Predicting Likelihood of Requirement Implementation within the Planned Iteration: An Empirical Study at IBM

Ali Dehghan, Adam Neal, Kelly Blincoe, Johan Linaker, Daniela Damian

What is the problem?

Some requirements take longer than planned

Product managers would like to detect them ahead of time

Why it matters?

Helps them in (re)planning and (re)allocation of resources.

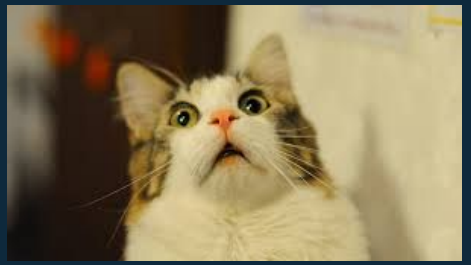
What has been done?

Prediction of fix time and fix effort of defects

Prediction of release readiness

But ...

> there's a lack of focus on high-level requirements



Research Questions

RQ1: Can we predict whether or not a requirement will be completed within the planned iteration?

RQ2: Can we optimize the predictive model to maximize precision of predictions, while maintaining an acceptable recall?

RQ3: What are the features that can be used in this prediction and how important are they relatively?

IBM Enterprise Platform

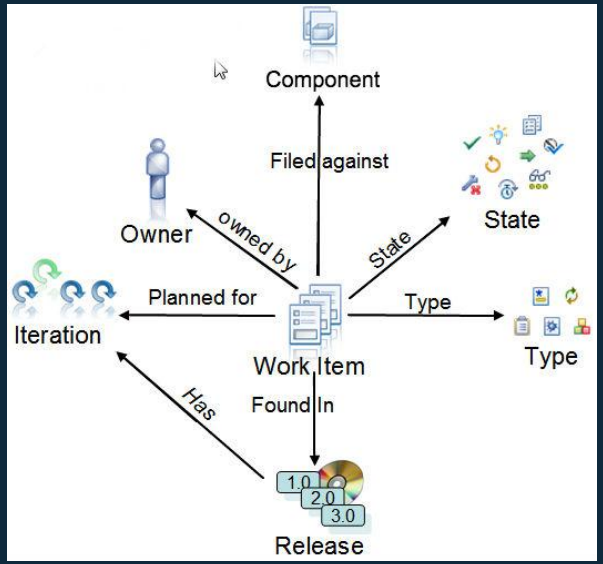
Includes a variety of products and a community of stakeholders

Work item: Represent a work that should be done

Histories: represent historical records of work items

Iteration: A time-box for development

Stakeholder types: Creator, Owner, Resolver, Commenter and Subscriber



Work item Hierarchy

The ideal hierarchy of workitems:

Plan Item 1->* **Story** 1- >* **Enhancement** 1->* **Task**

Plan item / Stories: Represent high-level requirements and functionalities

A work item of type Defect can be a child of any other work items

Methodology

Review of prior work (bug resolution time prediction and bug triaging)

A mixed-method empirical study:

- qualitative: Interviews with IBM analysts and developers
- quantitative: Machine learning techniques

18 semi-structured (2 face-to-face and 16 remote) interviews

- Define research goals and settings
- Choose 3 projects for the study
- Understanding the IBM ecosystem
- Understanding the data
- Iterative feature engineering

Feature Engineering

Engineered a set of 29 features

Each feature comes from one or multiple of:

- Prior work
- Suggestions made by IBM practitioners
- Domain knowledge achieved as a result of interviews

Feature engineering iteratively involved:

- Defining factors impacting completion time (complexity, priority, etc.)
- Data visualization
- Proposing a feature (e.g. by brainstorming, prior work)
- Running statistical tests on candidate features

Project Stats

Att. / Project	A	B	C
Start date	Jun 2006	Jan 2009	Jun 2006
End date	Oct 2016	Oct 2016	Oct 2016
# work items	75k	71k	177k
# histories	816k	835k	1.73m
# plan items	839	749	447
# stories	1,286	3,640	4,471
# comments	374k	312k	312k
#developers	594	481	796

Model Training – Settings

Predicting at 4 different stages:

- Day 1, and the end of 1st, 2nd and 3rd quarters of an iteration
- Results in 24 datasets (3 projects, 2 WI types, 4 stages)

Requirement selection criteria:

- Is already completed
- Is planned for an iteration with an end date
- For each requirement, the last history before the prediction stage is selected

A binary classification problem (NO is the positive class)

$$iteration\_met = \begin{cases} YES, & \text{if } completion\_date \leq end\_date \\ NO, & \text{otherwise} \end{cases}$$

Model Training – Techniques

Random Forest as the learning algorithm

Cost-insensitive learning for RQ1

Cost-sensitive learning to address RQ2 and RQ3

- Define a high penalty for False Positives
- Calculated based on
  - A variable penalty based on balance of the dataset
  - A constant penalty for favoring precision over recall

Feature importance calculation:

- For each dataset, in 29 iterations, exclude one feature and calculate WA
- Averaging importance of features across 24 datasets

Model Evaluation

Performance metrics:

- Precision, Recall and F-measure for RQ1
- Weighted Average for RQ2 and RQ3

$$WA = (3 * precision(NO) + recall(NO))/4$$

dataset	NO%			
	0th	1st	2nd	3rd
A - plan	0.53	0.44	0.49	0.49
B - plan	0.30	0.54	0.52	0.32
C - plan	0.35	0.38	0.39	0.39
A - story	0.50	0.38	0.40	0.40
B - story	0.44	0.20	0.19	0.18
C - story	0.50	0.37	0.36	0.36

Leave-One-Out Cross Validation for quantitative validation

Model Features

- Engineered a set of 29 features
- Features classified into 9 logical categories
- Categories are either:
  - Based on factors impacting completion time
  - Representatives of feature families
- Categories meant to guide further studies
- Each feature in the most relevant category

Feature Categories

1. General Features
2. Complexity Indicating Features
3. Progress Implying Features
4. Priority Implying Features
5. Problem Change Indicating Features
6. Process Change Indicating Features
7. Stakeholder Characteristics
8. Stakeholder Communication
9. Child Features

1- General Features

Feature	Type	Range	Cardinality	Missing
Creator ID	Nominal	-	15%	0
Creation Month	Nominal	-	2%	0
Owner ID	Nominal	-	10%	15%

2- Complexity Indicating Features

Feature	Type	Range	Cardinality	Missing
Subscriber Count	Integer	0 - 250	-	0
Filed Against (Component)	Nominal	-	15%	0
Iteration Change Count	Integer	0 - 12	-	0

3- Progress Implying Features

Feature	Type	Range	Cardinality	Missing
Iteration Remained Days	Integer	0 - 365	-	0
Days Since Creation	Integer	0 - 1400	-	0
Status	Nominal	-	1%	0

4- Priority Implying Features

Feature	Type	Range	Cardinality	Missing
Priority	Nominal	-	1%	40%
Severity	Nominal	-	1%	30%
Days Before First Assignment	Integer	0 - 1000	-	0
Days Since Last Comment	Integer	0 - 500	-	0
Owner Change Count	Integer	0 - 8	-	0
Days Since Last Assignment	Integer	0 - 500	-	0

5- Problem Change Indicating Features

Feature	Type	Range	Cardinality	Missing
Summary Change Count	Integer	0 - 9	-	0
Description Change Count	Integer	0 - 45	-	0
Days Since Last Summary	Integer	0 - 360	-	0
Days Since Last Description	Integer	0 - 260	-	0

6- Process Change Indicating Features

Feature	Type	Range	Cardinality	Missing
Days Since Requirement Type Change	Integer	0 - 260	-	0
Days Since Child to Parent Type Change	Integer	0 - 400	-	0

7- Stakeholder Characteristics

Feature	Type	Range	Cardinality	Missing
Days Since Last DE Comment	Integer	0 - 500	-	0
Count of Same Component Workitems Resolved by Owner	Integer	0 - 900	-	15%
Creator-Team Relationship	Nominal	-	3	0

8- Stakeholder Communication

Feature	Type	Range	Cardinality	Missing
Comment Count	Integer	0 - 70	-	0
Commenter Count	Integer	0 - 20	-	0

9- Children Related Features

Feature	Type	Range	Cardinality	Missing
Same Type Child Count (status = New)	Integer	0 - 6	-	0
Large Size Child Count (status = New)	Integer	0 - 15	-	0
Medium Size Child Count (status = New)	Integer	0 - 60	-	0

Res. – Cost Insensitive Model (RQ1)

stage	cnt	no%	prec.	rec.	f1	wa	cnt	no%	prec.	rec.	f1	wa	cnt	no%	prec.	rec.	f1	wa
	Project A - Plan Items						Project B - Plan Items						Project C - Plan Items					
0th	316	.53	.76	.69	.72	.74	231	.30	.69	.47	.56	.63	77	.35	.67	.24	.35	.56
1st	393	.44	.82	.70	.76	.79	224	.54	.63	.67	.65	.64	267	.38	.60	.58	.59	.60
2nd	462	.49	.84	.73	.78	.81	287	.52	.68	.70	.69	.68	280	.39	.67	.56	.61	.64
3rd	468	.49	.84	.73	.78	.81	473	.32	.70	.46	.56	.64	287	.39	.63	.51	.57	.60
	Project A - Stories						Project B - Stories						Project C - Stories					
0th	521	.50	.73	.71	.72	.72	2020	.44	.69	.65	.67	.68	1644	.50	.71	.71	.71	.71
1st	769	.38	.74	.58	.65	.70	2472	.20	.81	.51	.63	.74	1999	.37	.66	.52	.58	.62
2nd	842	.40	.73	.59	.65	.69	2759	.19	.85	.54	.66	.77	2310	.36	.68	.49	.57	.63
3rd	873	.40	.74	.62	.67	.71	2925	.18	.84	.55	.66	.76	2422	.36	.69	.50	.58	.64

Res. – Cost Sensitive Model (RQ2)

stage	cnt	no%	prec.	rec.	wa	cnt	no%	prec.	rec.	wa	cnt	no%	prec.	rec.	wa
	Project A - Plan Items					Project B - Plan Items					Project C - Plan Items				
0th	316	.53	.92	.56	.83	231	.30	.94	.24	.77	77	.35	1.0	.04	.76
1st	393	.44	.90	.50	.80	224	.54	.85	.28	.70	267	.38	.76	.28	.64
2nd	462	.49	.93	.56	.83	287	.52	.86	.28	.71	280	.39	.80	.15	.64
3rd	468	.49	.91	.56	.82	473	.32	.94	.11	.73	287	.39	.83	.18	.67
	Project A - Stories					Project B - Stories					Project C - Stories				
0th	521	.50	.88	.34	.74	2020	.44	.85	.30	.71	1644	.50	.85	.34	.72
1st	769	.38	.88	.29	.73	2472	.20	.97	.23	.79	1999	.37	.84	.16	.67
2nd	842	.40	.87	.31	.73	2759	.19	.94	.30	.78	2310	.36	.84	.16	.67
3rd	873	.40	.86	.29	.71	2925	.18	.95	.30	.79	2422	.36	.88	.17	.70

Res. – Feature Importance (RQ3)

Rank by frequency of being among top 10	cnt
iteration_days_remained	17
creator_identifier	15
days_without_owner	12
iteration_change_count	11
filed_against	11
creation_month	11
large_size_child_count_new	10
days_since_last_summary	10
summary_change_count	10
subscriber_count	10
component_resolver	9
comment_count	9
owner_change_count	8
days_since_last_type_s_p	8
creator_team_relationship	8
medium_size_child_count_new	7
commenter_count	7
owner_identifier	7
status	7
days_since_last_owner	7
days_since_last_description	7
severity	7
days_since_last_comment	6
days_since_last_de_comment	6
description_change_count	6
same_type_child_count_new	5
days_since_last_type_child	4
priority	4
days_since_creation	4

Rank by frequency of being among top 5	cnt
creator_identifier	13
iteration_days_remained	12
days_since_last_summary	8
iteration_change_count	7
filed_against	6
creation_month	6
days_without_owner	5
status	5
large_size_child_count_new	4
days_since_last_description	4
days_since_creation	4
days_since_last_comment	4
subscriber_count	4
owner_change_count	4
owner_identifier	4
creator_team_relationship	4
component_resolver	3
comment_count	3
description_change_count	3
days_since_last_owner	3
severity	3
medium_size_child_count_new	2
days_since_last_de_comment	2
priority	2
commenter_count	2
summary_change_count	2
same_type_child_count_new	1
days_since_last_type_s_p	0
days_since_last_type_child	0

Rank by average rank	avg
iteration_days_remained	8.7
creator_identifier	10.2
iteration_change_count	12.5
days_since_last_summary	12.8
creation_month	12.9
component_resolver	13.3
days_without_owner	13.4
subscriber_count	13.5
large_size_child_count_new	13.9
comment_count	14
days_since_last_description	14.1
description_change_count	14.2
days_since_last_owner	14.3
days_since_last_type_s_p	14.4
status	14.6
creator_team_relationship	14.6
filed_against	14.9
owner_change_count	15
summary_change_count	15.6
severity	15.7
commenter_count	16.1
medium_size_child_count_new	16.6
days_since_last_de_comment	16.7
days_since_last_comment	17.1
days_since_last_type_child	17.3
priority	17.5
same_type_child_count_new	18.4
days_since_creation	18
owner_identifier	18.9

Discussion

- We showed that effective predictive models could be trained to estimate completion time of a high-level requirement
- We showed that there are certain features that are often of high importance and high-level abstractions of features can be introduced to make them applicable to data of other firms
- We showed that there are cases that low-recall predictions are useful and ML techniques could be adopted to satisfy them

Threats to Validity

- Construct validity: Concerns our misinterpretation of IBM workflows. Addressed by prolonged close connection with IBM
- Internal validity: relates to whether the results follow from the data. Feed-back driven design of study and feature engineering
- External validity: Concerns lack of generalizability of our study, mitigated by studying several projects and feature categories
- Reliability: Specified design decisions and settings clearly

Summary

- Engineered a set of 29 features and classified them into 9 logical categories
- Proposed a cost-insensitive learning process which provides acceptable results based on F1-score
- Proposed a cost-sensitive learning process to maximize precision or predictions according to interest of IBM
- Ranked features based on their relative importance to the trained model

Future Work

- On-site validation of model
- Engineering further features in the nine categories
- Text analysis on description, summary and comments
- Use of ensembles to integrate models trained on different data
- Analysis on communications between stakeholders while incorporating their organizational and geographical distribution

Thanks!