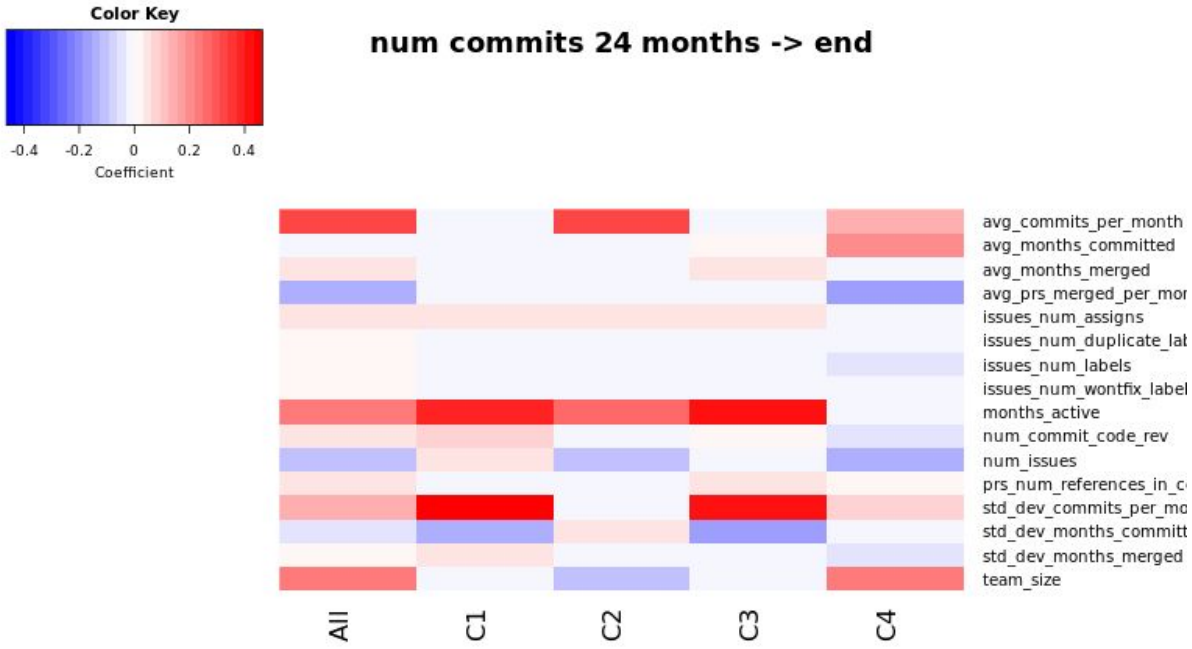
num_commits: Forecasting accuracy



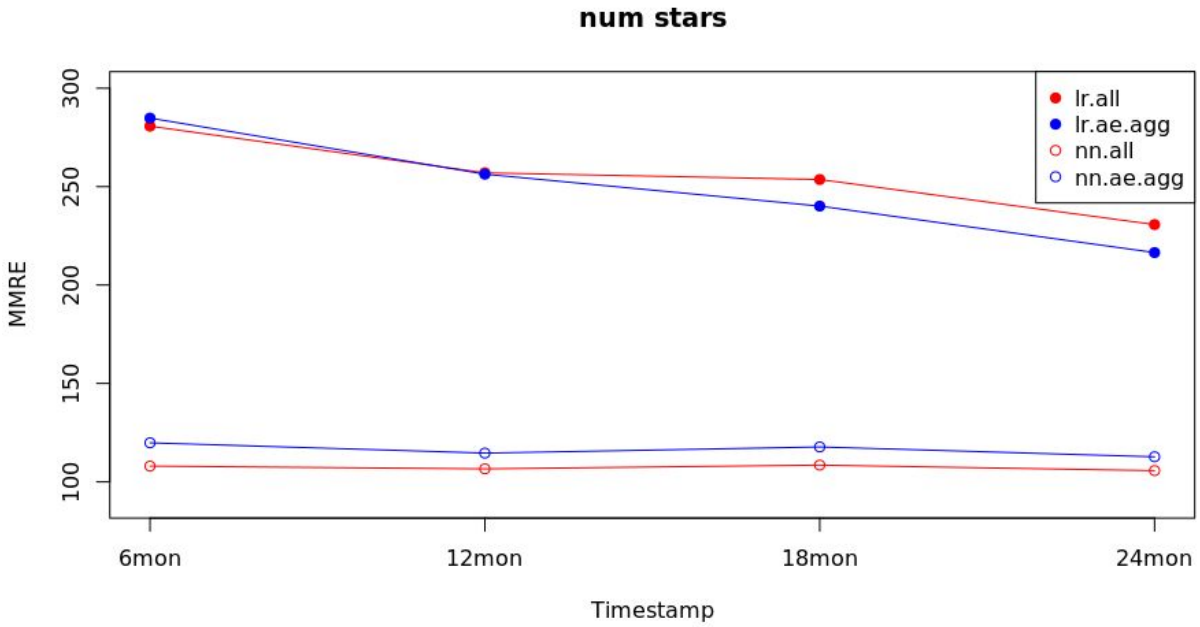
num_commits: Feature Importance Maps from the Autoencoder+MLR models


Color Key
-0.3
-0.1
0
0.1
0.2
0.3
Coefficient
num commits 12 months -> end
avg_commits_per_month
avg_months_committed
avg_prs_merged_per_mor
issues_num_assigns
issues_num_duplicate_lat
issues_num_labels
issues_num_wontfix_labe
months_active
num_commit_code_rev
num_issues
prs_num_references_in_c
std_dev_commits_per_mo
std_dev_months_committ
std_dev_months_merged
team_size
All
C1
C2
C3
C4

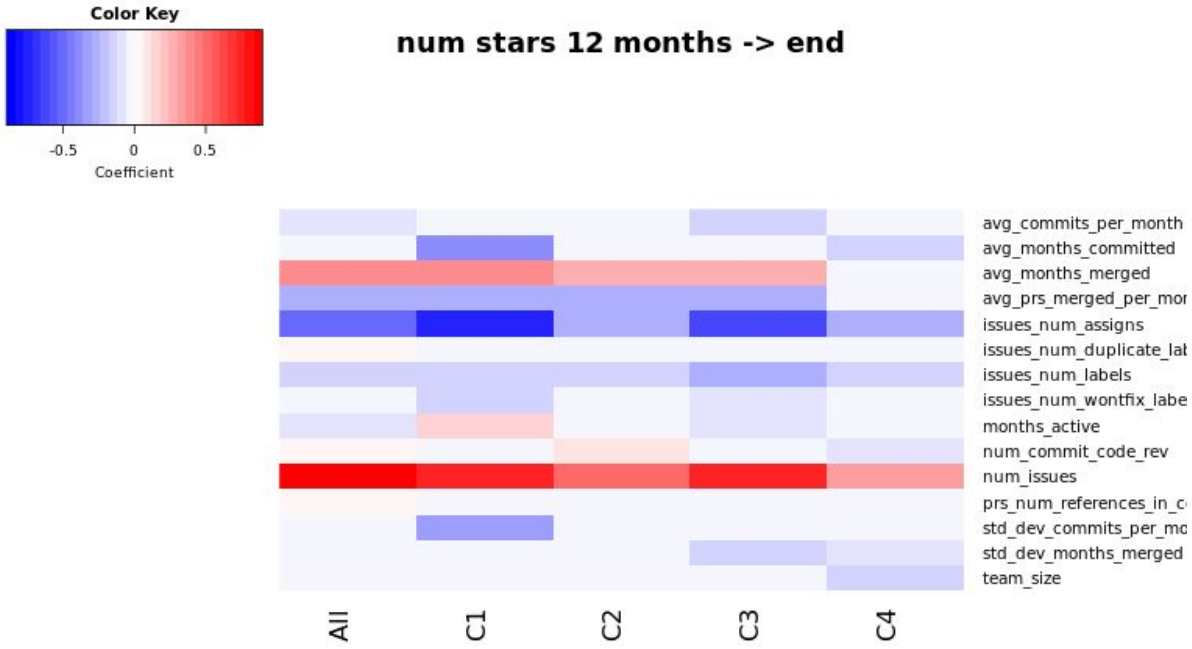

Color Key
-0.4
-0.2
0
0.2
0.4
Coefficient
num commits 18 months -> end
avg_commits_per_month
avg_months_committed
avg_months_merged
avg_prs_merged_per_mor
issues_num_assigns
issues_num_duplicate_lat
issues_num_labels
issues_num_wontfix_label
months_active
num_commit_code_rev
num_issues
prs_num_references_in_c
std_dev_commits_per_mo
std_dev_months_committ
std_dev_months_merged
team_size
All
C1
C2
C3
C4

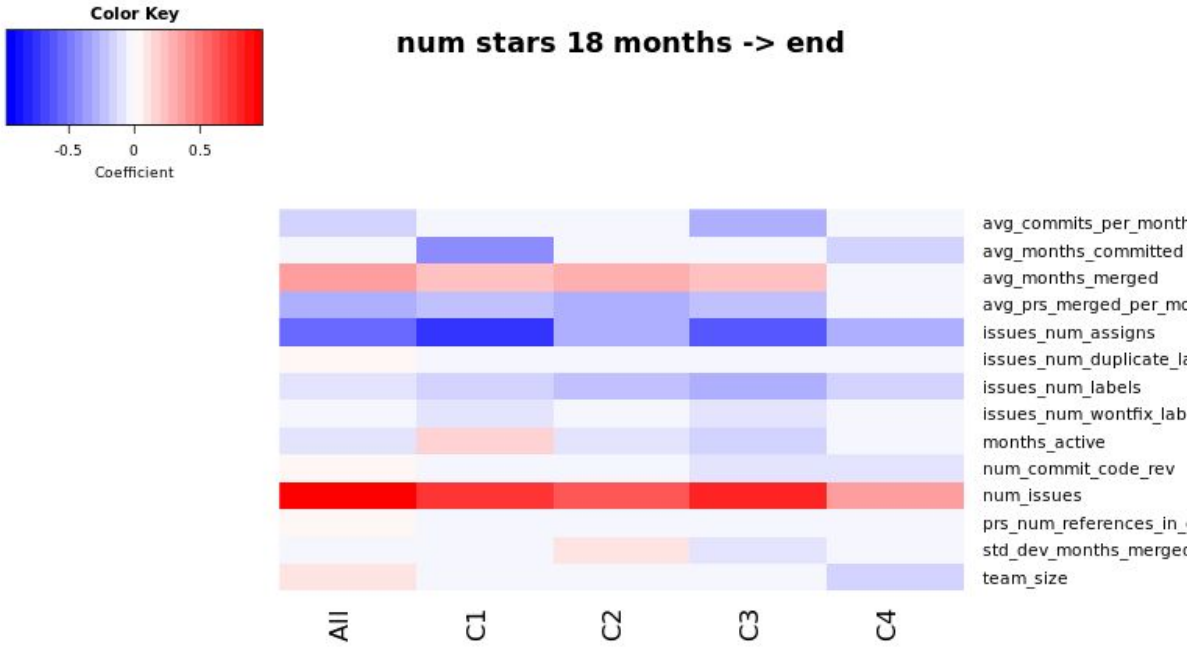

Color Key
-0.4
-0.2
0
0.2
0.4
Coefficient
num commits 24 months -> end
avg_commits_per_month
avg_months_committed
avg_months_merged
avg_prs_merged_per_mor
issues_num_assigns
issues_num_duplicate_lat
issues_num_labels
issues_num_wontfix_label
months_active
num_commit_code_rev
num_issues
prs_num_references_in_c
std_dev_commits_per_mo
std_dev_months_committ
std_dev_months_merged
team_size
All
C1
C2
C3
C4

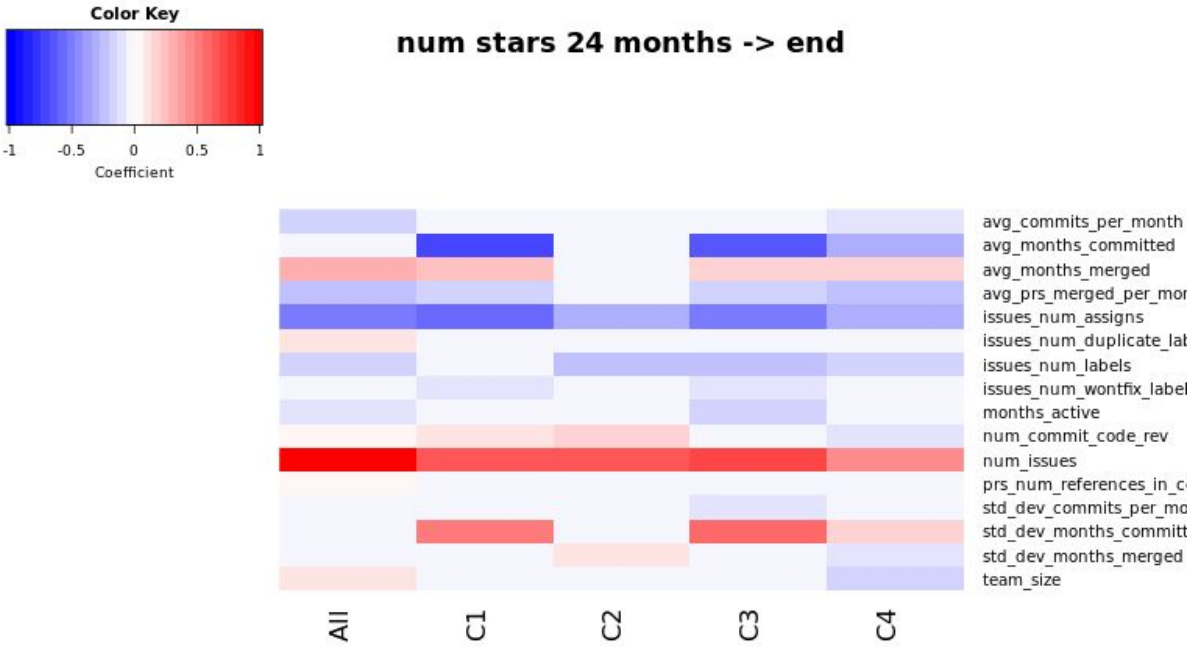
num_stars: Forecasting accuracy



num_stars: Feature Importance Maps from the Autoencoder+MLR models




Color Key
-0.5
0
0.5
Coefficient
num stars 18 months -> end
avg_commits_per_month
avg_months_committed
avg_months_merged
avg_prs_merged_per_mo
issues_num_assigns
issues_num_duplicate_la
issues_num_labels
issues_num_wontfix_lab
months_active
num_commit_code_rev
num_issues
prs_num_references_in_
std_dev_months_merged
team_size
All
C1
C2
C3
C4


Color Key
-1
-0.5
0
0.5
1
Coefficient
num stars 24 months -> end
avg_commits_per_month
avg_months_committed
avg_months_merged
avg_prs_merged_per_mor
issues_num_assigns
issues_num_duplicate_lab
issues_num_labels
issues_num_wontfix_label
months_active
num_commit_code_rev
num_issues
prs_num_references_in_c
std_dev_commits_per_mo
std_dev_months_committ
std_dev_months_merged
team_size
All
C1
C2
C3
C4