

**Department of Computer and Software Engineering**

**SOEN 6611-E (Software Measurement)**

**Winter 2019**

**Milestone 4**

“We certify that this submission is the original work of members of the group and meets the Faculty's Expectations of Originality”

|  |  |  |
| --- | --- | --- |
| **Submitted by (Team 10):** |  | **Submitted to:** |
| * Hari Narayanan Kannan * Harsh Divecha * Ramit Basra * Dhaval Chandreshkumar Modi * Venkata Lakshmi Sai Krishna Chadalavada | (**40047827)**  **(40084737)**  **(40045043)**  **(40083289)**  **(40087977)** | Dr. Rodrigo Morales Alvarado |

1. **Type of Study:**

We will be studying fault-proneness in relation to C&K. Source for mining repository has been selected as GitHub adhering, as it is a very well known and an extensively used revision control system.

**2. Related Studies:**

Objective of this research was to find Empirical evidence of the association between the badsmells and class error probability. In this study they designed a Bayesian inference for individual metrics which provide posterior probability for fault occurrence[9].

In this study, we collected data about faults found in object- oriented classes. Based on these data, we verified how much fault-proneness is influenced by internal (e.g., size, cohesion) and external (e.g., coupling) design characteristics of oo classes [12].

Relationship between object-oriented metrics and defect prediction was studied in terms of bug by, computing the accuracy of the proposed model on different datasets using logistic regression [18].

1. **Projects:**

Projects selected are all based on Java, as most of the open source software’s are available in this language and codes written in java require Object Oriented concepts to be implemented implicitly and it is one of the most extensively used and documented languages known. All the projects selected have more than 300k lines of code which increases the probability of the code being fault prone or consisting of defects in general.

|  |  |
| --- | --- |
| **Project Selected:** | **Tools and plug-ins [6]:** |
| * + Ant   + Mockito   + Wildfly | * Understand 5.0.9 by Scitools. * Jupyter notebook. * DesigniteJava. |

Revisions Selected for studying the most desirable projects:

|  |  |  |
| --- | --- | --- |
| **Mockito** | **Ant** | **Wildfly** |
| * 2.25.0 | * 1.01.0 | * 16.0 |
| * 2.21.1 | * 1.05.2 | * 12.0 |
| * 2.17.6 | * 1.10.5 | * 8.0 |

1. **Metrics:**

**5.1 C&K Metrics:**

The C&K Metric Suite has been implemented to analyse fault proneness in the projects under consideration [3] [4]. Chidamber and Kemerer (CK) et al. [2] gives the formal definition of the metrics as follows:

**Weighted Methods per Class (WMC):** This measures the sum of complexity of the methods in a class. The complexity of the class may be calculated by the cyclomatic complexity of the methods. The high value of WMC indicates that the class is more complex as compare to the low values.

**Depth of Inheritance Tree (DIT):** DIT metric is used to find the length of the maximum path from the root node to the end node of the tree. DIT represents the complexity and the behavior of a class, and the complexity of design of a class and potential reuse.

**Number of children (NOC):** According to Chidamber and Kemerer, the Number of Children (NOC) metric may be defined for the immediate sub class coordinated by the class in the form of class hierarchy. These points are come out as NOC is used to measure that “How many subclasses are going to inherit the methods of the parent class”. The greater the number of children, the greater the potential for reuse. The greater the number of children, the greater the likelihood of improper abstraction of the parent class.

**Coupling between Objects (CBO):** CBO is used to count the number of the class to which the specific class is coupled. The rich coupling decreases the modularity of the class making it less attractive for reusing the class and more high coupled class is more sensitive to change in other part of the design through which the maintenance is so much difficult in the coupling of classes.

**Response for class (RFC):** The response set of a class (RFC) is defined as set of methods that can be executed in response and messages received a message by the object of that class. Larger value also complicated the testing and debugging of the object through which, it requires the tester to have more knowledge of the functionality.

**Lack of Cohesion in Methods (LCOM):** This metric is used to count the number of disjoints methods pairs minus the number of similar method pairs used. Since cohesiveness within a class increases encapsulation it is desirable and due to lack of cohesion may imply that the class is split in to more than two or more sub classes. Low cohesion in methods increase the complexity, when it increases the error proneness during the development is so increasing.

**5.2 Bad Smells:**

Bad smell means a piece of code in program that signifies the design problem in software code structure. Bad smell is a bug not a run time error. These bad smells require refactoring to improve the code quality [8]. We have used DesigniteJava tool to detect bad smells in the projects. The following are the list of bad smells:

**5.2.1. Design Smells [17]:**

#### Abstraction Design Smells: The definition of ab-straction design principle is given as follows: “The principle of abstraction advocates the simplification of entities through reduction and generalization: reduction is by elimination of unnecessary details and generalizati-on is by identification and specification of common and important characteristics.” The following are the smells detected that violates the principle of abstraction: Imperative Abstraction, Multifaceted Abstraction, Unutilized Abstraction, Duplicate Abstraction.

**Encapsulation Design Smells:**

The definition of encapsulation design principle is given as follows: “The principle of encapsulation advocates separation of concerns and information hiding through techniques such as hiding implementation details of abstractions and hiding variations.”The following are the smells detected that violates the principle of encapsu-lation:

**Deficient Encapsulation**, **Unexploited Encapsulation**

**Modularization Design Smells:** The definition of modularization design principle is given as follows: “The principle of modularization advocates the creation of cohesive and loosely coupled abstractions through techniques such as localization and decomposition.” The following are the smells that violates the principle of modularization: Broken Modularization, Insufficient Modularization, Hub-like Modularization, Cyclically-dependent Modularization.

**Hierarchy Design Smells:** The definition of modulariz-ation design principle is given as follows: “The principle of hierarchy advocates the creation of a hierarchical organization of abstractions using techniques such as classification, generalization, substitutability, and ordering.”. The following are the smells that violates the principle of hierarchy: Wide Hierarchy, Deep Hierarchy, Multipath Hierarchy, Cyclic Hierarchy, Rebellious Hierarchy, Unfactored Hierarchy, Missing Hierarchy.

1. **Implementation Smells:**

**Duplicate Code**: Identical or very similar code exists in more than one location.

**Long Method**: A [method](https://en.m.wikipedia.org/wiki/Method_(computer_science)), function, or procedure that has grown too large.

**Long Parameter List**: More than three or four parameters for a method.

**Magic Number**: A magic number is a numeric value that’s encountered in the source but has no obvious meaning. This “anti-pattern” makes it harder to understand the program and refactor the code.

**Missing Default Case:**A default case is missing in a *case* or *selector* statement.

**Long Statement:** The code contains long statements (that typically do not fit in a screen).

**Complex Conditional:** The longer a piece of code is, the harder it’s to understand. Things become even more hard to understand when the code is filled with conditions.

**Virtual Method Call from Constructor**: Constructors of derived types are called after constructors of base types. On the other hand, calls to virtual methods are always executed on the most derived type. This means that if you call a virtual member from the constructor in a base type, each override of this virtual member in a derived type will be executed *before* the constructor of the derived type is called. If the override in the derived type uses its members, this can lead to confusion and errors.

**Empty Catch Block:** An empty catch block code smell undermines the purpose of exceptions. When an exception occurs, nothing happens and the program fails for an unknown reason. The application can be in an unknown state which will affect the subsequent processing.

**Long Identifier**: Using too long identifiers may actually decrease the readability a lot.

1. **Related Work**

**1.1. Summary**

All of the research work studied and presented here share a common goal, to investigate relation between CK metrics and fault-proneness for faulty classes, an attempt is also made for building a prediction model for fault prediction using the conclusions derived in each paper. As the software evolves, the code and the architecture become more complex, which can inevitably induce bugs in the system. Finding bugs later in the development stages can increase the cost of software remodeling. Software developers usually use various metrics for determining the quality of the system. Using these metrics for predicting faults during development as well as every stage can increase the efficiency of the overall development process[12].

Satwinder Singh [9], used Analyst4J tool for detecting metrics as well as code smells, for determining faulty classes, while, in paper ‘Software bug prediction using object-oriented metrics’[18], Ckjm tool is used for CK metrics collection. Method for code smell detection is not specified. Victor R. Basili[12], used a manual approach, the study was performed on codes written in C++ language by various students of differing skill-sets and skill levels, data about errors was collected during the testing phase and fixes during the repair phase. GEN++ tool was used for collecting CK metrics. All the papers had different ways to collect the data, but used the same logistic regression model. If the count of bug or smell detected was greater than 0, it was labeled 1 else 0. This is because logistic regression is the most appropriate model for binary classification and fault detection is a binary classification problem, hence same methodology is shared among all the papers.

Bayesian inference was performed on 2 out of 6 metrics which showed a threshold for both metrics, which is a posterior probability of occurrence of error. Likelihood probability of error increasing is concluded at 40% with density ~6 for CBO and 35% with density ~9 for NOC [9]. Regression Analysis was performed and using P-values for each metric, following are the metrics and their category for fault prediction, DIT, RFC, NOC : very significant, WMC, CBO : significant and LCOM : insignificant [12]. From these studies we can safely conclude that though the final metrics from the results may variate, the CK metrics suite is significant in terms of fault prediction.

**1.2. Comparison with the study proposed:**

**Similarities & Differences:**

The paper work of [18] is not shown in table as it is very similar to the work by Satwender[9].

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Basilli [12]** | **Satwender [9]** | **Our work** |
| Suite of Metrics | CK | CK | CK |
| Metrics tool: | Gen++ | JDeodarant. | Scitool’s Understand. |
| Language | C++ | Java | Java |
| Dependent Variable type: | Code faults and coding errors. | Implementat-ion Code Smells. | Design and Implementation smells. |
| Data Collection method | Manual Paper based | Analyst4J tool. | Designite Java. |
| model | Logistic Regression. | Bayesian Inference | Logistic Regression. |
| Nature : | Binary. | Binary | Binary |

The approach for classifying errors/smells into 1 and 0 is shared among all the papers and our work. A significant difference in this paper is that we are considering both the design and implementation code smells for detecting presence of bugs. Since we are using the latest automated tools we expect a much better quality in terms on analysis.

**Limitations:** The study in paper is performed on systems developed by students, which are taken as study subjects. These systems would consist of bugs and code smells which would have not been present if professionally developed systems were used in the first place. Additionally, faulty classes are being defined during testing manually which might incorporate errors induced by humans[12]. Results were concluded and analysis was performed on 2 out of 6 metrics in paper [9] which does not tell anything about the involvement of other metrics in prediction.

**Improvements:** Our study will be done on professionally developed and well known systems, we select 3 softwares in addition to three release for each product. Hence the analysis we perform can be said to be relevant in real working environment with some extent of certainty.These different softwares simulate different work culture and environments. Also, we are using specialized and updated software’s for the entire process which might improve the results of analysis. Inclusion of both implementation and design code smells, adheres to a more strict bug detection ruleset.

**2. Data Collection**

**2.1. Internal Metrics Implementation [15] [16]:**

**CBO (Count Class Coupled):** CBO represents the number of classes coupled to a given class. According to this metric “Coupling Between Object Classes” (CBO) for a class is a count of the number of other classes to which it is coupled. Theoretical basis of CBO relates to the notion that an object is coupled to another object if one of them acts on the other, i .e. methods of one use methods or instance variables of another (Chidamber et. al. 1994). As Coupling between Object classes increases, reusability decreases and it becomes harder to modify and test the software system. Chidamber and Kemerer state that their definition of coupling also applies to coupling due to inheritance, but do not make it clear if all ancestors are involuntarily coupled or if the measured class has to explicitly access a field or method in an ancestor class for it to count. In general, the definition of the CBO is ambiguous, and makes its application tough (Mayer et. al. 1999).

**NOC (Count Class Derived):**  Number of children (NOC) of a class is the number of immediate sub-classes subordinated to a class in the class hierarchy. Theoretical basis of NOC metric relates to the notion of scope of properties. It is a measure of how many sub-classes are going to inherit the methods of the parent class (Chidamber et. al. 1994).

**WMC (Sum Cyclomatic):** It is the sum of the cyclomatic complexity of each method for all methods of a class.If a Class C, has n methods and c1, c2 …cn be the complexity of the methods, then

WMC(C)= c1 + c2 +… + cn.

The specific complexity metric that is chosen should be normalized so that nominal complexity for a method takes on a value of 1.0. If all method complexities are considered to be unity, then WMC = n, the number of methods. WMC break an elementary rule of measurement theory that a measure should be concerned with a single attribute (Chidamber et. al. 1994).

**DIT (Max Inheritance Tree):** The DIT for class X is the maximum path length from X to the root of the inheritance tree. According to this metric Depth of Inheritance (DIT) of a class is “the maximum length from the node to the root of the tree. The definition of DIT is ambiguous when multiple inheritance and multiple roots are present.

**LCOM (Percent Lack of Cohesion):** The LCOM metric value is calculated by removing the number of method pairs that share other class field from number of method pairs that does not share any field of other class. Consider a Class C1 with n methods M1, M2 ..., Mn. Let {Ij} = set of instance variables used by method Mi. There are n such sets {I1},{I2}... {In}. Let P = {(Ii,Ij) | Ii ∩ Ij = ∅ } and Q = { (Ii ,Ij) | Ii ∩ Ij ≠ ∅ }. If all n sets {I1},{I2}... {In}. are ∅ then let P = ∅ [4]. Lack of Cohesion in Methods (LCOM) of a class can be defined as:

LCOM = |P| - |Q|, if |P| > |Q|

LCOM = 0 otherwise

The high value of LCOM indicates that the methods in the class are not really related to each other and vice versa. According to above definition of LCOM the high value of LCOM implies low similarity and low cohesion, but a value of LCOM = 0 doesn’t implies the reverse (Mayer et. al. 1999).

**Cyclomatic Complexity:** Cyclomatic complexity metric used for measuring the program complexity through decision-making structure such as if-else, do-while, foreach, goto, continue, switch case etc. expressions in the source code. Cyclomatic complexity only counts the independent paths by a method or methods in a program. Complexity value of program is calculated using following formula:

V (G) =E -N +P,

where

V (G) - Cyclomatic complexity

E - No. of edges of decision graph

N - No. of nodes of decision graph

P - No. of connected paths

If there is no control flow statement like IF statement then complexity of program will be 1. It means there is single path for execution. If there is single IF statement then it provides two paths: one for TRUE condition and other for FLASE condition, so complexity will be 2 for it with single condition.

**RFC (Count Decl Method All):** According to this metric Response for a Class (RFC) can be defined as | RS |, where RS is the response set for the class. The response set for the class can be expressed as:

RS = {M}∪ all i { R i }

where, {Ri} = set of methods called by method i and

{M} =set of all methods in the class.

The response set of a class is a set of methods that can potentially be executed in response to a message received by an object of that class. (Henderson et. al. 1991) But here the point to be noted is that because of practical considerations, Chidamber and Kemerer recommended only one level of nesting during the collection of data for calculating RFC. This gives incomplete and ambiguous approach as in real programming practice there exists “Deeply nested call-backs” that are not considered here. If CK intend the metric to be a measure of the methods in a class plus the methods called then the definition should be redefined to reflect this (Mayer et. al. 1999).

**2.2. Challenges Faced:**

Scitools Understand software is used for metrics collection, it provides the data of metrics not only on classes, but also on files and directory, resulting to an incorrect analysis, so extra cleaning had to be done on the data. Software systems selected for study are well known, widely used and large systems, retrieving data on personal laptops having lower ram memory led to multiple crashes and was time consuming to some extent. Arrangements had to be done for a laptop with RAM size of at least 12 GB for metrics collection. Initially we were using only CK metrics and with single criteria for each type of code-smell category, as our knowledge was limited. After some discussions and extra study, we concluded that there should be multiple criteria for each type of code-smell detection. We finally settled in for the use of an external tool, but we propose to study further and develop our own set of criteria for code-smell classification as a separate work. Data cleaning and Preparation was a major challenge as the output of Designite java separates the Packages, Methods and Types of classes then detects code smells. This step composes 80% of the work time. The output from Designite java was to be manipulated in a way that it could be merged with the output from Scitool’s Understand.

**2.3. Correctness of Metrics:**

C&K metrics are widely used and have been validated multiple times, so we mention the hypothesis for the construct of these metrics instead, with regards to our proposed study.

**WMC:** A class with significantly more member functions than its peers is more complex and, by consequence, tends to be more fault-prone [12].

**DIT:** A class located deeper in a class inheritance lattice is supposed to be more fault-prone because the class inherits a large number of definitions from its ancestors. In addition, deep hierarchies often imply problems of conceptual integrity, i.e., it becomes unclear which class to specialize from in order to include a subclass in the inheritance hierarchy [12].

**NOC:** Classes with large number of children (i.e., subclasses) are difficult to modify and usually require more testing because the class potentially affects all of its children. Furthermore, a class with numerous children may have to provide services in a larger number of contexts and must be more flexible. We expect this to introduce more complexity into the class design and, therefore, we expect classes with large number of children to be more fault-prone [12].

**CBO:** Highly coupled classes are more fault-prone than weakly coupled classes because they depend more heavily on methods and objects defined in other classes [12].

**RFC:** Classes with larger response sets implement more complex functionalities and are, therefore, more fault-prone [12].

**LCOM:** Classes with low cohesion among its methods suggests an inappropriate design (i.e., the encapsulation of unrelated program objects and member functions that should not be together) which is likely to be more fault-prone [12].

**2.4. Git Repository:**

* https://github.com/HarshDivecha/SOEN661

**3. External Metrics Collection:**

**3.1. Tools used:**

* Scitools Understand: For collecting C&K metrics.
* DesigniteJava: For Code-Smells
* Jupyter Notebook: For data analysis.

**3.2. Tools Explored:**

* BugZilla
* PMD
* JDeodrant
* InFusion

**Empirical Study**

Before performing statistical analysis the most important task is to clean the data. Since the data is produced by a software there should not be any missing values, hence the data is clean in terms of quality. But we have used different softwares for metrics data and code smell data collection. Check for the format outputted, as these differ a custom code was written to combine both the data-sets in a standard form. Then the major consideration is the imbalance in dataset.

|  |  |
| --- | --- |
| Count of Bugs | |
| 0 (Absence) | 232434 |
| 1 (Presence) | 11074 |

We have performed all the analysis on the raw data-set then performed modelling after dealing with imbalance in the data. This is because if we use the imbalanced data, then the accuracy of the model would always be greater than 90%, which is a sign of an over-fitting. The resulting model would miss-classify many 0’s as 1’s.

1. **Statistical tests**:

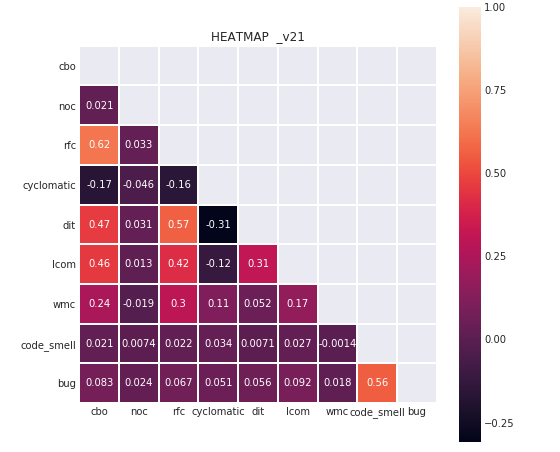
We are studying and building a bug prediction model, with dependent variable being presence of bug i.e. (if code\_smell > 0 : classify as 1, else: 0). Hence, we need a statistical test for binary variables.

We are using logistic regression as we have only two outputs, i.e. binary.

1. **Variables with high correlation**:

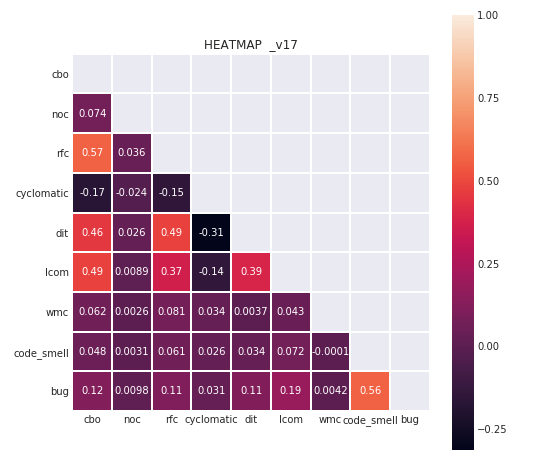
From the open source projects taken under consideration, the following are the high correlations that we observe. Variables executing correlation >45% are considered, depending on this we select variables such that, we eliminate the variables contributing repitetive data. They are listed are follows:

Mockito



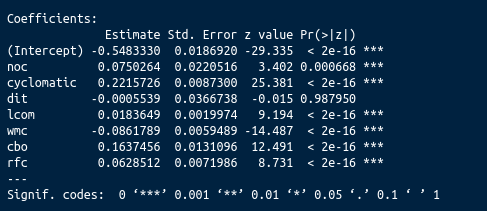
|  |  |
| --- | --- |
| CBO vs RFC (62%)  CBO vs DIT (47%) | CBO vs LCOM (46%)  RFC vs DIT (57%) |

WildFly



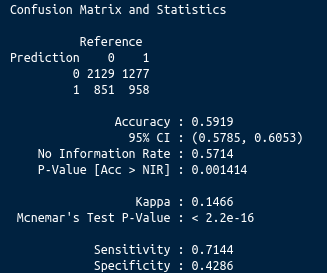
|  |  |
| --- | --- |
| CBO vs RFC (57%)  CBO vs DIT (46%) | CBO vs LCOM (49%)  RFC vs DIT (49%) |

**3. Regression with all independent variables in the model**:



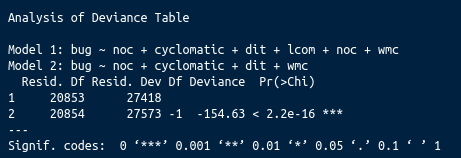
We use P-values to make conclusions in significance testing. More specifically, we compare the P-value to a significance level to make conclusions about our hypotheses. The above significance codes display levels of significance. We use the widely accepted threshold value of alpha for this purpose set at 0.5 on positive and negative sides of the normal distribution. If the P-values lies between 0 and alpha, it is accepted as significant. More closer to 0 higher the significance. We accept the variables with \*\*\* as they execute a P-value between (0.000, 0.001). Hence, we can discard ‘**dit**’ as its significance is **0.98**, which shows that it does not contribute much to our model.

**4**. **Hypothesis**:



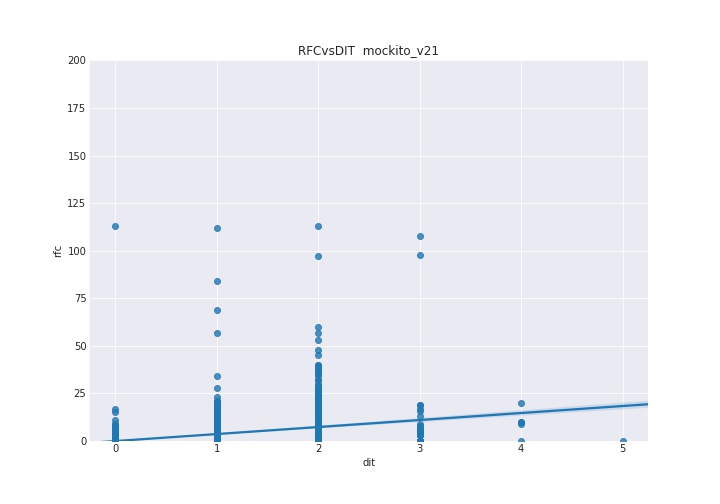
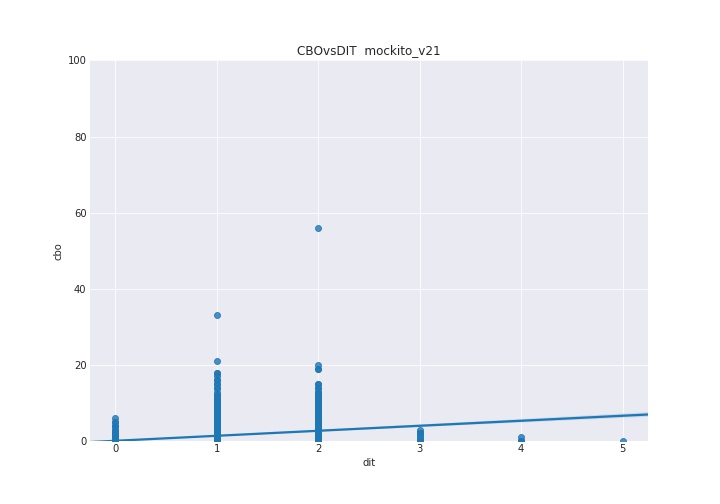
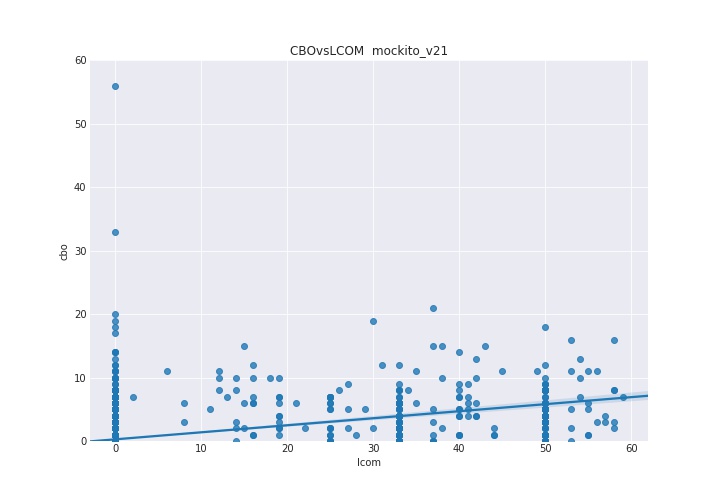
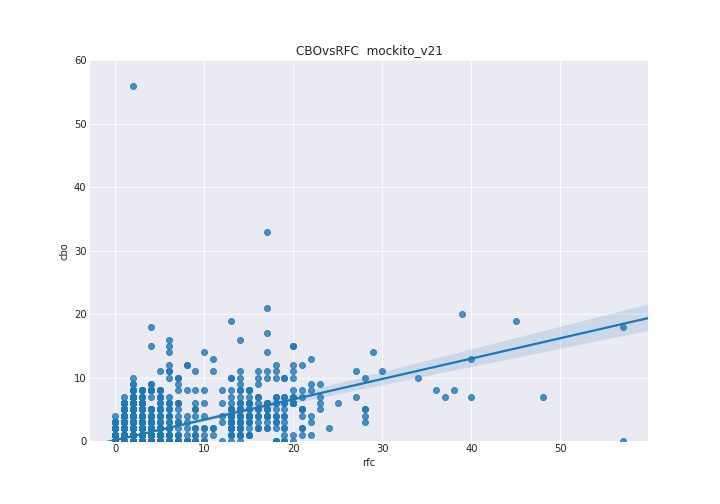
***Null hypothesis***: C&K metrics has no correlation with Code Smells/bugs.

***Alternative hypothesis***: C&K metrics is correlated with Code Smells/bus.



The basic model without hyper-parameter tuning has a 60% prediction accuracy, with specificity of 42% and sensitivity of 71%, shows that the model is not over-fitted. After eliminating insignificant variables, we see that the model 2 is significant compared to model 1.

We **reject** the null hypothesis.



**5. References:**

[1] Amit Sharma, Sanjay Kumar Dubey, “Comparison of Software Quality Metrics for Object- Oriented System”, IJCSMS International Journal of Computer Science & Management Studies, Special Issue of Vol. 12, June 2012 ISSN (Online): 2231 –5268.

[2] C. Shyam, Kemerer, F. Chris, "A Metrics Suite for Object- Oriented Design" M.I.T. Sloan School of Management, pp. 53-315, 1993.

[3] Istehad Chowdhury, Mohammad Zulkernine, “Using complexity, coupling, and cohesion metrics as early indicators of vulnerabilities”, Journal of Systems Architecture 57 (2011) 294–313.

[4] Ramanath Subramanyam and M.S. Krishnan, “Empirical Analysis of CK Metrics for Object-Oriented Design Complexity: Implications for Software Defects”, IEEE Transactions on Software Engineering, Vol. 29, No. 4, April 2003.

[5] E Fioravanti, P. Nesi, “A Study on Fault-Proneness Detection of Object-Oriented Systems”, [Proceedings Fifth European Conference on Software Maintenance and Reengineering](https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7309), 07 August 2002, 6905819.

[6] Rüdiger Lincke, Jonas Lundberg and Welf Löwe, “Comparing Software Metrics Tools”, Proceedings of the ACM/SIGSOFT International Symposium on Software Testing and Analysis, ISSTA 2008, Seattle, WA, USA, July 20-24, 2008.2.

[7] Radu Marinescu, “Measurement and Quality in Object-Oriented Design,” PhD thesis, Politechnica University of Timisoara, Romania, 2002.

[8] Sukhdeep Kaur1 & Dr. Raman Maini2, Analysis of Various Software Metrics Used to Detect Bad Smells, 2016

[9] Heena Kapila & Satwinder Singh, Analysis of CK Metrics to predict Software Fault-Proneness using Bayesian Inference, International Journal of Computer Applications (0975 – 8887)

[10] Victor R. Basili, A Validation of Object-Oriented Design Metrics as Quality Indicators.

[11] Software Measurement PPTs by Dr. Rodrigo Morales Alvarado.

[12] A Validation of Object-Oriented Design Metrics as Quality Indicators, Victor R. Basili, 1996

[13] Evaluation of code smells and detection tools: https://doi.org/10.1186/s40411-017-0041-1

[14] DesigniteJava: <https://github.com/tushartushar/DesigniteJava>

[15] “A Critical Suggestive Evaluation of CK Metric”,  [Pacific Asia Conference on Information Systems, PACIS 2005, Bangkok, Thailand, July 7-10, 2005](https://www.researchgate.net/conference-event/PACIS_Pacific-Asia-Conference-on-Information-Systems_2005/9595).

[16] “Analysis of Various Software Metrics Used to Detect Bad Smells”, Sukhdeep Kaur, Dr. Raman Maini, The International Journal of Engineering and Science (IJES), Volume 5, Issue 6, Pages PP -14-20 2016 ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805.

[17] “Refactoring for Software Design Smells: Managing Technical Debt”-

Ganesh Samarthyam, Girish Suryanarayana, and Tushar Sharma.

[18] Software bug prediction using object-oriented metrics DHARMENDRA LAL GUPTA 1,2, \* and KAVITA SAXENA 2