# Canada Revenue Agency

# Project Cost Difference & Delivery Success Forecast

**Algonquin College**
**BISI 2023 Fall**
**Team 5**

**Ziang Cui  Xiaomeng Xu  Li Zhuang  Bo Yang  Ang Li**

# Part A. Restates the Opportunity

## Background and Objectives

In today's ever-changing market, accurate budget forecasting and effective cost control are crucial to the success of any project. With this background, our client - Canada Revenue Agency(CRA) requires a comprehensive analysis of their projected budget estimates to ensure that their financial planning aligns with actual expenditures. Therefore,based on the materials and data provided by CRA, our project will focus on leveraging advanced machine learning technologies and the powerful analytics of Power BI to achieve three pivotal objectives: prediction of differences between estimated and actual costs, forecasting the likelihood of success of project deliveries, and generating data-driven insights for stakeholders.

These objectives are designed to enhance financial accuracy, improve project outcomes, and empower decision-makers with actionable intelligence, thereby ensuring a competitive edge in project management.

## Tools

The tools we used in this project mainly include Jira, Power BI, and Machine Learning with Python Pandas library, Python Numpy library, Python Matplotlib library, Python Seaborn library.

## Data Source

The dataset for this project is provided by clients of the Canada Revenue Agency and draws on real-world project data. However, for security and privacy reasons, the data has been synthetically altered by these clients. One limitation to note is the dataset's size—it comprises just 64 rows and 12 columns.

# Part B. Logical Solution

To approach the objectives, following a logical solution as outlined can be highly beneficial.

Firstly, it's essential to determine whether the prediction results should be classified as classification, regression, or both. This decision guides the selection of appropriate algorithms and methods for the model. Once the type of prediction is clarified, the next step involves researching various models that fit the criteria. This research helps in understanding the strengths and limitations of potential models and facilitates an informed decision on the optimal model for the task.

After selecting the model, the next crucial step is data cleaning. This involves removing or correcting any inaccuracies, handling missing data, and normalizing data to ensure it is suitable for analysis. With clean data, we can proceed to make initial predictions using the chosen model.

Following initial predictions, it is important to adjust parameters of the model based on the outcomes. This tuning can involve altering settings like the learning rate, the number of iterations, or other hyperparameters to optimize performance.

Subsequently, the model should be used to make another set of predictions to test the adjustments. Based on these results and the existing data, it is vital to determine the analysis direction, which might involve deeper exploration into specific data segments or features.

The next step would be to use tools such as Power BI for detailed data analysis. Power BI facilitates the examination of data through various lenses and helps in uncovering insights that might not be visible through raw analysis.

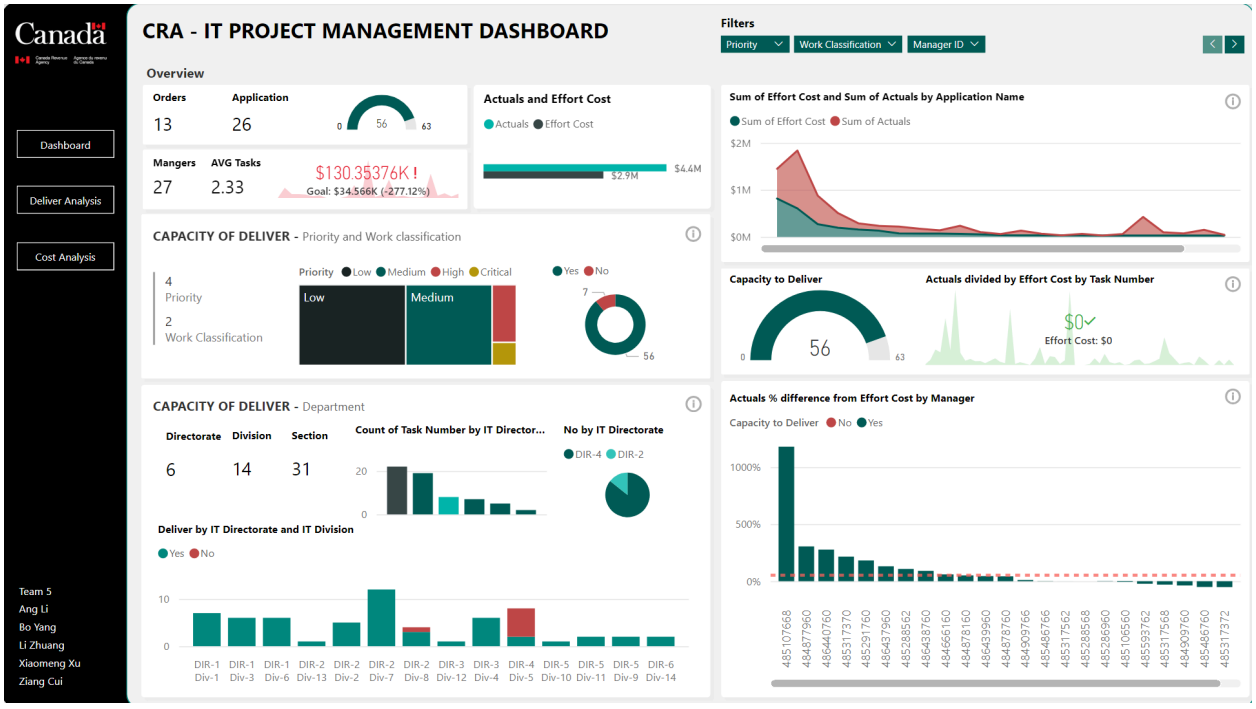
Finally, all the insights and data visualizations can be compiled into a dashboard. Creating a dashboard in Power BI involves designing visual components such as charts, maps, and tables that effectively communicate the findings and trends identified in the data. This dashboard serves as a comprehensive tool for stakeholders to view and interact with the data in a meaningful way, supporting decision-making processes.

Through these steps, we can ensure a structured and thorough approach to tackling a data analysis project, leading to actionable insights and efficient utilization of data.

# Part C. Physical Solution

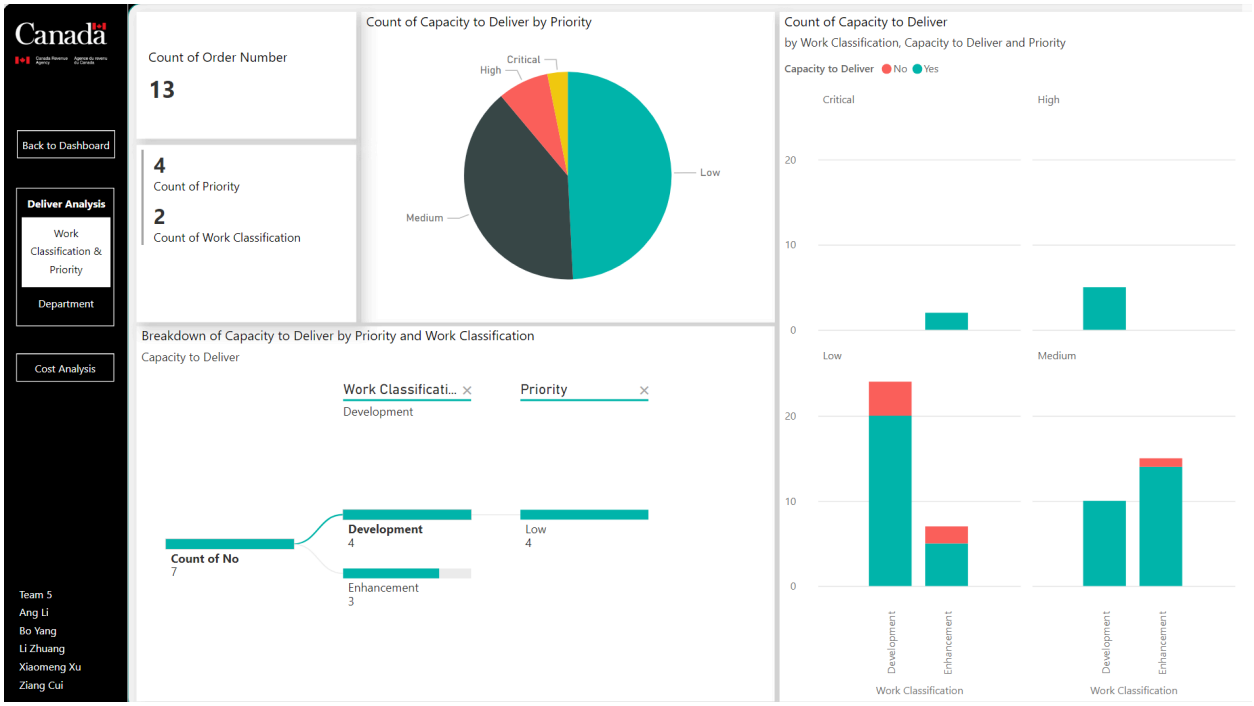
## Power BI

In our examination of performance within the Canada Revenue Agency's IT project management, a suite of detailed Power BI visual reports has been scrutinized to discern the pivotal factors that determine the capacity for project delivery and the variance in costs. These reports consolidate composite data on project priority, work classification, departmental structure, task identification, and managerial performance, showcasing the interplay between these variables and their influence on successful project delivery and cost containment. In this overview, we will reveal how each element singularly and collectively impacts various facets of project management and discuss how these insights can guide future project planning and resource allocation. By integrating analysis of these critical data points, we can deepen our understanding and optimize IT project management processes, ensuring budget fidelity and timely completion of projects.



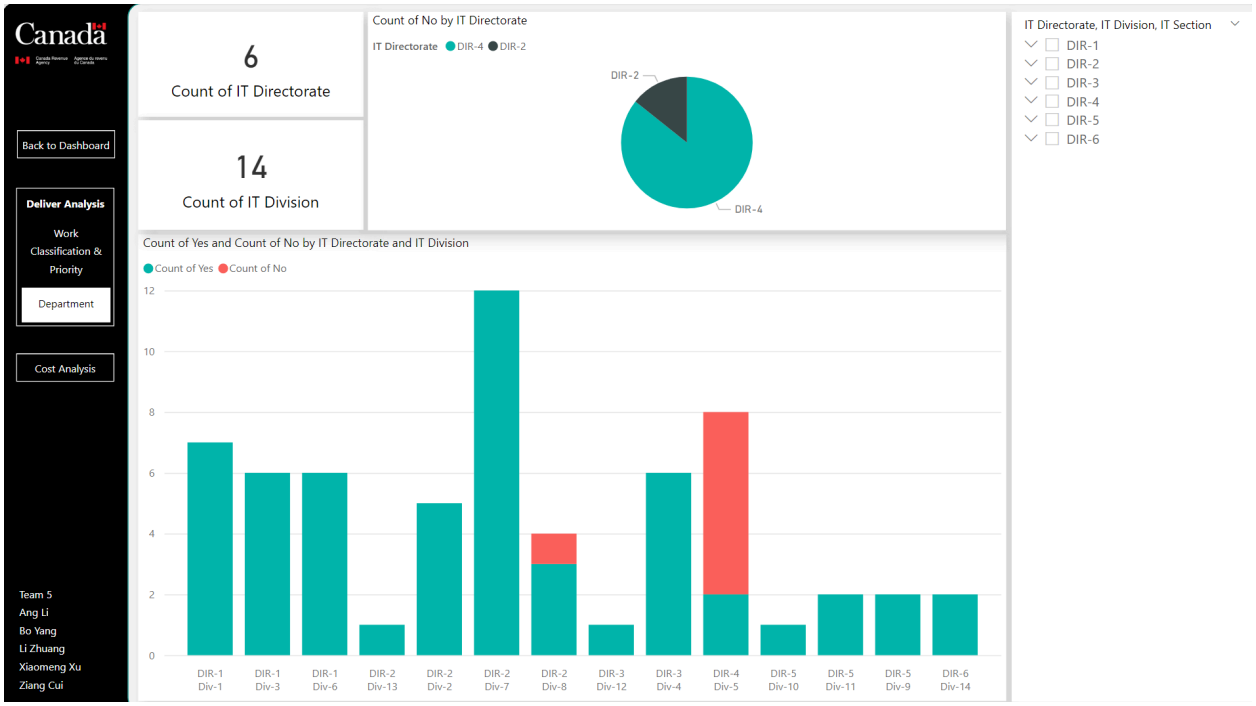

**Project Priority and Delivery Capacity**

The visual data indicate a significant impact of project priority on delivery capacity. Projects with 'Critical' and 'High' priority statuses are shown to have a notably higher likelihood of successful delivery compared to those marked as 'Low' and 'Medium'. This underscores that projects with a higher priority are likely to secure the necessary resources and managerial attention, emphasizing the importance of strategic prioritization and resource allocation in project planning.



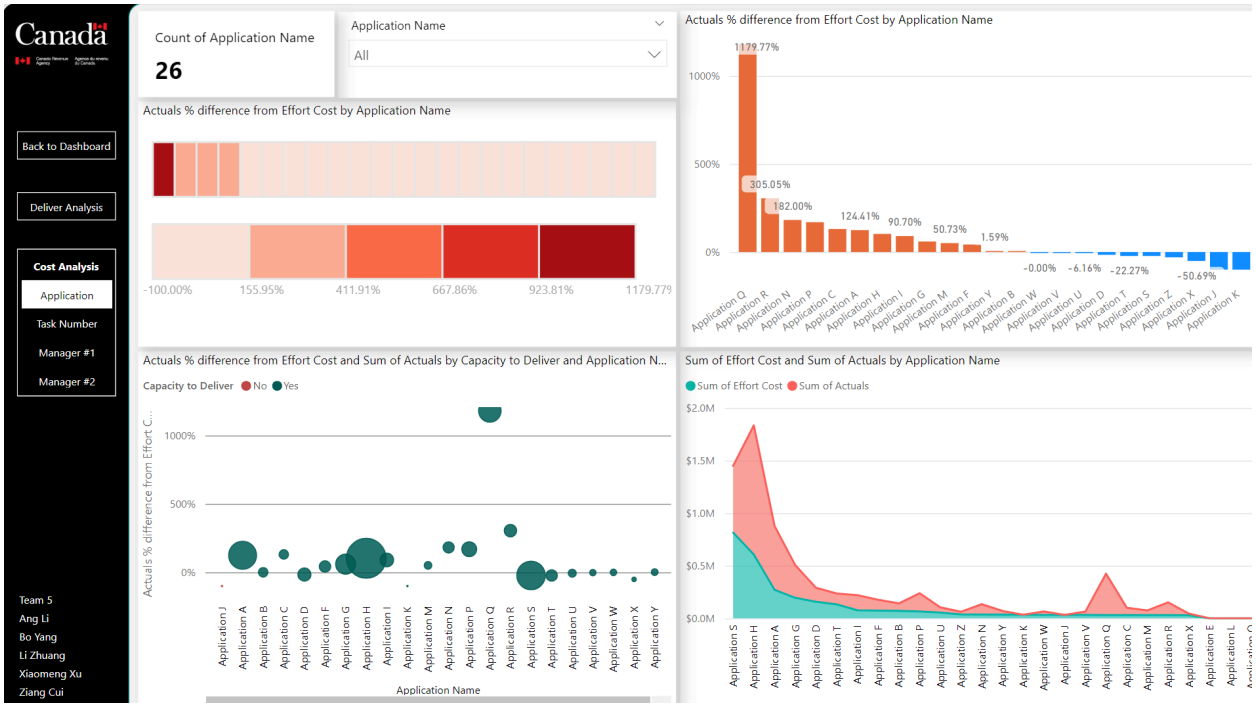

**Work Classification and Delivery Capacity**

Work classification seems to play a lesser role in predicting the successful delivery of projects. Although there are differences in the delivery rates for various work classifications, such as 'Development' and 'Enhancement', it is not a strong predictor. However, when combined with priority, the influence of work classification may become more pronounced, pointing to a more complex decision-making framework where multiple variables interact.



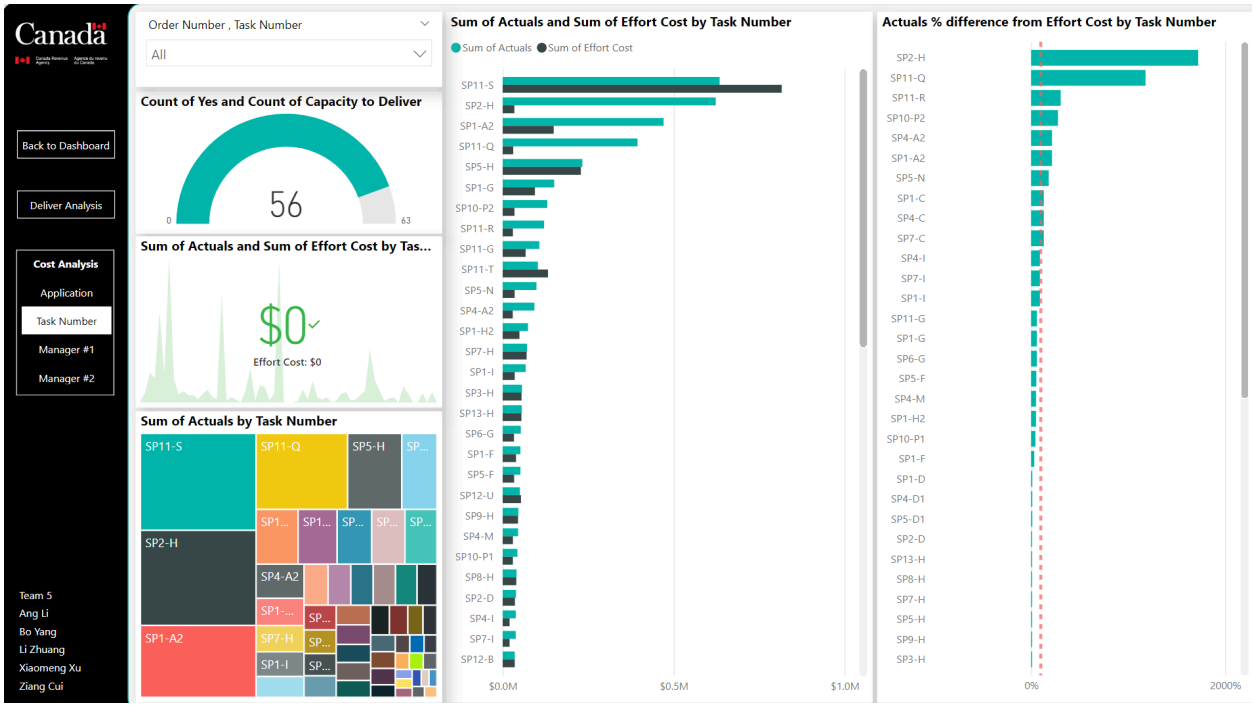

**Departmental Structure**

The data reveal that different departments and divisions display varying performances in project delivery, with certain departments achieving higher success rates. These discrepancies likely reflect differences in management practices, team structures, skills levels, and resource allocations across departments.



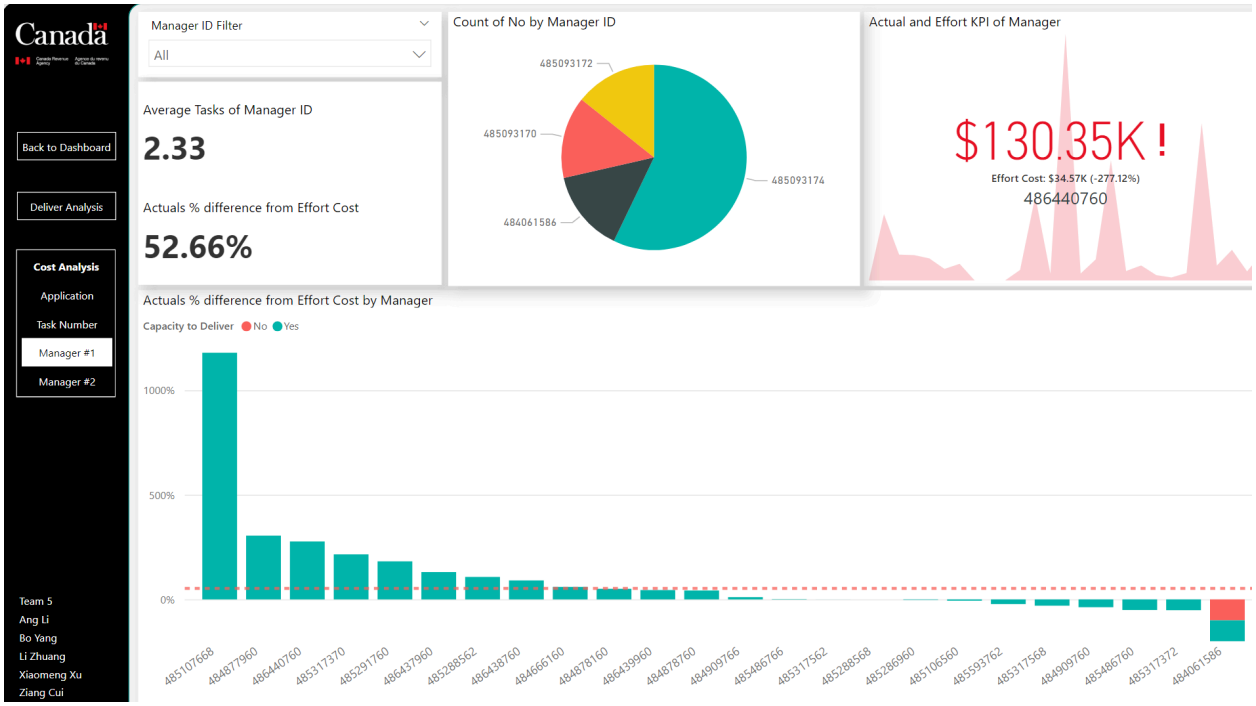
**Project Cost Variance**

Managers and application names have a lower significance in predicting cost variances, suggesting that budget overruns are possibly related to factors like project planning, changes during implementation, market condition fluctuations, or technical challenges. This also suggests that additional factors influencing cost variances, such as project scale, complexity, or team composition, should be explored.




**Task Number and Costs**

There appears to be a substantial disparity between budget and actual costs associated with specific task numbers. This could be due to unforeseen challenges encountered during the implementation of certain tasks or changes as the project evolves.




**Managerial Performance**

We observe differences among various manager IDs in terms of project cost variances and delivery capabilities. This underscores the significant impact of a project manager's skills and experience on the financial and delivery outcomes of a project. Some managers may excel in maintaining the budget, while others may be better at ensuring timely project completion.

**Application Cost-Effectiveness**

The percentage differences between budgeted and actual expenditures for different applications suggest that the cost of developing or maintaining some applications might have exceeded initial estimates or changed during project execution. These changes could be due to factors such as technical difficulty, price fluctuations from external vendors, or underutilization of internal resources.

In conclusion, these visualizations provide deep insights into which factors are most critical for the delivery capacity and cost control in IT project management. Understanding these key elements can help organizations more effectively plan and implement IT projects and also provide data-backed support for potential future interventions.

## Analysis

Based on the visualization data provided from the Power BI reports, we can deduce key findings regarding the determinants of 'capacity of deliver' and 'cost difference' within the IT project management at the Canada Revenue Agency.

### Capacity of Deliver

The 'capacity to deliver' an IT project successfully was measured against various factors such as priority, work classification, and the department within the agency.

| | Priority | Work Classification | Department |
|---|---|---|---|
| Major influencer | ✅ | ❌ | ✅ |
| Impact multiplier | 6.19 | - | 41.25 |

※ Capacity of Deliver could be Predict well

1. **Priority**: Projects classified as 'Critical' and 'High' priority had a better 'capacity to deliver' rate compared to 'Medium' and 'Low' priority projects. This suggests that resource allocation and management focus are effectively aligned with the project's prioritization, bolstering successful outcomes.

2. **Work Classification**: The differentiation between 'Development' and 'Enhancement' work classifications showed that the type of IT work has less predictive power on delivery capacity. This indicates that regardless of the nature of IT work, other factors are more influential in determining the successful delivery of projects.

3. **Department**: The 'capacity to deliver' was found to be significantly influenced by the department structure, with Directorates and Divisions showing varied performances. For instance, certain divisions within the IT Directorate have demonstrated a higher success rate in

project delivery, suggesting that organizational structure and departmental practices have substantial impacts on project outcomes.

## Cost Difference

| | Manager | Application Name |
|---|---|---|
| Major influencer | ✗ | ✗ |
| Impact multiplier | - | - |

※ Cost Difference could NOT be Predict well

The analysis of cost variance focused on the actual spend versus the budgeted amount for IT projects.

1. **Manager and Application Name**: The investigation into cost differences revealed that individual managers and specific applications were not significant predictors of budget variance. This finding suggests that while managerial practices and application types are essential aspects of project management, they do not substantially contribute to budget deviations.

2. **Task Number**: Further examination into cost variance by task numbers highlighted that specific tasks experienced greater discrepancies between budgeted and actual expenditures. It implies that certain tasks within projects are prone to cost overruns, which could be due to unforeseen complexities or scope changes during the project lifecycle.

## Conclusion

The visual analytics from the Power BI reports conclude that the 'capacity to deliver' is chiefly influenced by the priority level and the department's structure, with the department being a major influencer and having a high impact multiplier. However, 'cost difference' could not be predicted as effectively by the manager or the application name, indicating that cost overruns are influenced by other unexamined factors. It is recommended for future analysis to consider other variables such as project complexity, scope changes, and resource availability to gain more insights into the factors affecting cost variance in IT project management.

# Machine Learning

## Objective

The objective for this step is to predict 2 values, one is based on the difference of effort cost and actual cost, another one is based on delivering success.

Output1: Cost Difference Ratio = (Actual Cost - Effort Cost) / Effort Cost
Output2: Capacity to Deliver

## Data Source

The data source is supported by Canada Revenue Agency clients and based on real world project data. It is synthetic and modified by clients for security and privacy reasons.
But there is a defect: the whole dataset is only 64 rows and 12 columns. This feature limits the development to some extent, but we can still conduct the methodology based on the small dataset.

The attribute includes Order Number, Change Request Title, Work Classification, Priority, Application Name, IT Directorate, IT Division, IT Section, Manager ID, Effort Cost, Capacity to Deliver, Actuals.

| Order Number | Change Request Title | Work Classification | Priority | Application Name | IT Directorate | IT Division | IT Section | Manager ID | Effort Cost | Capacity to Deliver | Actuals |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 161454 | Sub-Project-1 | Development | Low | Application A | DIR-1 | Div-1 | Sec-1 | 485317370 | 149149.4219 | Yes | 470545.583 |
| 161454 | Sub-Project-1 | Development | Low | Application B | DIR-2 | Div-2 | Sec-2 | 485317562 | 35203.48047 | Yes | 8830.17 |
| 161454 | Sub-Project-1 | Development | Low | Application C | DIR-1 | Div-3 | Sec-3 | 486437960 | 30286 | Yes | 69890.00667 |
| 161454 | Sub-Project-1 | Development | Low | Application D | DIR-3 | Div-4 | Sec-4 | 484909766 | 34726 | Yes | 13018.63333 |
| 161454 | Sub-Project-1 | Development | Low | Application E | DIR-4 | Div-5 | Sec-5 | 485093174 | 0 | No | 0 |
| 161454 | Sub-Project-1 | Development | Low | Application A | DIR-1 | Div-1 | Sec-6 | 485317372 | 30544 | Yes | 14789.15059 |
| 161454 | Sub-Project-1 | Development | Low | Application F | DIR-1 | Div-5 | Sec-7 | 484878760 | 39062.37988 | Yes | 51832.86 |
| 161454 | Sub-Project-1 | Development | Low | Application G | DIR-1 | Div-6 | Sec-8 | 484666160 | 94317.5 | Yes | 150693.8376 |
| 161454 | Sub-Project-1 | Development | Low | Application H | DIR-1 | Div-1 | Sec-9 | 484878160 | 49126 | Yes | 74047.01104 |
| 161454 | Sub-Project-1 | Development | Low | Application I | DIR-1 | Div-3 | Sec-10 | 486438760 | 35262 | Yes | 67243.40739 |
| 161454 | Sub-Project-1 | Development | Low | Application H | DIR-2 | Div-7 | Sec-11 | 485288562 | 29437 | Yes | 30002.12558 |
| 161097 | Sub-Project-2 | Enhancement | Medium | Application B | DIR-2 | Div-2 | Sec-2 | 485317562 | 35203.48047 | Yes | 8830.17 |

Only 2 columns are numerical values: Effort Cost and Actuals. Others are all categorical values. This is another feature for the dataset.

## Machine Learning Model:

Several machine learning steps were conducted for this project.


In the next, some core steps will be introduced.

### Data Observation:

Some basic python code was used to observe the data.

For example: df.shape, df.info(), df.duplicated(), etc.

```
In [3]: df.shape
Out[3]: (64, 12)
```

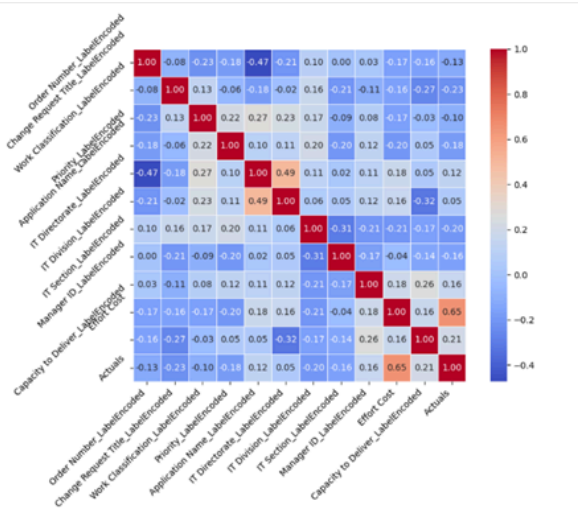
```
In [4]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Order Number          64 non-null     int64
 1   Change Request Title  64 non-null     object
 2   Work Classification   64 non-null     object
 3   Priority              64 non-null     object
 4   Application Name      64 non-null     object
 5   IT Directorate        64 non-null     object
 6   IT Division           64 non-null     object
 7   IT Section            64 non-null     object
 8   Manager ID            64 non-null     int64
 9   Effort Cost           64 non-null     float64
 10  Capacity to Deliver   64 non-null     object
 11  Actuals               64 non-null     float64
dtypes: float64(2), int64(2), object(8)
memory usage: 6.1+ KB
```

```
In [5]: df.duplicated()
Out[5]: 0     False
        1     False
        2     False
        3     False
        4     False
              ...
        59    False
        60     True
        61    False
        62    False
```

Also, visualization methods were used to observe the data in another way.

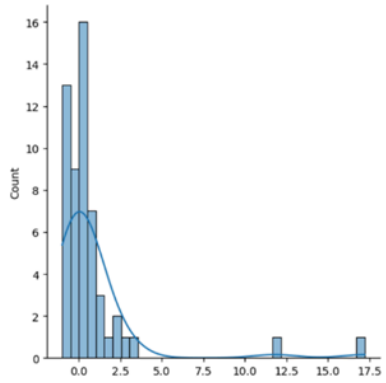
For example: Heat map, Pie chart, Bar chart, Scatter chart, etc.

```
corr_matrix = df_en.corr()

# Plot the heatmap
plt.figure(figsize=(10, 8))  # Set the figure size (optional)
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap='coolwarm',
            cbar=True, square=True, linewidths=.5)

# Optional: Adjust the layout
plt.xticks(rotation=45, ha="right")
plt.yticks(rotation=45)
plt.tight_layout()  # Adjust the layout to make room for the rotated x-axis labels

# Show the plot
plt.show()
```

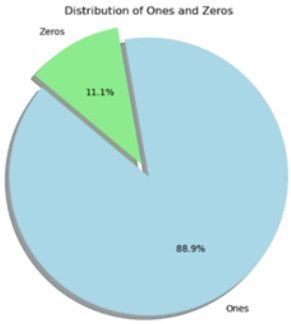


```
#dist plot for cost difference ratio
sns.displot(y, kde=True)
```

```
<seaborn.axisgrid.FacetGrid at 0x2de367b7010>
```

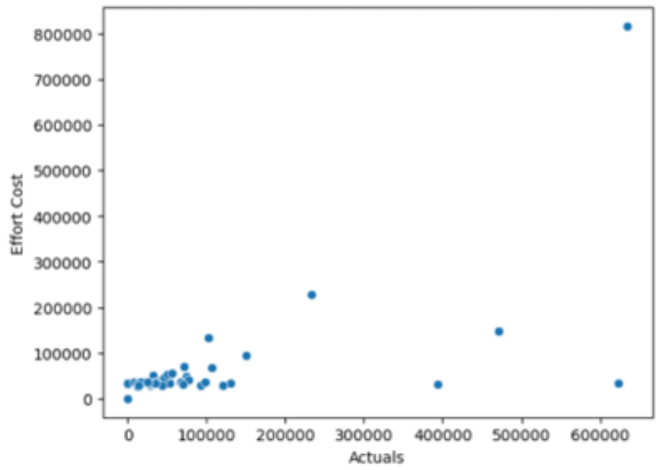



```
labels = ['Ones', 'Zeros']
sizes = [count_of_ones, count_of_zeros]
colors = ['lightblue', 'lightgreen']
explode = (0.1, 0)  # Explode the first slice (Ones)

# Plot
plt.figure(figsize=(6, 6))
plt.pie(sizes, explode=explode, labels=labels, colors=colors, autopct='%1.1f%%', shadow=True, startangle=140)
plt.title('Distribution of Ones and Zeros')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()
```




```
sns.scatterplot( data=df_en, x='Actuals', y='Effort Cost')
```

```
<Axes: xlabel='Actuals', ylabel='Effort Cost'>
```




## Model Selection:

The size of the dataset is small. There are 10 categorical columns and 2 numerical columns. The values in the numerical column have large range values. Based on these features, we dropped the neural networks model first.

For the output of Capacity to Deliver, it is a classification problem. The output value is yes or no, numerically 1 or 0 on behalf. Based on the data features, we chose the Logistic Regression model, Decision Tree model, and Random Forest model.

For the output of Cost Difference Ratio, it is a Regression problem. The output value is a ratio. Based on the data features, we chose Linear Regression model, Decision Tree model, and Random Forest model.

**Model Selection Consideration Matrics**

| | Problem Type | Data Features | Models |
|---|---|---|---|
| Capacity to Deliver | Classification Forecast Yes or No | • Data Size: Small<br>• Data Dimensions<br>10 categorical columns<br>2 numerical columns<br>• Data Value<br>Numerical columns have large range values | • Logistic Regression<br>• Decision Tree<br>• Random Forest |
| Cost Difference Ratio | Regression Forecast Ratio | | • Linear Regression<br>• Decision Tree<br>• Random Forest |

### Input/Output Configuration:

For output as Cost Difference Ratio, most of the columns were used as input except Order Number and Actuals. Order number indicates the same thing with Change Request Title, so it was deleted. As for eliminating the Actuals, it is because the output also contains both actuals and effort cost. If both values are employed as input, it doesn't make sense for Machine Learning.

For output as Capacity to Deliver, most of the columns were used as input except Order Number, Capacity to Deliver and Actuals. In this case, the Order Number and Capacity to Deliver was eliminated for the same reason as the first case, order number is a repeating item, Capacity to Deliver is the output. But why do we eliminate the Actuals? It is because in the real world, the actual value can only be obtained after the project is finished. When we estimate the Capacity to Deliver, we won't have the Actual cost value. This is a future value for the model. If we use it in the model, it is a kind of data leakage in machine learning. So the Actuals attribute was eliminated.

### Model Build up:

Take the random forest model as an example.

Before building up a model, use train_test_split to split the data into x_train, x_test, y_train, y_test. Then use the StandardScaler to standardize the data which can help with the prediction.

There is an integrated library sklearn.ensemble which can help build up both the classification and regression model directly.

```
# create an instance of the model
rf = RandomForestClassifier(n_estimators=200, criterion='gini')
```

```
rf = RandomForestRegressor(n_estimators=100, random_state=42)
```

### Model Tuning Effort:

Model tuning adjusts machine learning model settings using training and testing data, aiming to improve performance by finding the best configuration. Several methods we used in this part:

1. Hyperparameter Tuning:
   - Changed max depth of the tree for Decision Tree
   - Changed number of trees for Random Forest

2. Encoding:
   - Label Encoding
   - One-hot Encoding
   - Target Encoding

3. PCA - condenses complex information by focusing on its main trends, making it easier to grasp and utilize

4. K-Fold Cross Validation:
   - Split the dataset into k number of subsets (known as folds)
   - Perform training on all the subsets.

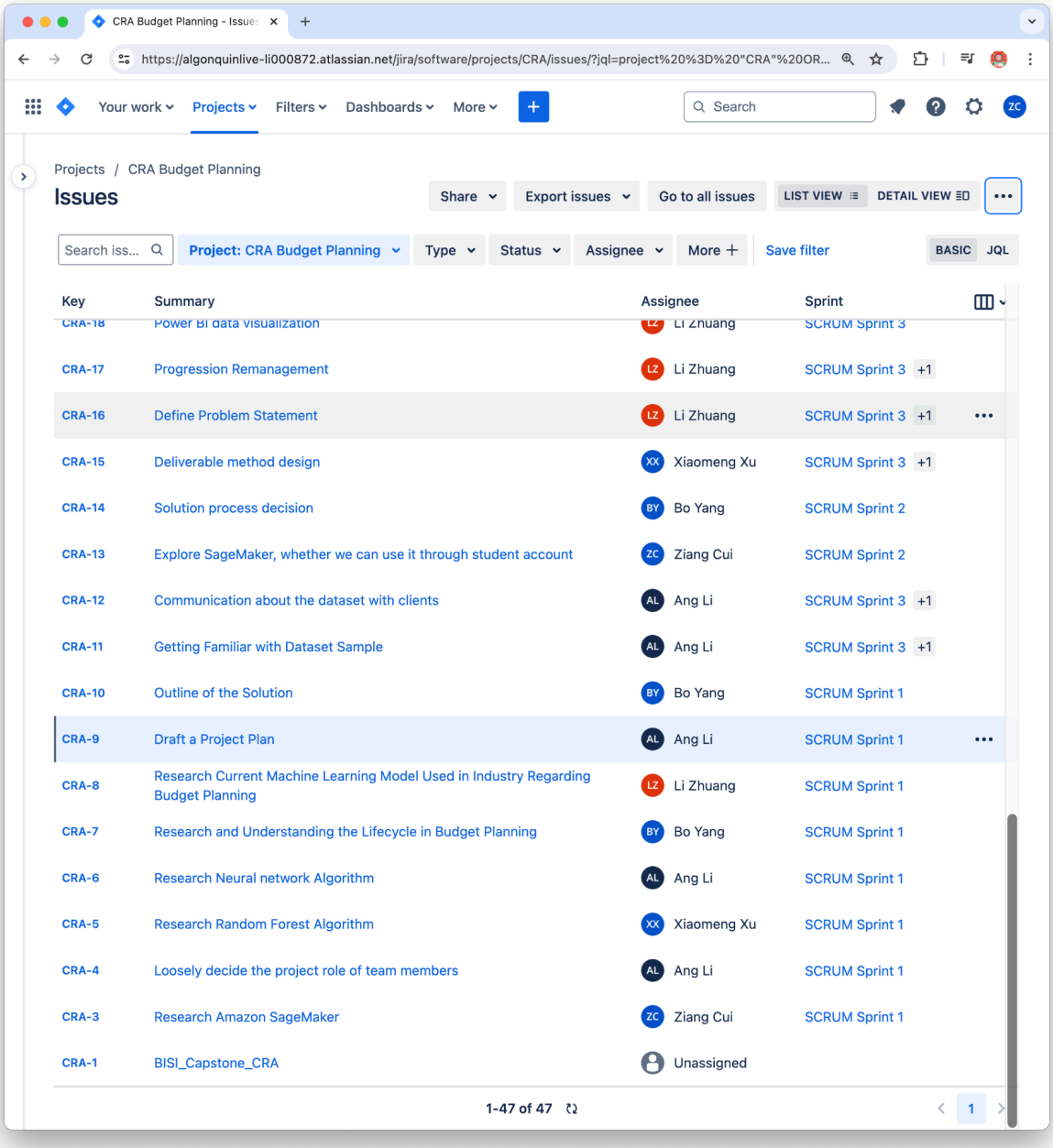
# Part E. Jira progress

We use jira as our project management tool to assign tasks to each team member. It also serves as a display tool for project progress to the clients..
All of the team members are the scrum team members. Specifically, Professor Jim plays the role of the Product Owner, Ziang and Ang the scrum master.
In the long term, we also built a roadmap for the project at the very beginning as below, which made the entire project move in the right direction.

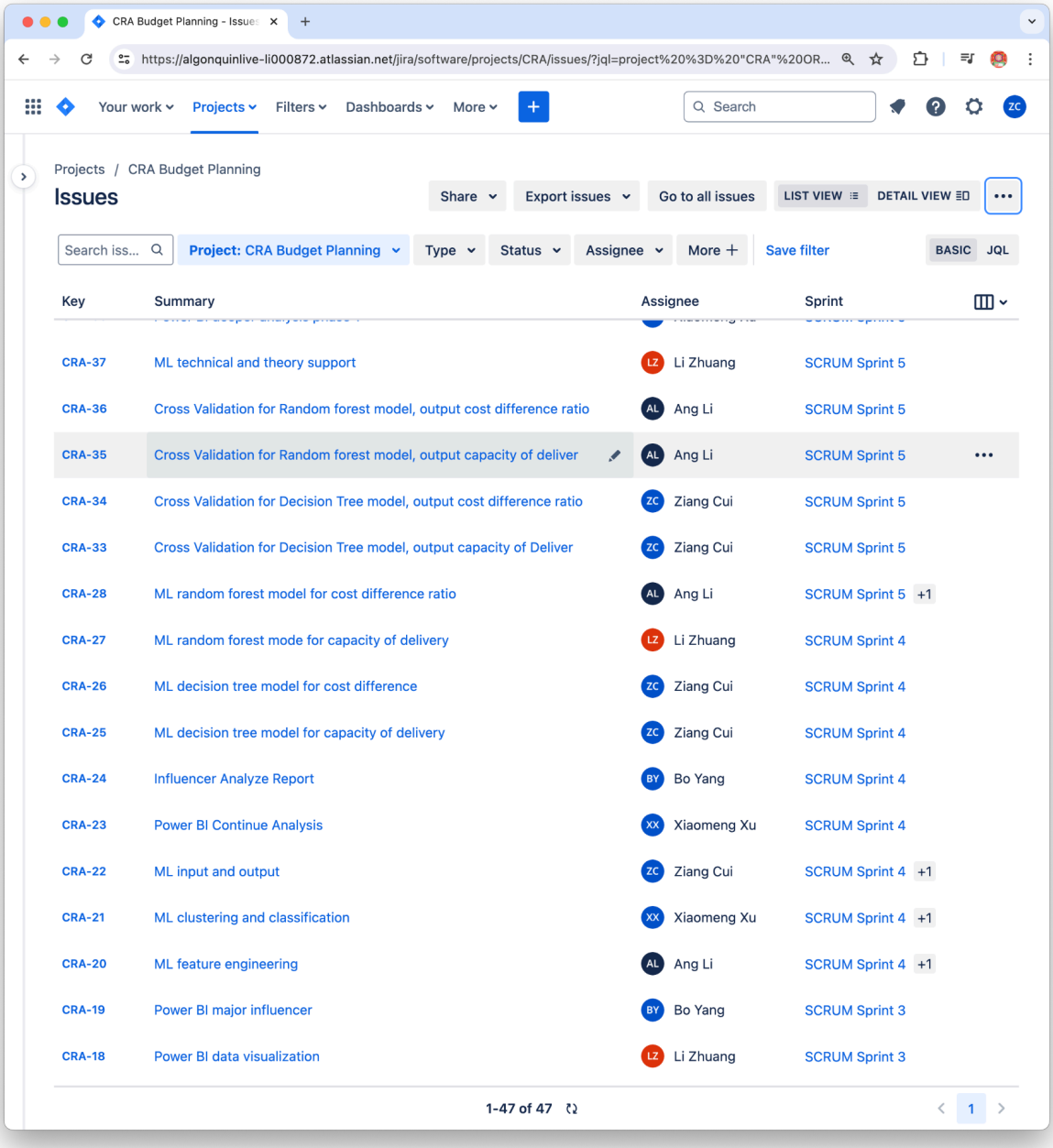
| Week | Goal |
|---|---|
| 1 | Research |
| 2 | Research, Understand the Data,Data Preparation |
| 3 | Feature Engineering, Data Visualization |
| 4 | Choose ML Algorithm, ML Test |
| 5 | Machine Learning Fine Tuning |
| 6 | Insights Analysis |
| 7 | Report, Dashboard |
| 8 | Final Report |

At a weekly level, based on agile methodology, we arrange sprint planning meetings and sprint retrospectives every week to plan tasks for next week and review tasks from the previous week. After each sprint begins, team members select tasks related to themselves based on their abilities and roles.

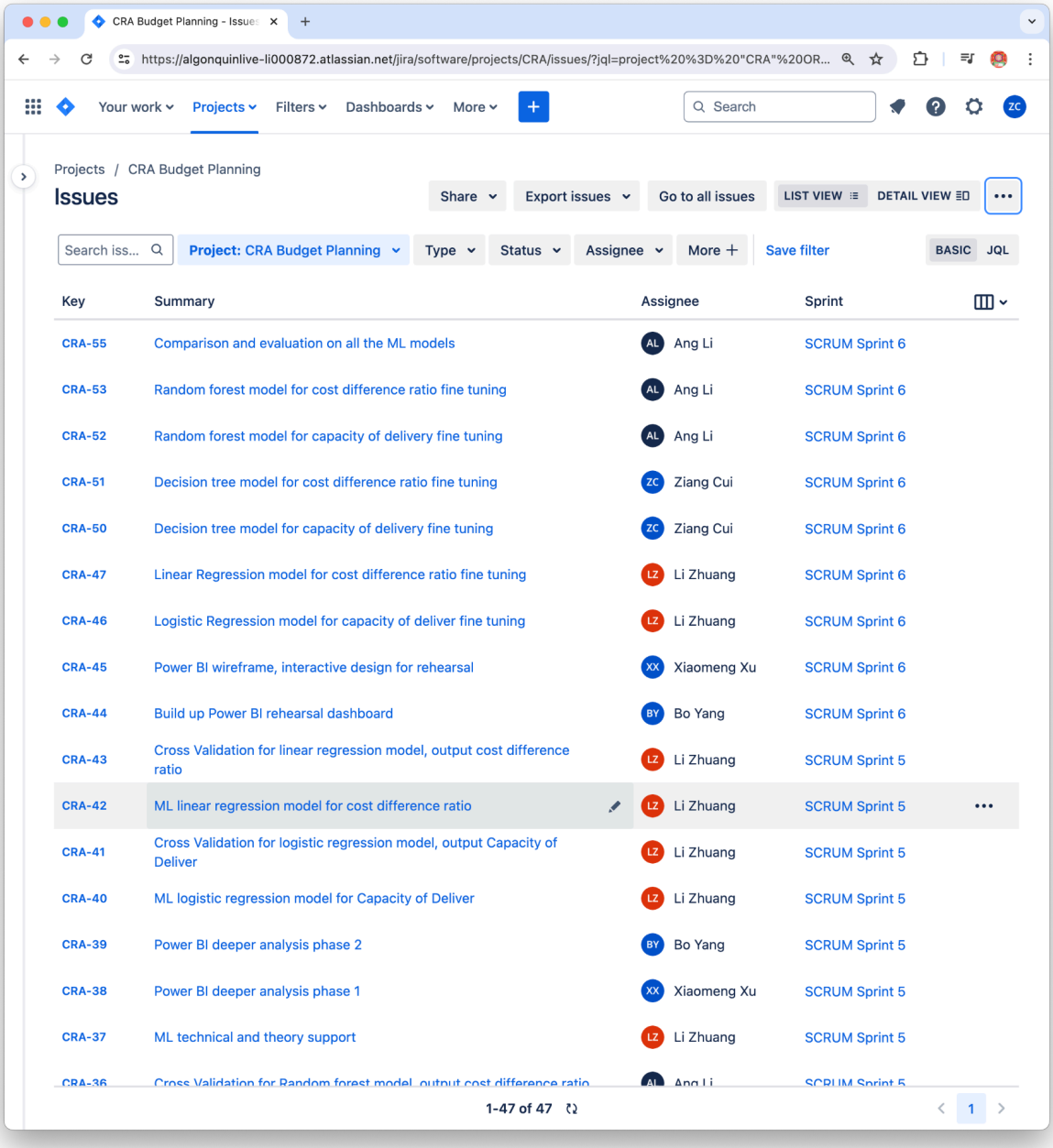
The screenshots below demonstrates all the details of the issues, including the assignee, sprint number etc.




Issue 1 to issue 18




Issue 19 to issue 37




Issue 41 to issue 55

Although we have delays due to various factors during the project, we believe this is normal for an agile project.

# Part F. Results, Analysis, Proposed Data Story, Value, Cautions, Lessons Learned

## Power BI

### Proposed Data Story:

Our narrative begins by understanding that every IT project is a unique tapestry woven from numerous threads such as priority, departmental influence, managerial prowess, and the nature of the task itself. The Power BI visualizations serve as a window into this complex interplay. High-priority projects and those housed within certain departments consistently showed higher delivery success rates, painting a picture of a well-oiled machine where importance aligns with performance. However, the nuanced tale of cost variances unravels a different thread, one that is not as straightforward and is less influenced by managerial oversight or the type of application being developed.

### Value:

The value of this analysis lies in its ability to inform and transform. By recognizing which factors are significant contributors to success, the Canada Revenue Agency can strategically target its resources and training to bolster the areas that yield the highest impact. Knowing what does not heavily influence cost variance is equally valuable, suggesting a reevaluation of current budgeting practices and a potential pivot towards a more agile and adaptable approach to project cost estimation.

### Lessons Learned:

From this exploratory journey, several lessons emerge:

1. Priority setting can make or break a project's timely delivery, suggesting that strategic decisions at the project initiation stage are critical.

2. The impact of departmental structures on project success rates underscores the importance of cohesive and supportive departmental cultures that foster project completion.

3. Managerial impact on cost but not on delivery capacity indicates that project success involves more than just effective leadership; it requires a supportive infrastructure and a responsive system to change.

4. The minimal impact of work classification on delivery capacity teaches us that flexibility in resource allocation across various types of work may be more beneficial than rigid structures.

5. The unpredictability of cost variance, regardless of the manager or application, teaches that budgeting for IT projects may need more contingency planning and real-time adjustments.

In summary, the proposed data story and the lessons learned provide a roadmap for future projects, highlighting where to maintain course and where to steer towards uncharted but potentially more successful waters.

# Machine Learning

## Models Evaluation and Comparison

Model Evaluation Metrics for Classification Problem:

Accuracy, Precision, Recall, F1-score: Higher value is better. Range from 0 to 1.

Model Evaluation Metrics for Regression Problem:

Mean Squared Error (MSE), Mean Absolute Error (MAE): Lower value is better.

Range from 0 to ∞.

R2: Higher value is better. Range from -∞ to 1.

| Output=capacity of deliver | | | |
|---|---|---|---|
| | Logistic Regression | Decision tree | Random forest model |
| Accuracy | 0.95 | 0.84 | 0.94 |
| Precision | 0.98 | 0.94 | 0.97 |
| Recall | 0.96 | 0.87 | 0.97 |
| f1 score | 0.97 | 0.92 | 0.96 |
| Best Model | √ | | |

| Output=cost difference ratio | | | |
|---|---|---|---|
| | Linear Regression | Decision tree | Random forest |
| MSE | 8.01 | 7.74 | 8.10 |
| MAE | 1.59 | 1.08 | 1.20 |
| R2 | -3.74 | -0.41 | -0.46 |
| Best Model | | √ | |

Although it seems like the best model was selected from both problems. But as you look at the values, for the output as cost difference ratio, all these three model's R2 value is smaller than 0.

This suggests that the model performs worse than simply predicting the mean of the target variable. It indicates that the model is poorly fitted to the data. So, we can not find a good model to do the prediction for output as cost difference ratio.

## Conclusion

Finally, the conclusion is that for the output as Capacity to Deliver, it can be predict by using our Logistic Regression Predict Model. But for the output as the Cost Difference Ratio, it can not be predicted by our models.
There are several reasons why it is hard to build up a model to predict the cost difference ratio.

1. The complex relationship between effort cost and actual cost.
2. The dataset is too small.
3. Most of the categorical data is less correlation with cost values.
4. The dataset is recorded by different managers, the recording method for each manager is different.

## Challenges and Limitation

One of the major challenges is the data quality of the project recording data. The foundation of taking advantage of Machine Learning is the large amount of accurate and real data. However, most of the organizations are facing problems in collecting, recording, maintaining, abstracting data in a standard and proper way. Thus, the quality and consistency of data can not be ensured to train the ML model. The performance can be impacted significantly by using these uncertainty values. The example project was constrained by using the small dataset and the bad quality of data.

# Part G. Future Perspective

When predicting the cost difference ratio, we often face challenges related to dataset size. Despite our best efforts to optimize regression models for forecasting, the accuracy of our predictions remains limited due to the constraints imposed by a small dataset. Under these circumstances, it is difficult to capture the full complexity and variability of the data, which can significantly impact the precision of our forecasts.

In light of these challenges, it is advisable to consider potential modifications for future projects, especially when larger datasets become available. A larger dataset would provide a more robust foundation for the analytical models, allowing for a better representation of trends and anomalies. With more data, we can re-evaluate and potentially modify the existing models or explore alternative models better suited for the expanded dataset.

Additionally, adjusting the parameters of the current models might also yield better results. By doing so, we can refine the model's ability to learn from a larger quantity of data without overfitting. This approach would likely enhance the model's predictive accuracy and reliability.

In summary, while we strive to achieve the best possible outcomes with the current dataset, embracing opportunities to incorporate larger datasets and revising our modeling strategies will be crucial in improving the precision of our cost difference ratio predictions in the future.