MEF UNIVERSITY

SOFTWARE PROJECTS CLUSTERING AND SELECTION BY MACHINE LEARNING METHODS

Capstone Project

Elif Torun

İSTANBUL, 2018

MEF UNIVERSITY

SOFTWARE PROJECTS CLUSTERING AND SELECTION BY MACHINE LEARNING METHODS

Capstone Project

Elif Torun

Advisor: Prof. Semra Ağralı

İSTANBUL, 2018

MEF UNIVERSITY

Name of the project: SOFTWARE PROJECTS CLUSTERING AND SELECTION BY MACHINE LEARNING METHODS
Name/Last Name of the Student: Elif Torun
Date of Thesis Defense: 03/09/2018

I hereby state that the graduation project prepared by Elif Torun has been completed under my supervision. I accept this work as a "Graduation Project".

03/09/2018
Prof. Semra Ağralı

I hereby state that I have examined this graduation project by Elif Torun which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

03/09/2018
Prof. Özgür Özlük

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of Elif Torun and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Prof. Semra Ağralı	………………………..
2. Prof. Özgür Özlük	………………………..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University’s ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name 03/09/2018 Signature

Elif Torun

EXECUTIVE SUMMARY

SOFTWARE PROJECTS CLUSTERING AND SELECTION BY MACHINE LEARNING METHODS

Elif Torun

Advisor: Prof. Semra Ağralı

AUGUST, 2018, 31 Pages

In today's hyper volatile business world, software development projects play key roles in maintain the current situation of the company and they are vital in taking the company one step further. Selecting the right project to invest is a critical decision point regarding the hard competition, diminishing profitability and high cost of the projects. The main aim of this study is clustering the projects and deciding which project to invest by using machine learning methods. We use IT project demands data of one of the biggest banks due to the capital, number of transactions and number of customer portfolio in Turkey. The data includes 2048 Information Technology related project demands occurred in 2017 and 2018. For the clustering part of the project both unsupervised and supervised learning methods are used and success rates are compared. We observe that supervised learning methods are more successful than the unsupervised ones. For the project selection part all process of the bank and output of the all steps are reviewed. According to our results, second workshop, which is the last step of the project assessment and selection process, has almost 50% of the total process effort and gives the precise effort estimation as an outcome, can be eliminated, and the project selection decision can be made with around 90% success ratio with machine learning methods. The result of this study provides an efficient way to select projects and a platform to see the complexity of the project portfolio.

Key Words: Project clustering, project selection, demand management, K- Means, Logistic Regression, Support Vector Machine.

ÖZET

MAKİNE ÖĞRENMESİ METODLARI İLE YAZILIM GELİŞTİRME PROJELERİ SINIFLANDIRMASI VE SEÇİMİ

Elif Torun

Tez Danışmanı: Prof. Semra Ağralı

AĞUSTOS, 2018, 31 Sayfa

Günümüzün hızla değişen iş dünyasında, yazılım geliştirme projeleri şirketlerin mevcut durumlarını korumak için anahtar oyuncularken, şirketi bir adım ileri götürmek için de zorunludur. Sıkı rekabet koşulları, azalan karlılıklar ve projelerin yüksek maliyetleri göz önüne alındığında, doğru projeye yatırım yapmanın kritik bir karar noktası olduğunu görüyoruz. Bu çalışmanın amaçları, makine öğrenmesi metotları kullanılarak projeleri karmaşıklık seviyelerine göre sınıflandırmak ve hangi projelere yatırım yapılması gerektiğine karar vermektir. Çalışmada Türkiye'nin sermaye, işlem sayısı ve müşteri portföyü açısından en büyük bankalarından birinin Bilişim Teknolojileri proje talepleri kullanılmıştır. Veride 2017 ve 2018 yıllarında talep edilen 2048 proje bulunmaktadır. Sınıflandırma problemi için gözetimsiz ve gözetimli öğrenme teknikleri kullanılarak, başarı oranları karşılaştırılmıştır. Sonuç olarak gözetimli öğrenme tekniğinin proje sınıflandırmasında daha başarılı olduğu tespit edilmiştir. Proje seçim problemi için de tüm süreç ve çıktılar gözden geçirilmiştir. Sonuç olarak proje planlama sürecinin toplam eforunun yaklaşık %50'sini alan ikinci çalıştay kaldırılarak, makine öğrenmesi teknikleri ile proje seçiminde yaklaşık %90 başarıya ulaşıldığı görülmüştür. Bu çalışmanın sonucu, proje seçim sürecinde verimlilik artışı sağlamakta ve proje portföyündeki karmaşıklığı gösteren bir platform sunmaktadır.

Anahtar Kelimeler: Proje kümeleme, proje seçimi, talep yönetimi, K- Means, Logistic Regression, Support Vector Machine.

TABLE OF CONTENTS

1. INTRODUCTION

Technology is in the center of the ordinary people's everyday life. In order to protect their presence in the market, all companies need to be adapted to new technologies. On the other hand, technology is changing rapidly so it is not enough to have the existing technologies, it is important to see the future trends and invest in essential projects in line with their strategies. Ohame (1982) defines strategy as:

> "The way, in which a corporation endeavors to differentiate itself positively from its competitors, using its relative corporate strengths to better satisfy customer needs"

Strategic initiative execution is essential for the organization's competition. While this is the case, according to Economist Intelligence Unit's survey that includes 587 senior executives globally (EIU, 2013), 61% of the participants state that executing the strategies is the main difficulty in this process. In the last three years only, 56% of the strategic initiatives executed successfully.

According to the 2014 PwC survey (PwC, 2014), more than %66 of the CEO's declared that there is a need for change and/or they are developing strategies to change. So as all the figures show that change management is a critical responsibility of all managers.

As a result, organizations talk a good game about strategy but without the right projects and programs to carry them out, even the most forward-thinking strategies fail.

In this research, the main focus is the clustering of the project by their complexity and decision of performing or not performing the projects by using machine learning methods. For project clustering problem, K- means method is used as an unsupervised learning method, and Multiclass Logistic Regression and Support Vector Machines methodologies are used as supervised learning. The current decision process of the bank is reviewed and outputs of the all steps fed to the Logistic Regression and Support Vector Machines models 90% success with improved efficiency in the process is achieved.

1.1. What is a "Project"

The required features to define a study as Project are (i) being temporary that has a defined beginning and end, and (ii) produce a unique product or service with a defined scope. Here unique states that every project has its own specific set of activities that is different from the routine operations. The software development for a defined business process or construction of a building is all well - known examples of the project.

1.2. Project Selection Process

Projects need to be selected carefully and managed by experts to deliver on time and within budget constraints. In order to achieve this, the idea behind the project needs to be evaluated and approved. At this point, Project Management Office (PMO) and Project Management Methodologies are taking place to evaluate all projects objectively.

Project Management defined as the application of tools and techniques by experienced project managers who have the skills and knowledge of the methodologies.

While all the methodologies are focusing on the management of the selected process to deliver them successfully, there is no standard approach for project selection process. In this study, the Bank's overall process has been reviewed and detailed analysis is made based on the current process.

1.2.1. Constraints

Three main constraints scope, budget and time are applicable for all kinds of project internationally. In the Bank, budget means both internal resource capacity and cash out payments such as outsourcing, hardware investments or software investments. In addition to those, Bank's strategies and regulations are the critical constraints for the project selection and resource usage decision. The bank defines 5 years strategic plan and every year in August, that year's strategic initiatives are defined.

1.2.2. Decision Process

Decision process starts with collecting the demands from business units in August. Work stream Group of the projects is assigned as the strategic initiative, running business and compliance. After that, all the business units prioritize their demands, project management office defines the main domain and related business units of the project and main domain defines turnkey effort ratio, total effort range**,** related IT units, vendor unit count and high-level budget requirements. Compliance projects are reviewed by the compliance office, clustered due to their criticality as High, Medium and Low. Complexity clustering is defined with all these information by PMO. As a result of this phase, preliminary information for all projects is ready.

In order to make a precise estimation, 2nd workshops are organized with the leadership of the project managers and participation of the demand owner and related IT domains. High-level requirements of the project are reviewed, detailed and precise effort and budget estimation is completed.

2. EXPOLATORY DATA ANALYSIS

In this study, the project demands in 2017 and 2018 are used to solve the project clustering and selection problem. Properties of the dataset are explained in this section.

2.1. About the Data

The dataset includes 24 columns and 2084 rows. Please find the explanations of the columns below.

Year (int64) is the year that project requested. In the dataset, there is only 2017 and 2018.

Project Type (object) there are 3 kinds of projects, Running, Masterplan and Insertion.

1) **Running:** means the project has been planned previous year and ongoing in the planned year. For example, the project was requested in masterplan 2017 period and planned but developments are still ongoing in 2018. To see the overall portfolio, running projects are also in the dataset.
2) **Masterplan:** New project requested in the masterplan and subject to masterplan assessments.
3) **Insertion:** This project is neither in the last years nor existing year's masterplan. This is an urgent project, inserted in the masterplan last year and still ongoing.

Project Name (object) is masked because of the data privacy rules of the bank. You can see the names as Project1, Project2... etc.

Department (object) refers to requested business unit.

BU Ranking (int64) is the prioritization of the requested business unit among their projects.

Workstream (object) is the categorization of the projects due to their impact area such as, Mandatory, Compliance, Commercial, Efficiency, Risk Mitigation and Reputation

Workstream Group (object) is the mapping of the projects with the workstream of the projects, strategic initiative, running business and compliance of the organization

Program name (object) is also masked but that gives you if the project is a part of a program or not.

Turnkey Effort Ratio (%) (int64) means it is suitable to give this project as a turnkey to a vendor. If it is, then you can see the ratio of the portion that can be given as turnkey. This cannot be %100 because none of the vendors can do any of the projects with "0" effort of the bank. On the other hand, if it is “0” that means this project cannot be given to a vendor.

Total Effort Range (object) gives a range of the effort after preliminary analysis of the requirements.

1st WS Average Effort (int64) interval of the total effort range that estimated by expert judgment of the main domain in the first workshops.

IT Unit Count (float) is the no of the domains that are related with the projects.

BU Unit Count (int64) is the no of the Business Units related with the projects

Vendor Unit Count (int64) is the no of vendors that will be working on the project

Main Domain Analysis (object) gives the main responsible analysis team in IT

Main Domain Development (object) gives the main responsible development team in IT

Compliance Priority (object) is the result of compliance office department assessment for the work stream marked as Mandatory, Compliance as High Medium and Low.

Complexity (object) is the clustering of the project based on their complexity as A, B, C, D. A refers to the most complex projects that need to be managed by experienced project managers and D refers to simple projects that might be managed by domain teams or junior project managers.

Score (int64) gives you the complexity score calculated based on the complexity of the architecture, project definition, related channels ... etc.

2nd Workshop Effort (int64) Effort is the precise estimation of the project effort due to well-defined requirements

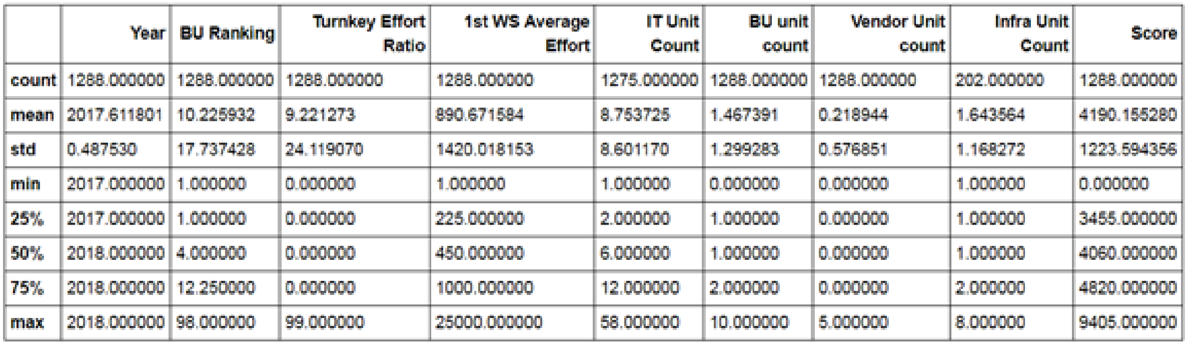
Budget Required (K TL) (int64) is the detailed estimation of the required budget to achieve detailed business goals.

Master Plan In / Out (object) is the key legend of the data that gives you this project planned in the masterplan or not.

2.2. Data Cleaning Process

When I review the process, 1st WS Average Effort is a critical variable for both clustering and project selection. I deleted the rows that have none, blank and zero on 1st WS Average Effort and 2nd Workshop Effort. I replace 1st WS Average Effort with the 2nd Workshop Effort for the cases that have 2nd Workshop Effort but do not have 1st WS Average Effort.

As a result, my data has 1.288 rows and 24 columns. As a result of all the cleaning process, the dataset is reduced and diversity decreased. There is a condensation on the Master Plan "In" projects. On the models for project selection, that condensation could cause an overfitting problem for Master Plan "In" cases.



	Year	BU Ranking	Turnkey Effort Ratio	1st WS Average Effort	IT Unit Count	BU unit count	Vendor Unit count	Infra Unit Count	Score
count	1288.000000	1288.000000	1288.000000	1288.000000	1275.000000	1288.000000	1288.000000	202.000000	1288.000000
mean	2017.611801	10.225932	9.221273	890.671584	8.753725	1.467391	0.218944	1.643564	4190.155280
std	0.487530	17.737428	24.119070	1420.018153	8.601170	1.299283	0.576851	1.168272	1223.594356
min	2017.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
25%	2017.000000	1.000000	0.000000	225.000000	2.000000	1.000000	0.000000	1.000000	3455.000000
50%	2018.000000	4.000000	0.000000	450.000000	6.000000	1.000000	0.000000	1.000000	4060.000000
75%	2018.000000	12.250000	0.000000	1000.000000	12.000000	2.000000	0.000000	2.000000	4820.000000
max	2018.000000	98.000000	99.000000	25000.000000	58.000000	10.000000	5.000000	8.000000	9405.000000

3. PROJECT DEFINITION

3.1. Problem Statement

There is no standardized method for software development projects clustering and the decision of performing or not performing the projects. Technology is changing rapidly, companies have their strategic goals and regulators need to be satisfied to sustain the current position and move forward. In order to achieve these right projects need to be selected to invest and complexity of the project need to be defined properly.

3.2. Project Objective

First target of the research is clustering the projects by their complexity. As the projects have their unique products, management of the project becomes more difficult when the complexity increases. Project manager and project management methodology need to be defined due to the complexity of the project.

Second target of the research is to select the right projects to do by using machine learning tools and techniques.

3.3. Project Scope

Clustering of the projects due to complexity and project selection (doing/ not doing decision of the projects) is in the scope of the study. The main focus is the effective way of deciding projects to plan by using machine learning tools and techniques. Project effort estimation which is the key variable in the decision is not in the scope of this study. It needs to be detailed as a separate research question. On the other hand, the planning of the project due to the constraints is not in the scope of this study. Optimal planning algorithm of the selected projects needs to be separate research question.

4. METHODOLOGY

4.1. Project Clustering Problem

Microsoft Azure Machine Learning platform is a tool that we can import our dataset, make data cleaning, build models and compare results. For clustering problem, both supervised and unsupervised techniques are implemented and results are compared. All steps can be seen on Figure 1.

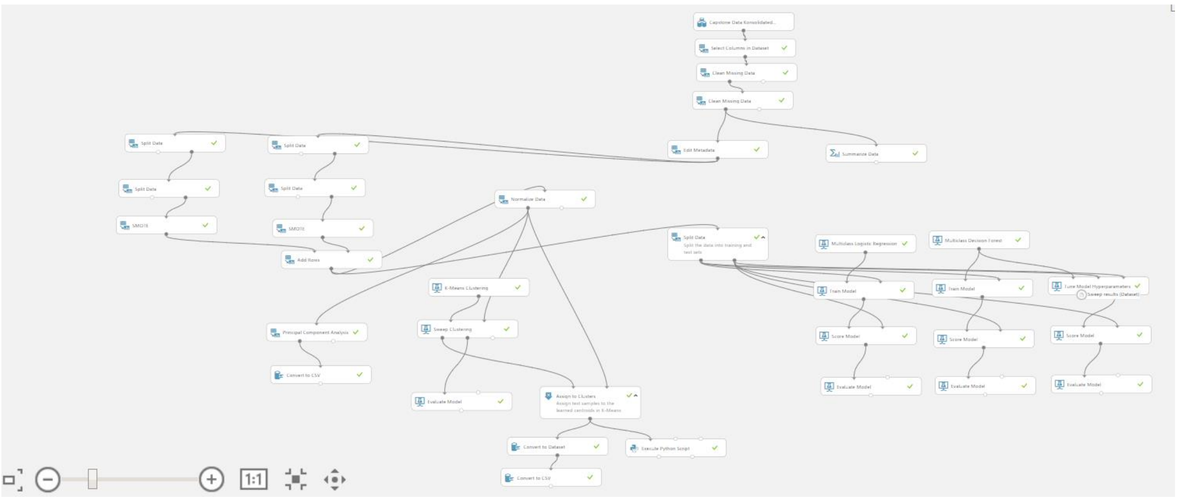


Figure 1. Microsoft Azure Experiment for Project Clustering Problem

Data is imported to Azure platform, and in the Select Columns in Dataset step, Project Name, 1st WS Average Effort, IT Unit Count, BU unit count, 2nd Workshop Effort, Budget Required, Complexity, BU Ranking, Vendor Unit count are selected for further analysis.

As the complexity is the label used for clustering the projects, the rows with empty complexity information are deleted in the Clean Missing Data step. Also, the missing values on the IT Unit Count column are replaced with mean. Complexity is changed as categorical at the Edit Metadata step.

As the dataset is imbalanced and there is a condensation on the B type projects, SMOTE is used to balance the dataset. SMOTE is a statistical technique to increase the number of rare cases while the number of major cases stays the same with the original data set. Since it works in binary format, dataset is first divided into two sets by using split data.

Then the split sets are combined with add rows function. Distribution of the original dataset and SMOTE dataset are given below in Figure 2.

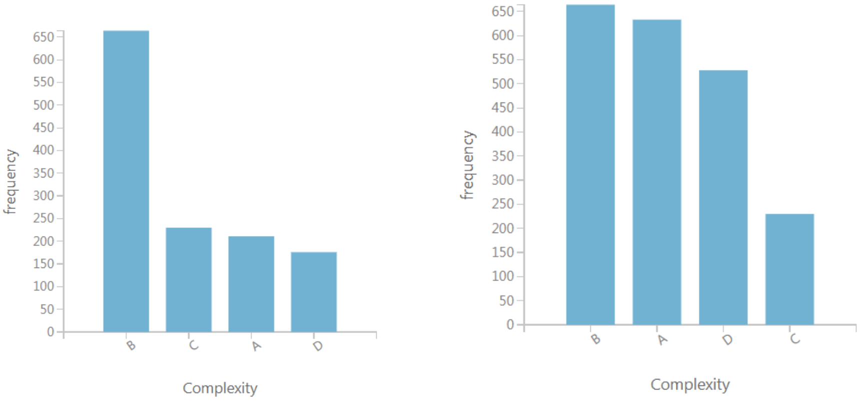


Figure 2. (a) Original dataset, (b) SMOTE dataset

Then model development begins in three branches by using unsupervised and supervised learning models.

4.1.1. Unsupervised Learning Methods, Tools and Techniques and Results

K- Means clustering is used as an unsupervised learning methodology. Before building the model, dataset is normalized with the logistic transformation method. K- Means clustering is defined with 4 Centroids, 200 iterations and Euclidean metric, without giving any label. Sweep Clustering is used to measure the clustering result by simplified silhouette metric. The result of the clusters is shown in Figure 3.

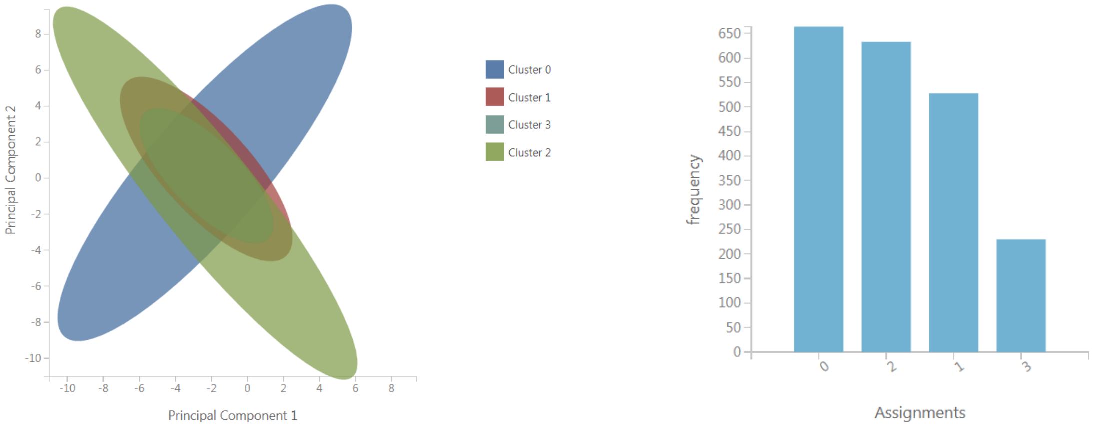
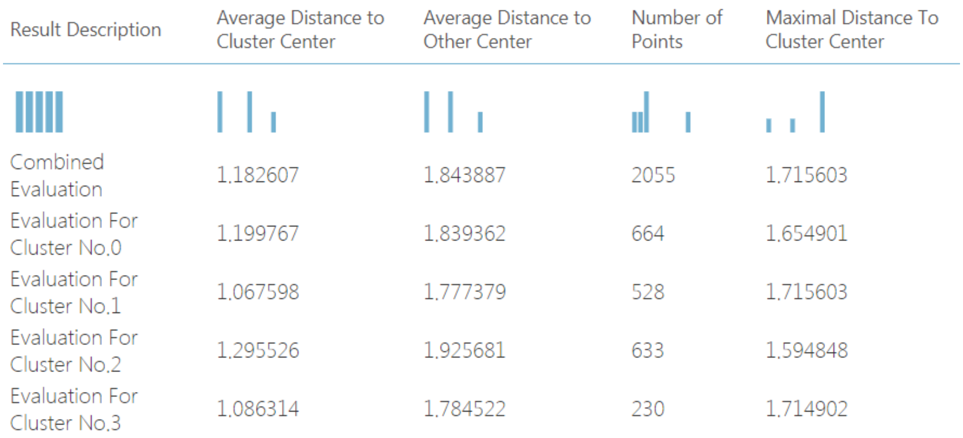


Figure 3. K – Means Unsupervised Clustering Results

Also clustering evaluation results of the K- Means are given in Figure 4.



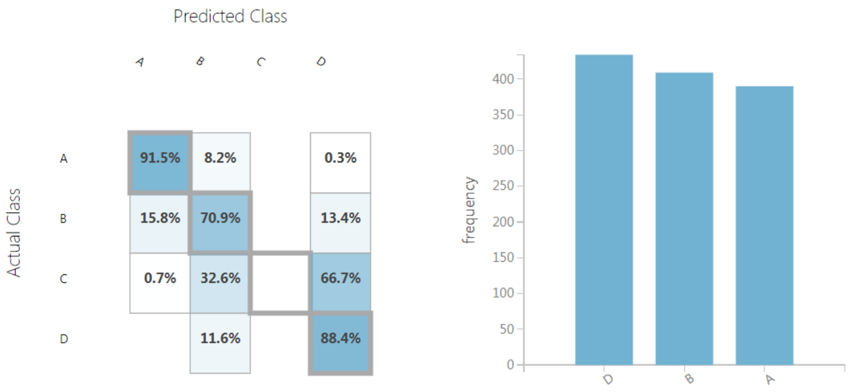
Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Combined Evaluation	1.182607	1.843887	2055	1.715603
Evaluation For Cluster No.0	1.199767	1.839362	664	1.654901
Evaluation For Cluster No.1	1.067598	1.777379	528	1.715603
Evaluation For Cluster No.2	1.295526	1.925681	633	1.594848
Evaluation For Cluster No.3	1.086314	1.784522	230	1.714902

Figure 4. K- Means Evaluation Results

4.1.2. Supervised Learning Methods, Tools and Techniques, and Results

Multiclass Logistic Regression and Multiclass Decision Forest (i.e., Random Forest) methods are used as supervised learning methods. Data is split into train (0.4) and test (0.6) sets. Complexity is selected as the label.

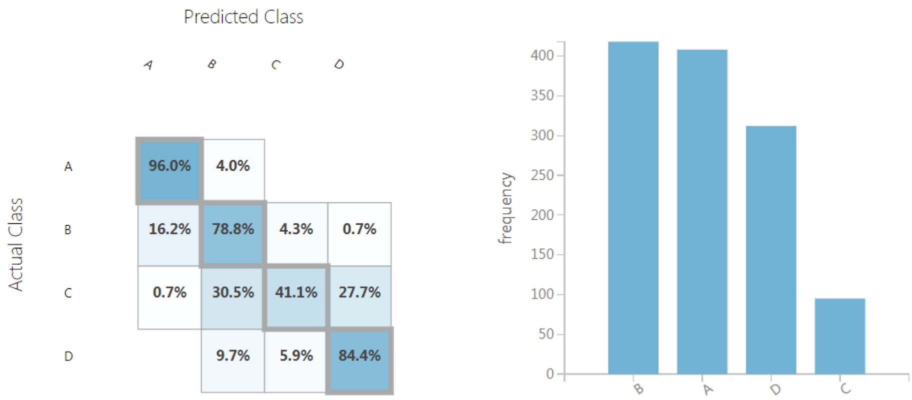
For Multiclass Logistic Regression, overall accuracy of the model is 0.73 while average accuracy is 0.86. As you can see the confusion matrix and predicted classes on Figure 5, accuracy are enormous for class A and Class D, while class C is not predicted.



Actual Class \ Predicted Class	A	B	C	D
A	91.5%	8.2%		0.3%
B	15.8%	70.9%		13.4%
C	0.7%	32.6%		66.7%
D		11.6%		88.4%

Figure 5. Multiclass Logistic Regression Model Results

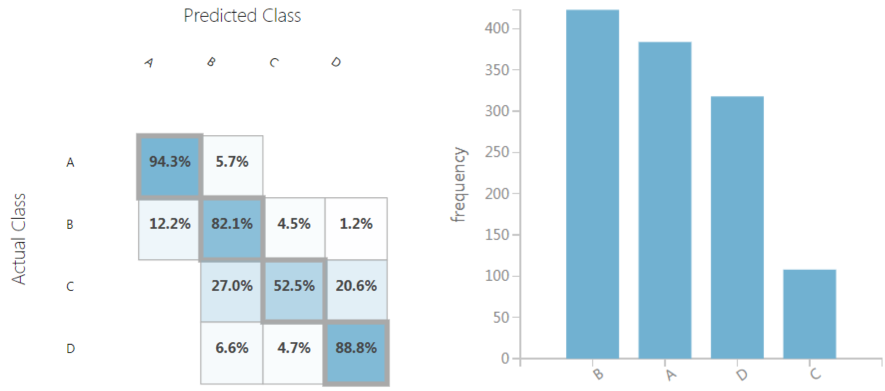
As a result of the Multiclass Decision Forest with the same dataset, overall accuracy of the model is increased to 0.80 while average accuracy is increased to 0.90. Predicted classes are given in Figure 6.



	A	B	C	D
A	96.0%	4.0%		
B	16.2%	78.8%	4.3%	0.7%
C	0.7%	30.5%	41.1%	27.7%
D		9.7%	5.9%	84.4%

Figure 6. Multiclass Decision Forest Model Results

Multiclass Decision Forest Model results are better than Multiclass Logistic Regression. With the tuned parameters, overall accuracy of the model is increased to 0.84 while average accuracy is increased to 0.92. Predicted classes are given in Figure 7.



	A	B	C	D
A	94.3%	5.7%		
B	12.2%	82.1%	4.5%	1.2%
C		27.0%	52.5%	20.6%
D		6.6%	4.7%	88.8%

Figure 7. Multiclass Decision Forest Model Results with tuned parameters.

4.1.3. Value Delivered

In conclusion, project can be clustered with multiclass decision forest model with 0.92 accuracy by using only 7 variables; BU ranking, 1st WS Average Effort, IT Unit Count, BU Unit Count, Vendor Unit Count, 2nd Workshop Effort, Budget Required.

The concentration of the original data for type-B projects is remarkable in the original dataset. Definition of type-B projects needs to be analyzed in detail for more homogenous distribution of the project complexity. As the data is imbalanced, SMOTE is used to increase the minority cases and rows added for more balanced dataset before building any model.

Decision forest is an ensemble-learning model for classification that builds a series of decision trees in order to increase the accuracy of the model and learn from tagged data. This is the main reason of the increased accuracy of the model than Multiclass Logistic Regression. In addition to that, smoothing the dataset increased the efficiency of the model. So with the smooth dataset, we may use this model for project classification prediction with the accuracy of 0.92.

On the other hand, on the cases that you do not have a labelled dataset (you may be a new company or do not have historical data), K – means clustering results are also satisfying. In order to prove that, we can use the K – Means result dataset as an original data and implement multiclass decision forest model.

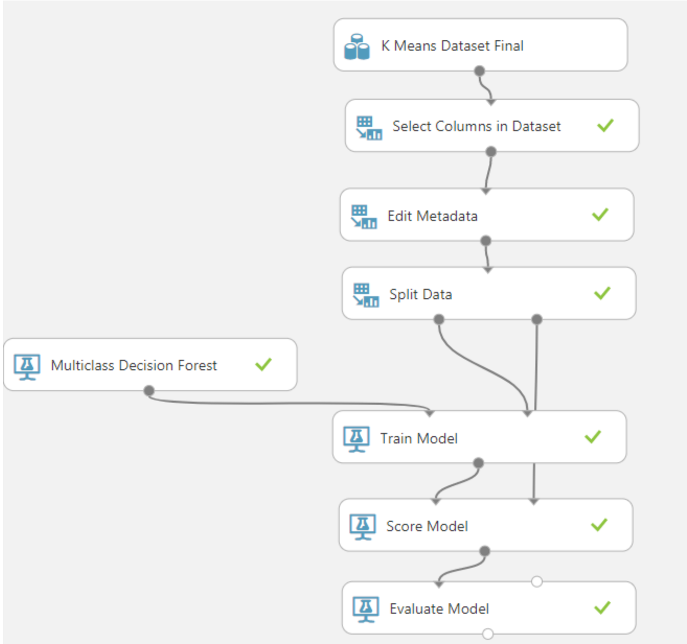
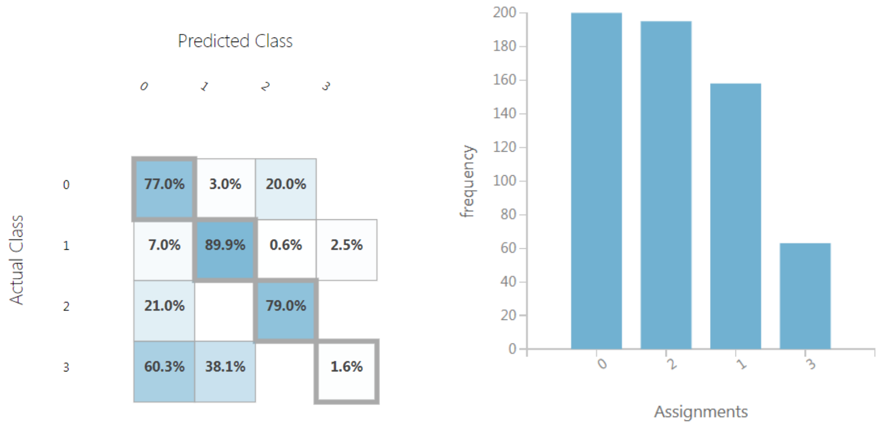


Figure 8. Multiclass Decision Forest Model with K-Means Result Dataset

Even the prediction is less than expected for class 3 which is equal to C Type projects, overall accuracy of the model is 0.73 while average accuracy is 0.87. Predicted classes are given in Figure 9.



Predicted Class

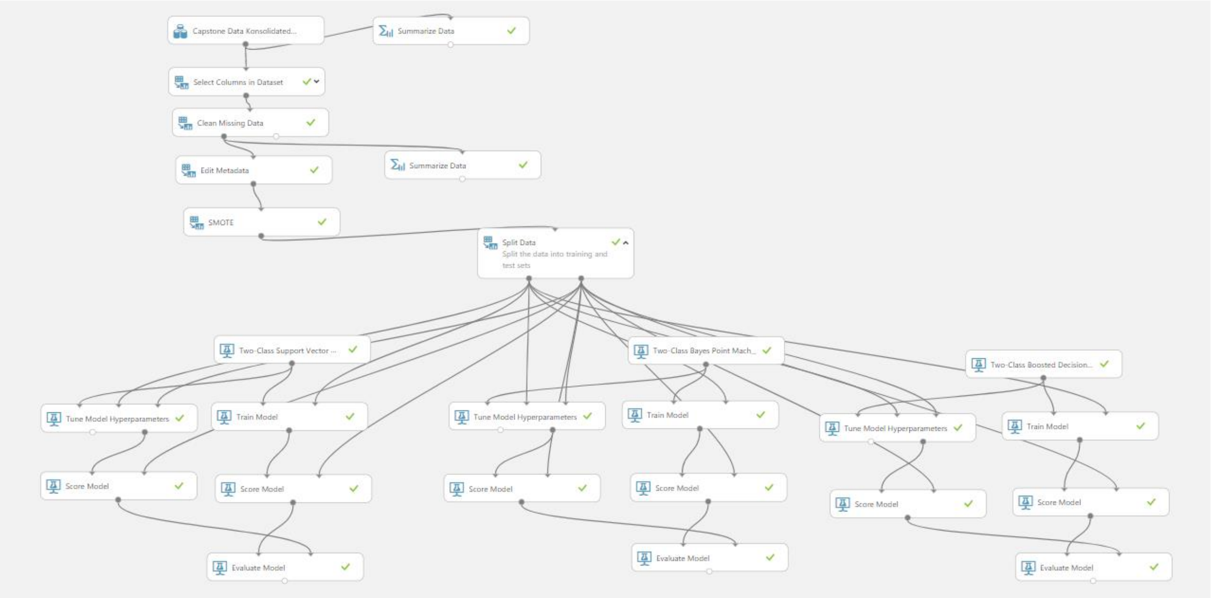
Actual Class	0	1	2	3
0	77.0%	3.0%	20.0%	
1	7.0%	89.9%	0.6%	2.5%
2	21.0%		79.0%	
3	60.3%	38.1%		1.6%

Figure 9. Multiclass Decision Forest Model Results with K- Means Clustering

4.2. Project Selection Problem

4.2.1. Methods, Tools and Techniques

Microsoft Azure Machine Learning platform is used for model implementation as given in Figure 10. For project selection problem, supervised learning techniques; Two-Class Support Vector Machine and Two- Class Bayesian Point and Two- Class Boosted Decision are used.



Capstone Data Konsolidated...
Summarize Data
Select Columns in Dataset
Clean Missing Data
Edit Metadata
Summarize Data
SMOTE
Split Data
Split the data into training and test sets
Two-Class Support Vector ...
Two-Class Bayes Point Mach...
Two-Class Boosted Decision...
Tune Model Hyperparameters
Train Model
Tune Model Hyperparameters
Train Model
Tune Model Hyperparameters
Train Model
Score Model
Score Model
Score Model
Score Model
Score Model
Score Model
Evaluate Model
Evaluate Model
Evaluate Model

Figure 10. Project In/ out Decision Model Building

Project Name, 1st Workshop Average Effort, IT Unit Count, BU Unit Count, Vendor Unit Count, Budget Required, MP In/ Out, Complexity and BU Ranking columns are selected from the dataset, before building the models. Complexity is a critical parameter for the decision process, the lines with empty complexity cell, deleted from dataset. MP In/Out and Complexity variables are changed as categorical. In the dataset, 687 (89%) projects are labelled as "In", which means the projects are selected to be included the master plan, and 82 (11%) projects are labelled as "out", which means the projects are not included in the master plan. SMOTE technique is used to balance the dataset by increasing the number of rare cases, i.e., "out" projects. Please find the original dataset and SMOTE dataset in Figure 11

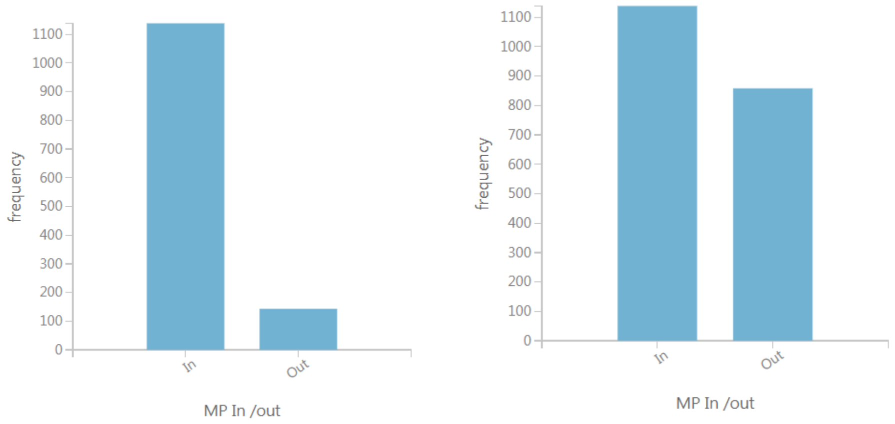
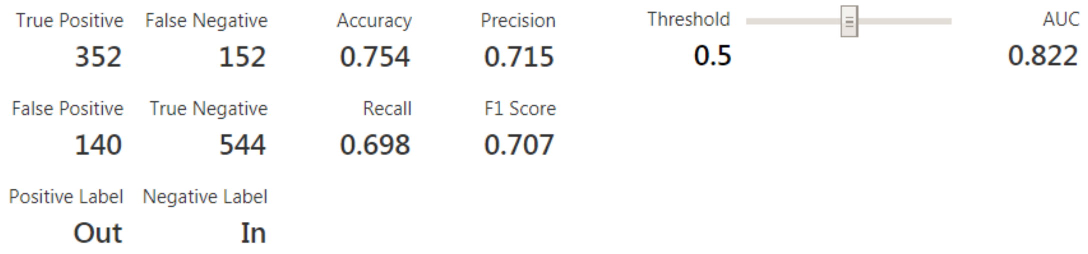


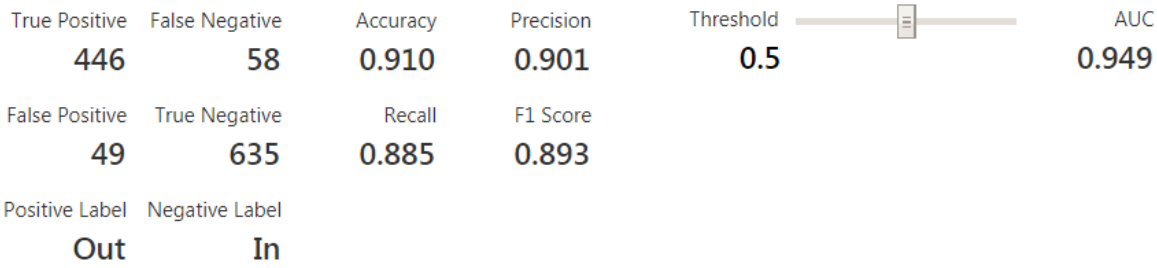
Figure 11. (a) Original dataset, (b)SMOTE dataset with MP In/ out label

First, the model is tested without second workshop effort. The result of the Two-Class Support Vector Machine model, tuned parameters results and the comparison of the models are given in Figures 12, 13 and 14, respectively. The model with tuned parameters gives the best result. The accuracy of the model is 0.910 as it has been increased from 0.75 to 0.91



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
352	152	0.754	0.715	0.5	0.822
False Positive	True Negative	Recall	F1 Score		
140	544	0.698	0.707		
Positive Label	Negative Label				
Out	In				

Figure 12. Two-Class Support Vector Machine Result



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
446	58	0.910	0.901	0.5	0.949
False Positive	True Negative	Recall	F1 Score		
49	635	0.885	0.893		
Positive Label	Negative Label				
Out	In				

Figure 13. Two-Class Support Vector Machine Result with Tuned Parameters

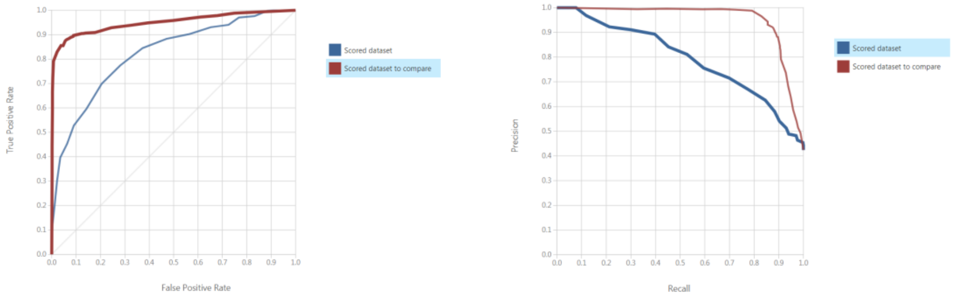
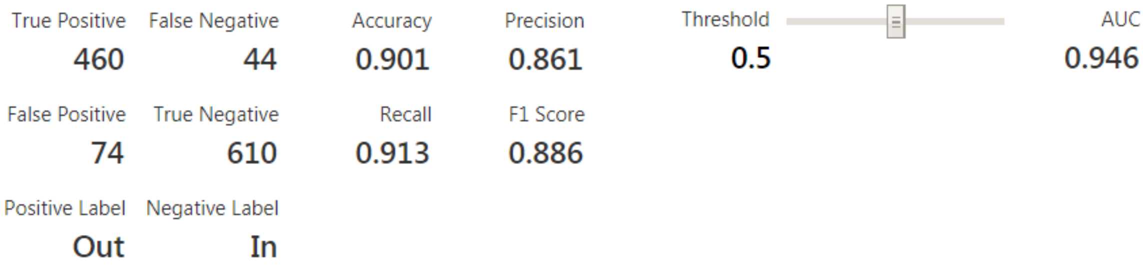


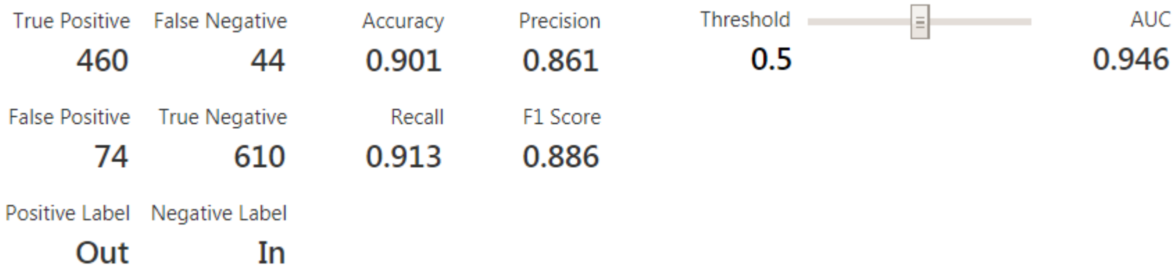
Figure 14. Two Class Support Vector Machine Evaluation Results Comparison

We apply the same analysis by using the same data with Two- Class Bayes Point model. Model results are given in Figure 15. In Figure 16, results are given with the tuned parameters, and in Figure 17 please find the comparison of the model with tuned parameters. The model gives a similar result with tuned parameters, and the accuracy of the model is 0.901 both with and without tuning the parameters.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
460	44	0.901	0.861	0.5	0.946
False Positive	True Negative	Recall	F1 Score		
74	610	0.913	0.886		
Positive Label	Negative Label				
Out	In				

Figure 15. Two Class Bayes Point Evaluation Results



True Positive 460	False Negative 44	Accuracy 0.901	Precision 0.861	Threshold 0.5	AUC 0.946
False Positive 74	True Negative 610	Recall 0.913	F1 Score 0.886		
Positive Label Out	Negative Label In				

Figure 16. Two Class Bayes Point Evaluation Results with Tuned Parameters

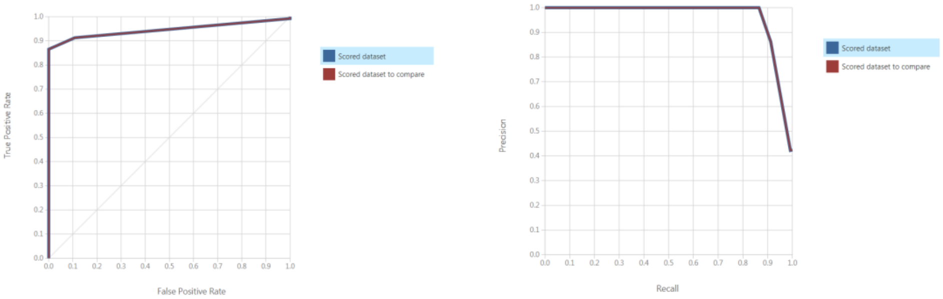
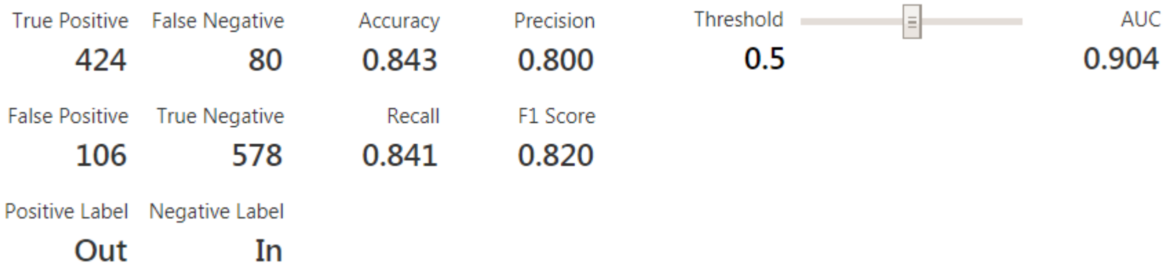


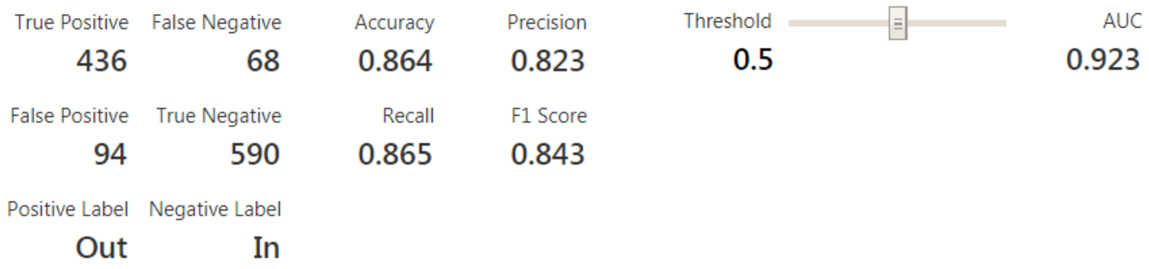
Figure 17. Two Class Bayes Point Evaluation Results Comparison

As a last step, we build the two – class boosted decision model. Model results are given in Figures 18 and 19. The model gives the same result with tuned parameters, and the accuracy of the model is 0.843.



True Positive 424	False Negative 80	Accuracy 0.843	Precision 0.800	Threshold 0.5	AUC 0.904
False Positive 106	True Negative 578	Recall 0.841	F1 Score 0.820		
Positive Label Out	Negative Label In				

Figure 18. Two Class Boosted Decision Model Evaluation Results



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
436	68	0.864	0.823	0.5	0.923
False Positive	True Negative	Recall	F1 Score		
94	590	0.865	0.843		
Positive Label	Negative Label				
Out	In				

Figure 19. Two Class Boosted Decision Model Evaluation Results with Tuned Parameters

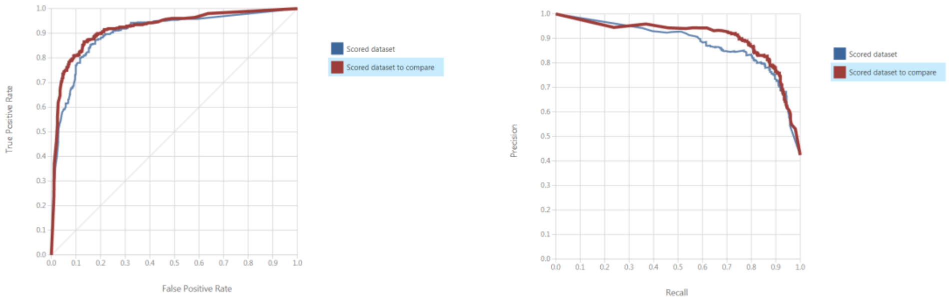
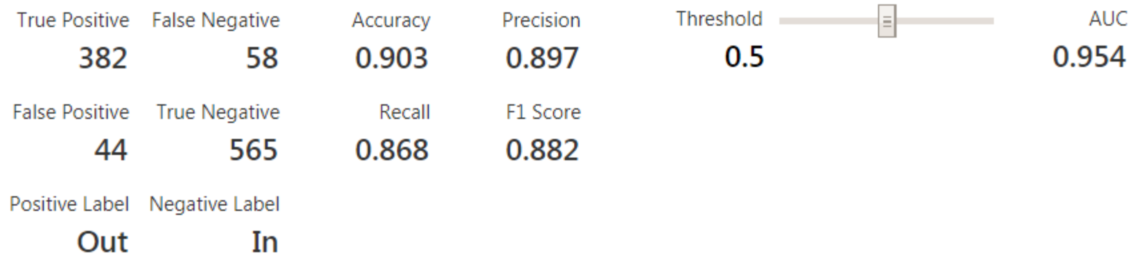


Figure 20. Two Class Boosted Decision Model Evaluation Results Comparison

The research question is how to increase the efficiency of MP in / out decision by using machine learning models. As a result of the analysis above, decision can be made by applying Support Vector Machine with tuned parameters, and the accuracy of the model is satisfactory even without second workshop effort. Then, we analyze the result of the same data with same models but this time with including second workshop efforts.

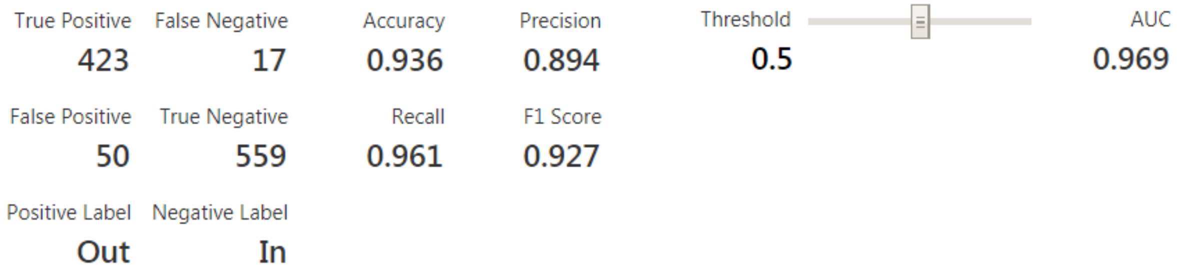
Two-Class Support Vector Machine model is developed with 2nd workshop effort and tuned parameters. The evaluation result of the model is given in Figure 21. The accuracy of the model is 0.903, which is closed to the accuracy of the same model without 2nd workshop effort (0.910).



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
382	58	0.903	0.897	0.5	0.954
False Positive	True Negative	Recall	F1 Score		
44	565	0.868	0.882		
Positive Label	Negative Label				
Out	In				

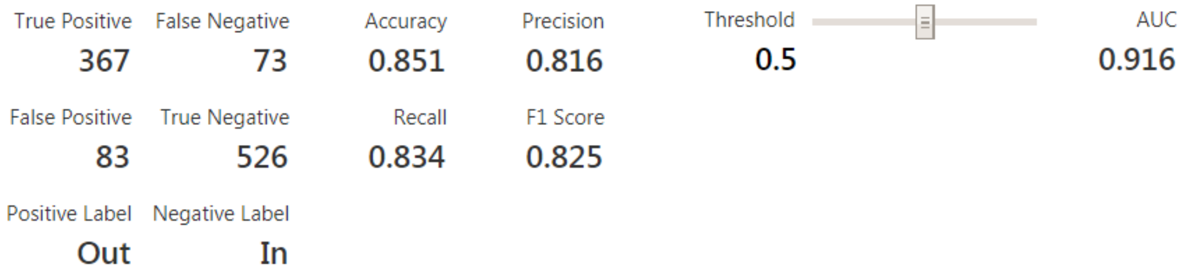
Figure 21. Two Class Support Vector Machine Evaluation Results with Tuned Parameters (with 2nd Workshop Effort)

When we compare the results of the other models with and without 2nd workshop effort, accuracy changes in a range of +/- 0.03.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
423	17	0.936	0.894	0.5	0.969
False Positive	True Negative	Recall	F1 Score		
50	559	0.961	0.927		
Positive Label	Negative Label				
Out	In				

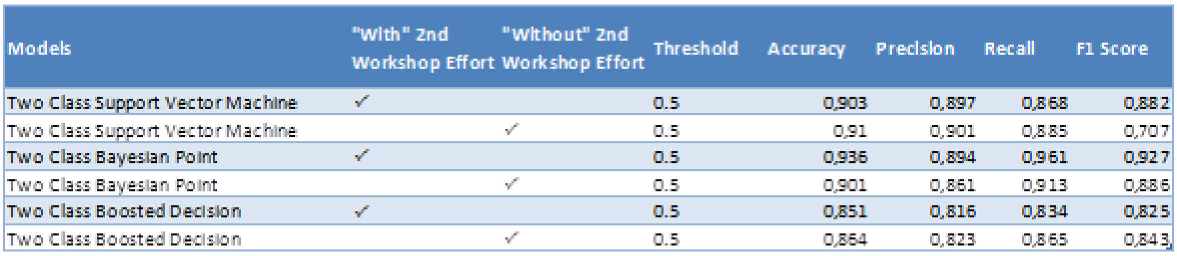
Figure 22. Two Class Bayes Point Evaluation Results (with 2nd Workshop Effort)



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
367	73	0.851	0.816	0.5	0.916
False Positive	True Negative	Recall	F1 Score		
83	526	0.834	0.825		
Positive Label	Negative Label				
Out	In				

Figure 23. Two Class Boosted Decision Model Evaluation Results (with 2. Workshop Effort)

4.2.2. Value Delivered



Models	"With" 2nd Workshop Effort	"Without" 2nd Workshop Effort	Threshold	Accuracy	Precision	Recall	F1 Score
Two Class Support Vector Machine	✓		0.5	0,903	0,897	0,868	0,882
Two Class Support Vector Machine		✓	0.5	0,91	0,901	0,885	0,707
Two Class Bayesian Point	✓		0.5	0,936	0,894	0,961	0,927
Two Class Bayesian Point		✓	0.5	0,901	0,861	0,913	0,886
Two Class Boosted Decision	✓		0.5	0,851	0,816	0,834	0,825
Two Class Boosted Decision		✓	0.5	0,864	0,823	0,865	0,843

There is no significant change in the accuracy of the model with and without 2nd Workshop effort. Moreover, the model accuracy is higher without 2nd Workshop effort. Approximately %50 of the total planning effort is spent in the 2nd Workshops. The contribution of this detailed study about in/ out decision is limited. This result leads us to a conclusion that this step can be eliminated, and approximately 50% productivity can be gained. However, this detailed effort is not only used for in/out decision, but it is also critical for capacity planning. So before eliminating this step, a detailed analysis needs to be performed to decrease the gap between 1st Workshop Effort and 2nd Workshop Effort. The most productive way to achieve this goal is to estimate the efforts in smaller ranges, plan the project with contingency, and review the scope and detail the estimations before starting the project or Q based.

4.3. Conclusion

In this study, we implement unsupervised and supervised learning methods in project proposal classification problem. Then, we implement supervised learning methods to decide on the acceptance of project proposals.

For classification, we observe that Multiclass Decision Forest Model results are satisfactory. On the other hand, for the cases that you do not have a labelled dataset (you may be a new company or do not have historical data), K – means clustering results are also satisfying.

About project selection process, we conclude that 2nd Workshop step can be eliminated, 1st Workshop effort ranges can be given narrower to predict a closer value and productivity can be achieved in the process. Two Class support vector machine model accuracy is satisfactory and can be used for the decision without 2nd workshop effort.

4.4. Social and Ethical Aspects

In this research, project classification and project selection is done by using machine learning methods and without human interaction. In this way, human resources are not wasted, fast and reliable results are achieved. Project success rates can be increased by assigning the appropriate project management methodologies and project managers to right project classes. In addition, project decision is given without any personal relationship purpose. This approach supported company’s strategic roadmaps with the right project selection, isolated from personal relationships.

On the other hand, machine learning methods need to be updated continuously to be able to inline with the fast-changing environment.

5. REFERENCES

EIU, Economist Intelligence Unit, “Why Good Strategies Fail: Lessons for the C-suite”, sponsored by PMI, 2013.

PwC Survey, (2014), 17th Annual Global CEO Survey: “Fit for the future: Capitalizing on global trends”

PMI, (2014), The Project Management Office: Aligning Strategy & Implementation. Retrieved from, https://www.pmi.org/business-solutions/white-papers/align-strategy-implementation, last access on 03/08/2018.

Westland, J., The Project Management Life Cycle, Cambridge, Cambridge University Press, 2006.

Heldman, K., PMP: Project Management Professional Exam Study Guide, Canada,Wiley Publishing.Inc, 2009.

Wang, Z. and Xue, X., Multi- class Support Vector Machine, book chapter in Support Vector Machines Applications, Editors: Yunqian Ma, Guodong Guo, China, Springer, 2014

Shmilovici, A., Support vector machines, Data mining and knowledge discovery handbook, Springer, 2009.

Yang T. Y. and Kuo L. (2001), Bayesian binary segmentation procedure for a Poisson process with multiple change points. Journal of Computational and Graphical Statistics 10: 772–785.

Jung Y. G., Kang M. S., Heo J. (2014). Clustering performance comparison using K-means and expectation maximization algorithms. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4433949/, last access 03.08.2018

PMI, (2018). European Journal of Public Health, 22 (4), 598–601. Retrieved from https://academic.oup.com/eurpub/article/22/4/598/483617?searchresult=1, last access on 12/07/2018.

Figueiredo, M. A. T., Jain A. K., Unsupervised Learning of Finite Mixture Models, IEEE Transactions on PAMI, 24 (3), 381-396, 2002.

Dvir D., Lipovetsky S., Shenhar A., Tishler A., In Search of Project Classification: a non-universal approach to project success factors, 27(4), 915-935, Elsevier, 1998. Retrieved from https://www.sciencedirect.com/science/article/pii/S0048733398000857

Fearnhead, P., Exact and efficient Bayesian inference for multiple changepoint problems, Statistics and Computing, 16(2), 203–213, Springer, 2006.

PMI, What is Project Management. Retrieved from https://www.pmi.org/about/learn-about-pmi/what-is-project-management