MONTANA
Department of Transportation
Research

ARTIFICIAL INTELLIGENCE (AI) BASED TOOL TO ESTIMATE CONTRACT TIME

FHWA/MT-23-001/9929-819

Final Report

September 2023

prepared for

THE STATE OF MONTANA DEPARTMENT OF TRANSPORTATION

in cooperation with

THE U.S. DEPARTMENT OF TRANSPORTATION
FEDERAL HIGHWAY ADMINISTRATION

September 2023

prepared by

David Jeong, PhD
Hamed Alikhani

Texas A&M University
College Station, TX

Artificial Intelligence (AI) Based Tool to Estimate Contract Time

Prepared by

David Jeong, PhD
Professor
Construction Science
&
Hamed Alikhani,
Graduate Research Assistant

Prepared for the
MONTANA DEPARTMENT OF TRANSPORTATION
in cooperation with the
U.S. DEPARTMENT OF TRANSPORTATION
FEDERAL HIGHWAY ADMINISTRATION

September 2023

A report from
Texas A&M University
College Station, TX 77843-3137
Ph: (O) 979-458-9380,
Email: djeong@tamu.edu

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA/MT-23-001/9929-819		

4. Title and Subtitle Artificial Intelligence (AI) Based Tool to Estimate Contract Time	**5. Report Date** September 2023 **6. Performing Organization Code**
7. Author(s) H. David Jeong (https://orcid.org/0000-0003-4074-1869), Hamed Alikhani (https://orcid.org/0000-0003-2937-2038)	**8. Performing Organization Report No.**
9. Performing Organization Name and Address Texas A&M University, 400 Bizzell St, College Station, TX 77843	**10. Work Unit No.** **11. Contract or Grant No.** MDT Project # 9929-819
12. Sponsoring Agency Name and Address Research Programs Montana Department of Transportation (SPR) 2701 Prospect Avenue PO Box 201001	**13. Type of Report and Period Covered** Final Report (April 2021 – Sept 2023) **14. Sponsoring Agency Code** 5401

15. Supplementary Notes

Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration. This report can be found at https://www.mdt.mt.gov/research/projects/const/ai_based_contracting_tool.aspx DOI: https://doi.org/10.21949/1518324

Recommended Citation: Jeong, H. David, and Alikhani, Hamed (2023). Artificial Intelligence (AI) Based Tool to Estimate Contract Time. Helena, MT: Montana Department of Transportation. DOI: https://doi.org/10.21949/1518324

16. Abstract

MDT is in the process of modernizing their contract time determination processes by developing user-friendly tools to facilitate the estimation process of project duration and contract time. As part of this modernization effort, a top-down project duration estimation tool was developed in this research project. This tool is particularly useful when there is limited project information available during the early preconstruction stages. This research project involved the development of an Artificial Intelligence (AI) based model that can predict the most probable duration of a construction project using an early cost estimate, major controlling work items, and their estimated quantities as input values. Additionally, a regression model with the same set of input variables was developed as a companion to the AI model. The models were trained and tested using historical project data of more than 1,000 highway projects from 2008 to 2019. To operationalize the models, a user-friendly Microsoft Excel tool named AI-PDET (Artificial Intelligence based Project Duration Estimation Model) was created. AI-PDET can be used throughout the preconstruction phases to quickly determine a reasonable project duration for proper project planning and delivery. Furthermore, it can serve as a reality check tool alongside bottom-up tools during the procurement stage. This report also provides specific guidance on updating the models and the database with new project data in AI-PDET.

17. Key Words	18. Distribution Statement
Contract time determination – project duration estimation – highway projects – Artificial Neural Networks – Regression analysis - daily work report — data analysis	No restrictions. Enter any other agency mandated distribution statements. Remove NTIS statement if it does not apply.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	53	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Disclaimer Statement

This document is disseminated under the sponsorship of the Montana Department of Transportation (MDT) and the United States Department of Transportation (USDOT) in the interest of information exchange. The State of Montana and the United States assume no liability for the use or misuse of its contents.

The contents of this document reflect the views of the authors, who are solely responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the views or official policies of MDT or the USDOT.

The State of Montana and the United States do not endorse products of manufacturers.

This document does not constitute a standard, specification, policy or regulation.

Alternative Format Statement

Alternative accessible formats of this document will be provided on request. Persons who need an alternative format should contact the Office of Civil Rights, Department of Transportation, 2701 Prospect Avenue, PO Box 201001, Helena, MT 59620. Telephone 406-444-5416 or Montana Relay Service at 711.

Table of Contents

List of Figures

List of Tables

1 Introduction

Getting a construction project done on time is a major performance goal that many DOTs including Montana DOT (MDT) constantly pursue and monitor. However, most DOTs continue to struggle to meet the schedule performance target of their highway projects. For example, in 2017, approximately $144.5 millions of road projects in Montana experienced schedule delay (Fraser, 2017). Estimating and tracking construction project duration is crucial in the project development process since it not only directly affects the agency's key performance indicator, but also affects the contractor selection, construction costs, construction quality, safety, and public satisfaction. Both unreasonably short or long construction contract times can result in negative consequences such as high bid prices, lack of qualified bidders, poor work quality, claims and disputes, prolonged inconvenience to the traveling public, lack of innovations, increased administration costs, and safety issues (FHWA, 2002; Hildreth, 2005; H. S. Jeong et al., 2009).

MDT is in the process of modernizing their contract time determination processes by developing user-friendly tools to facilitate the estimation of project duration and ultimately the determination of contract time. MDT has successfully developed the Production Rate Estimation Tool (PRET) for controlling activities and visual construction sequence logic diagrams for common types of highway projects (Jeong et al., 2019; Jeong & Alikhani, 2020). These tools are bottom-up tools that can help support specific work tasks during the scheduling and contract time development processes. As a continuation of this modernization effort, there is a need to develop a top-down tool that can estimate a project's duration when a limited amount of project information is available during the preconstruction stages. This tool can be used throughout the preconstruction phases to quickly determine a reasonable project duration for proper project planning and delivery, and it can also be used as a reality check tool along with the bottom-up tools during the procurement stage.

In recent years, artificial intelligence (AI) technologies have improved their technical capabilities for pattern recognition and prediction. Promising AI techniques such as Artificial Neural Networks (ANNs) are capable of processing various types of data and learning complex patterns to make a prediction with reliable accuracy. An AI-based data-driven model can leverage historical project characteristics and performance data to estimate a reliable project duration for a new project. In this research, an AI-based model and its tool were developed using historical highway project data. The model identifies the most influential factors that affect project duration such as estimated construction cost, major controlling work items and their quantities. It uses those factors as input variables to estimate the project duration with a certain level of confidence. Additionally, a regression model with the same set of input variables was developed as a companion to the AI model. A robust Microsoft Excel implementation tool was also developed to support a quick and reliable estimation of a project's duration.

1.1 Project Objectives

The goal of this research project was to develop a) an AI-based estimation model that takes key highway project characteristics and estimates a reliable project duration, and b) a Microsoft Excel-based tool that can operationalize the AI model. The specific objectives of this research project include i) obtain and analyze historical project data, ii) identify the most influential factors that affect the duration of highway projects, iii) develop an AI-based project duration estimation model and validate the results, and iv) develop an MS Excel-based tool that provides a user-friendly interface for using the AI model.

1.2 Work Tasks

To achieve the goals of this project, the following work tasks were performed.

1.2.1 Task 1: Critical Review of Current Leading Practices

The research team reviewed current top-down project duration estimation methods developed by some DOTs such as the Kentucky Transportation Cabinet, Indiana DOT, Ohio DOT, and Colorado DOT. The strengths and weaknesses of those methods in terms of estimation accuracy, technical approach, data requirements, and user interface were analyzed and documented. The research team also reviewed the state-of-the art AI techniques such as artificial neural networks to identify the most effective algorithm for project duration estimation. The analysis and review results of the current top-down project duration estimation methods used by some DOTs were utilized in evaluating the feasibility and suitability of different types of AI techniques for this research.

1.2.2 Task 2: Data Collection, Preliminary Analysis, and Meeting with MDT Schedulers

In this task, the research team obtained the historical project data from MDT. The obtained data were cleaned, processed, normalized, and organized to be suitable for this research. The research team first applied explanatory statistical methods to understand the collected data, determine data characteristics and define any visible patterns of data attributes such as correlation. Statistically significant variables were identified in this process and these variables became candidate variables for developing an AI-driven project duration estimation model.

The research team conducted a virtual meeting with MDT schedulers and representative district engineers to discuss the preliminary findings of the research and obtain their feedback. The meeting was used to confirm the influential factors identified for project duration estimation. The research team demonstrated a preliminary AI model to obtain their feedback and identify areas of improvement of the model to assure its practicality.

1.2.3 Task 3: AI Model Development

In this task, the findings and results from Tasks 1 and 2 were utilized to develop a fully functioning AI model for project duration estimation. The model receives input variables and passes them to hidden layers, where the input variables are processed together with a non-linear function to predict the output value. To ensure the validity of the models, approximately 80% of the dataset was used for training the model and the remaining dataset was used to validate the reliability and accuracy of the model. A multivariate regression model was also developed with the same set of input variables as a companion of the AI model.

1.2.4 Task 4: Tool Development

The research team developed a user-friendly tool based on the AI and regression models developed in Task 3 that operates in Microsoft Excel environment for easy implementation. The tool is named AI-PDET (Artificial Intelligence based Project Duration Estimation Tool). The research team used Visual Basic for Applications (VBA) in Microsoft Excel to automate computational tasks and create a user-friendly interface. The research team developed a user's manual with real examples to demonstrate how to use the tool to estimate the duration of a project.

1.2.5 Task 5: Guidance on Tool Maintenance and Database Update

A detailed guide on tool maintenance and database updates was developed in this task as the tool needs to stay relevant to new projects. It explains how to obtain and clean the new project data and describes different coding parts and actions that need to be taken to transfer the attributes of the two models to AI-PDET.

1.2.6 Task 6: Final report, Project Webinar, Final Presentation and Implementation Meeting

In this task, a draft of the final report that encompasses all task results, findings, and products was prepared for the technical panel's review. All comments from the technical panel on the draft were incorporated into the final report. Other required deliverables such as the project summary report and the performance measures report were submitted with the approval of the final report. A project webinar along with a final presentation was provided to the technical panel for rapid dissemination of the research findings. Also, an implementation meeting was conducted with the project technical panel to review the research team's implementation recommendations and to determine implementation recommendations. The research team documented the discussion results in the form of an implementation report.

2 Literature Review

The FHWA requires State Transportation Agencies (STAs) to have adequate written procedures for the determination of contract time. Contract time is defined as the maximum time allowed in the contract for completion of all work contained in the contract documents (FHWA, 2002). Current practices of Contract Time Determination (CTD) identified by a recent survey indicated 68% of DOTs across the US participating in the survey had a formal procedure for CTD (Taylor et al., 2017). Fifty three percent of the DOTs have developed agency-specific production rates of controlling work items and 39% of the DOTs use a project-specific sequence logic to estimate contract time.

Two main approaches are mostly used for project duration estimation to estimate the contract time: the bottom-up and the top-down approach. The bottom-up approach develops a pre-construction schedule to compute the total project duration that includes a) estimating the durations of work items using the production rates and b) determining activity relationships using activity sequence logics (FHWA, 2002; Daradkah et al., 2018; Jeong & Alikhani, 2020). DOTs have developed specific tools such as spreadsheet-based production rate estimation tools, production rate adjustment tools for weather and site factors, activity sequencing logic diagrams and contract time determination templates to help the scheduler develop a bar chart or a CPM-based schedule to determine the project duration.

For example, Virginia DOT categorizes their highway projects into six types and uses production rates and sequence logics for estimating project duration (Gondy & Hildreth, 2007). The Kentucky Transportation Cabinet and Texas DOT have developed a series of tools to support production rate estimation and construction activity sequencing using the critical path method concept (Connor, 2004). MDT developed an Excel-based tool to estimate production rates of major controlling activities considering factors affecting production rates such as work item quantity, work hours, work types, different seasons of work, districts, area types (urban/rural), and budget types (Jeong et al., 2019). The tool is used to determine a reliable duration of each controlling activity in project schedule. Also, MDT developed activity sequence logic templates that include major controlling activities for each common project type to help schedulers determine the sequence of activities in project schedules (Jeong & Alikhani, 2020). Although the bottom-up technique provides a reliable estimation of project duration before construction, it requires that the project is well-defined and the project information such as work items and their quantities is known with high certainty. However, such detailed information is not usually available in the early pre-construction stages.

A top-down approach for project duration estimation can be used in the pre-construction phases when the project's scope is not well defined. However, even at this early stage of project development, a DOT needs a reasonable project duration estimation for project planning and budgeting purposes. Therefore, a top-down project duration estimation is desirable during the pre-construction stages when a limited amount of project information is available, and the project design is not finalized. Also, this top-down tool will be handy to check the reasonableness of the project duration and the contract time calculated using the detailed bottom-up methods at the end of the design stage.

Some DOTs such as the Kentucky Transportation Cabinet, Indiana DOT, Ohio DOT, and Colorado DOT have developed a top-down project duration estimation tool based on the fact that there is strong correlation between project duration and key project characteristics such as project type, estimated cost, project location, and bid quantitates (Attal, 2010; KYTC, 2014; Taylor et al., 2017; Ohio DOT, 2020). Regression models were mostly used to establish the statistical relationship between key project parameters and project duration. In a survey, some DOTs reported that the regression method was more accurate and easier to use than their previous contract time estimation methods that are based upon production rates and generic precedence logic diagrams (Taylor et al., 2017). For example, Ohio DOT (2020) developed a regression model for each project type (in total, 19 types) using eight years of project data including project cost, project type, project location, and starting season to estimate project duration. Ohio DOT uses these models to estimate the duration of a project in early preconstruction stages by determining a mean duration with 90% and 95% confidence level. The agency uses such regression tools for preliminary estimation of contract time and uses production rate charts and scheduling tools for final setting of contract time.

2.1 Current Leading Practices

The CTD system of Texas DOT (TxDOT) was initially developed by Hancher et al. (1992) and further improved by Connor (2004) who developed a production rate estimation system. TxDOT categorized its highway projects into 13 project types. The CTD system asks the user to input work item quantities and the program finds a default production rate for each work item. The default production rates of the system can be adjusted for a particular project by the user. The system takes factors such as the location, traffic condition, complexity, soil conditions, and weather that affects the production rates and duration of work items. The system allows the user to determine the relationship of activities to finally generate a bar chart presenting the activities, relationships, and their duration (Hancher et al., 1992; Texas DOT, 2018).

Kentucky Transportation Cabinet (KYTC) (2014) divided highway projects into small size (lower than \$1M) and large size (higher than \$1M) and developed unique regression models for ten project types of small size and five project types of large size. The small projects account for more than 90% of the KYTC highway projects. The regression models require project identification number, construction estimate, letting date, and the selected design project type as input variables. The model returns an estimated lower, mean, and upper range of completion dates and working days with 95% of confidence level. For large projects, the input variables for the model include the construction estimate and key bid item quantities (KYTC, 2014).

A Microsoft Excel tool was developed to facilitate the estimation process as shown in Figure 2.1. Once the input variables are entered, the tool calculates and provides a mean, lower, and upper bound of the estimated project duration.

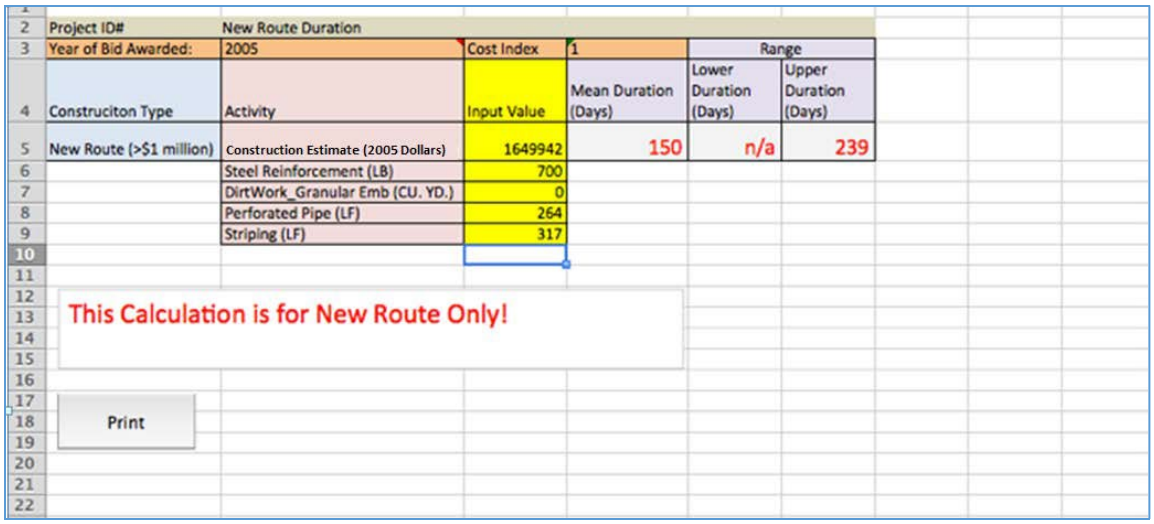


Figure 2.1 Screenshot of KYTC Excel tool (KyTC, 2014)

Ohio DOT (2020) developed a unique regression model for each project type (total of 19 types) using eight years of data. The regression models take multiple project variables such as project cost, project type, project location, and starting season to estimate project duration with 90% and 95% confidence interval. Ohio DOT applies such regression models for a preliminary estimation of contract time and uses production rates and scheduling tools for final setting of contract time that is illustrated in Figure 2.2 (Taylor et al., 2017). The project duration estimation tool is an Microsoft Excel tool that includes three steps. The first step is to input major work items for a new project then the tool automatically calculates the production rates and adjusts them with adjusting factors. The adjusting factors are determined based on factors such as the location of the project (rural/urban), traffic, project complexity, soil condition, and the size of the project. After production rates are calculated, the user inserts the major work items in the Bar Chart and identifies the work item relationships. Step 2 computes the overall duration of the project. Step 3 captures the total duration and the starting month and considers weekends, holidays, and weather days to adjust the project duration and estimate the completion date of the project (Ohio DOT, 2020).

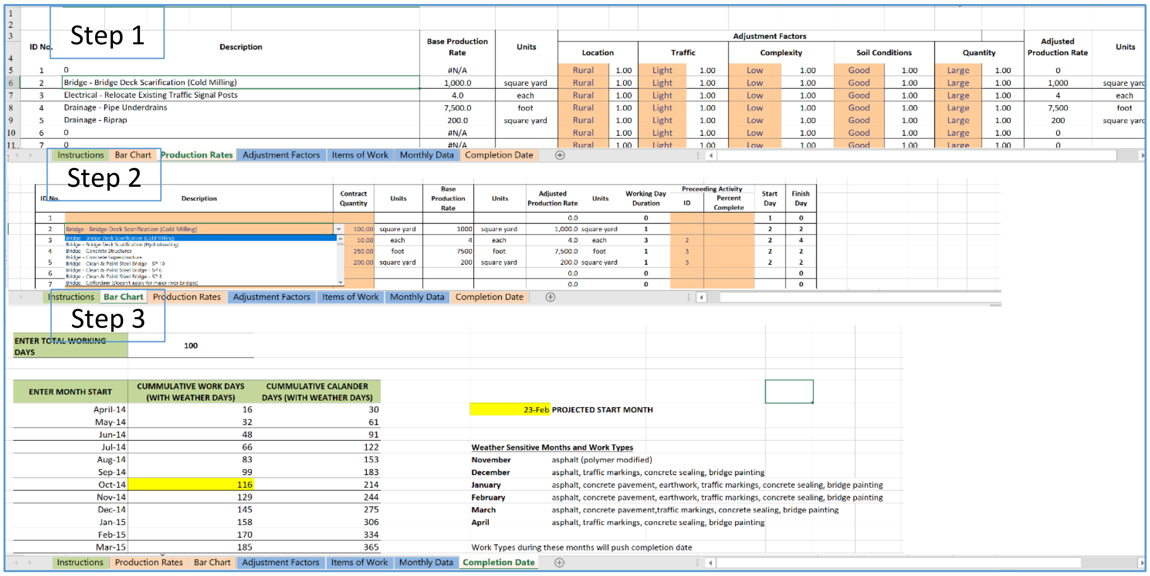


Figure 2.2 Screenshot of Ohio DOT Excel Tool (Ohio DOT, 2020)

Virginia DOT (Gondy & Hildreth, 2007) categorizes its highway projects into six types and uses production rates and sequence logics for CTD. As Figure 2.3 shows, for less complex projects, the sum of the durations for major work items accounts for the project duration. But, for more complex projects, a scheduling technique such as CPM or Bar Chart is used to develop a schedule for CTD after determining durations for major work items.

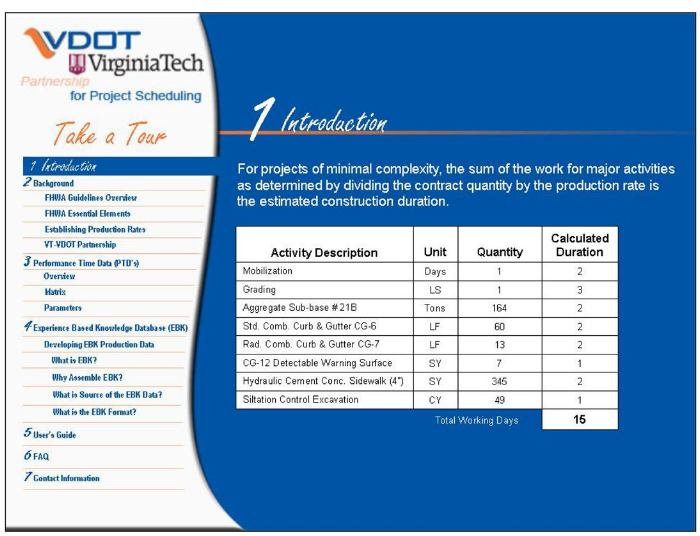


Figure 2.3 Screenshot of Virginia DOT CTD Tool (Gondy & Hildreth, 2007)

Nevett et al. (2020) collected highway project data from Colorado DOT that included information about quantities of work items, item level costs, and durations of 15,000 projects. They analyzed 22 variables and developed a multi linear regression model that can receive ten project characteristics such as project type, cost, traffic condition, and work item quantities to predict project duration.

Jeong et al. (2008) developed a comprehensive automated scheduling system for Oklahoma DOT (OKDOT). The research team categorized OKDOT highway projects into three tiers based on complexity (tier I has the highest complexity) and developed the scheduling system for Tier II and III that account for more than 90% of OKDOT highway projects. They developed a standalone computer application and linked Microsoft Project to a database of project types and production rates in Microsoft Access. The application receives estimated quantities of controlling work items from the user and finds the associated production rates of the work items from the database and computes durations of activities that can be adjusted by the user (Figure 2.4). Then, the application exports the information to a Microsoft Project that includes the pre-established activity sequence logics for different project types to determine the relationship of activities using CPM and create a reliable schedule that can be used as a basis for contract time (Figure 2.5).

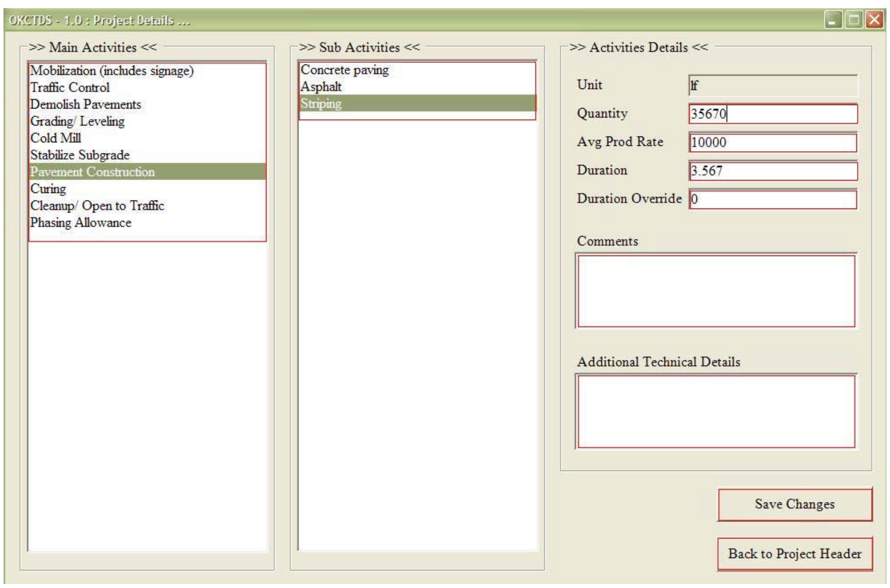


Figure 2.4 Screenshot of OKDOT CTD Tool (Jeong et al., 2008)

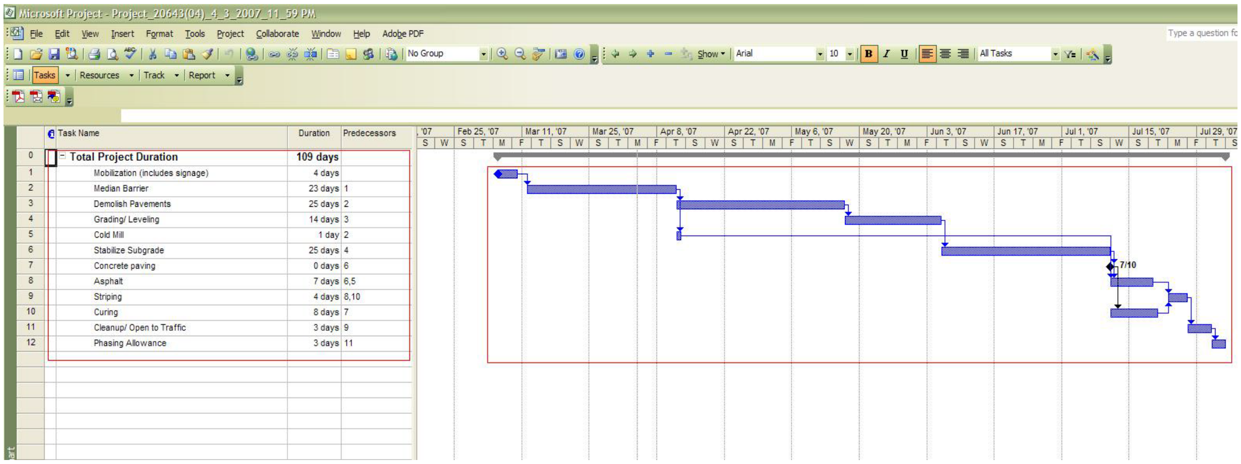


Figure 2.5 Total project duration and CPM diagram in Microsoft Project (Jeong et al., 2008)

2.2 Discussion on the Leading Practices

The top-down approach of project duration estimation is used as a quick and reliable approach in different DOTs for determining the contract time before the beginning of construction. DOTs developed different computer systems that take key project characteristics as the inputs and predict project duration as the output. To operationalize the models, different software programs were developed. A summary on the tools, methods, and project attributes is provided below.

- **Project attributes:** key project characteristics such as project location, type, size, estimated cost, traffic condition, soil condition, weather, as well as work item quantities were considered as input variables to predict the project duration.
- **Software programs:** Microsoft Excel is widely used as the main tool (Ohio DOT, 2020; Taylor et al., 2013; Texas DOT, 2018). Researchers used Visual Basic as a programming tool in Microsoft Excel to further improve the capability of project information analysis in DOTs (Jeong, Shane, et al., 2019). Other software programs such as Lotus, Flash-up, and Microsoft project were used to link the database to the main tool and present the bar chart schedule (Jeong et al., 2008; Texas DOT, 2018). Due to its simplicity, high capabilities, and availability of Microsoft Excel in DOTs, MS Excel will be used in this research as the main tool.
- **Modeling approach:** Most DOTs have used statistical methods such as regression models as the major modeling approach. Some DOTs used a stochastic approach to estimating a range of project durations within a certain confidence level (Ohio DOT, 2020; KYTC,

2014; Taylor et al., 2013). The stochastic approach can provide a more realistic insight of project duration since there are many uncertainties associated with project variables.

2.3 Influential Factors on Project Duration

Recent studies identified key project characteristics as predictors of project duration estimation as shown in Table 2.1. Controlling work items are activities that are highly likely to be on the critical path of a project. Controlling work items and their quantities were identified as key influential factors by previous researchers. Project location is another contributing factor to project duration. For example, traffic congestion in urban areas may prolong the duration of an urban project. Also, difficulties of carrying materials to mountainous areas may also increase the project duration. Project work type is critical in project duration estimation since each project type has its specific activities and sequences. Project's estimated cost is another factor determining the size of the project that is influential in project duration, since larger projects tend to have more expenses, larger material procurement, and more equipment pieces. Weather condition is highly influential. Cold weather can slow down the construction process especially in districts where the winter is long and severe.

Table 2.1 Highway project attributes used for project duration estimation.

No.	Attribute	References
1	Project location	Ohio DOT (2020), Abdel-Raheem et al. (2020), Taylor et al. (2013), Attal (2010), Hegazy and Ayed (1998), Hancher et al. (1992)
2	Project size	Nevett et al. (2020), KYTC (2014), Jeong et al. (2008)
3	Estimated cost	Nevett et al. (2020), Ohio DOT (2020), KYTC (2014), Wilmot and Mei (2005)
4	Controlling work item quantities	Ohio DOT (2020), Mensah et al. (2016), KYTC (2014), Williams and Heldreth (2009), Jeong et al. (2008), Hancher et al. (1992)
5	Scope of work	Attal (2010)
6	Contract execution date	Ohio DOT (2020), Attal (2010)
7	Design method	Hoffman et al. (2007)
8	Project type	Nevett et al. (2020), Ohio DOT (2020), Attal (2010), Jiang & Wu, (2004), Skitmore & Ng (2003)
9	Population of the area	Leu & Yang (1999)

10	Number of lanes	Mahmood et al. (2017), Williams & Heldreth (2009)
11	Traffic condition	Ohio DOT (2020), Nevett et al. (2020), Jiang & Wu, (2004), Hancher et al. (1992)
12	Production rates	Connor (2004), Jiang & Wu (2004)
13	Weather conditions	Ezeldin & Sharara (2006), Jiang & Wu (2004)

Use of project characteristics in a project duration estimation model is highly dependent on the availability of data. In general, applying more project features results in a higher accuracy and reliability of the estimated duration.

2.4 Review of Statistical Methods and Artificial Neural Networks (ANNs)

Project duration estimation using key project characteristics requires a statistical model that can estimate the relationship between multiple numerical and categorical independent variables (i.e., project characteristics) and one numerical dependent variable (i.e., the project duration). In statistics, regression analysis is primarily used for such purposes.

Regression models were mostly used to establish the statistical relationship between key project parameters and the project duration. Some DOTs reported that their regression method was more accurate and easier to use than their previous contract time estimation methods that were primarily based upon production rates and generic precedence logic (Taylor et al., 2017). Kentucky, Indiana, and Ohio recently developed single variate and multivariate regression models that use cost estimates and selected bid item quantities to estimate contract time (Jiang & Wu, 2004; Zhai et al., 2016). Nevett et al. (2020) collected highway project data from Colorado DOT that included information about construction quantities, cost, and contract time of 15,000 projects. They analyzed 22 variables and developed a multi linear regression model that uses ten influential variables and predicts project duration. Ohio DOT (2020) developed a regression model for each project type (in total, 19 types) using eight years of project data including project cost, project type, project location, and starting season to estimate project duration. Ohio DOT uses these models to estimate the duration of a project in early preconstruction stages by determining a mean duration with 90% and 95% confidence level. The agency uses such regression tools for preliminary estimation of contract time and use production rate charts and scheduling tools for the final setting of contract time. Ohio DOT uses these models to estimate an early construction duration with 90% and 95% confidence level (Taylor et al., 2017). KYTC (2014) developed regression models for ten project types of small size and five project types of large size that receive project attributes such as work item quantities and project cost to predict project duration with a defined level of certainties. Nevett et al. (2020) analyzed 22 variables and developed a multi linear regression model that can receive ten project characteristics to predict the project duration for Colorado DOT.

Although regression analysis models have been used vastly in literature for project duration estimation, Artificial Neural Networks (ANNs) have been applied recently because of their capability of recognizing complex non-linear relationships between inputs and output (Attal 2010; Mensah et al., 2016; Petruseva et al., 2019; Cheng et al., 2019; Karaca et al., 2020; Alikhani et al., 2020). Attal (2010) identified six key project characteristics in highway projects of Virginia DOT and used them as input variables and applied ANNs to predict project duration. They achieved the accuracy of 91% for ANNs in predicting contract time. They compared the accuracy of ANNs with regression analysis, which was 91% and 89% respectively, and concluded that the ANNs had a better performance in prediction. Al-saadi et al. (2017) used ANNs for predicting contract time and achieved the accuracy of 90% and compared the method with other techniques and concluded that the ANN worked more accurate than other methods. Cheng et al. (2019) used ANN to obtain schedule to completion of construction projects and achieved 99% of accuracy. Gransberg et al. (2017) conducted a research study for the MDT and used ANNs for early cost estimation of highway projects. They proposed a top-down estimating approach and embedded the ANN model into a Microsoft Excel tool to facilitate the usability of the model. They concluded that MDT could enhance the accuracy of cost estimation using the proposed ANN model compared with their existing method.

3 Data Collection and Preliminary Analysis

This chapter discusses the characteristics of collected data and preliminary analysis results. Historical highway project bid data were obtained from MDT and analyzed to provide insights about MDT highway projects and create a database to develop project duration estimation models. This chapter provides an overview of highway project data of MDT by describing available data, major project factors, and statistics on project characteristics.

3.1 Data Overview

Historical bid data of 1,090 highway projects from 2008 to 2019 were collected from MDT. Data attributes include project numbers, location (urban/rural), bid price, bid duration, adjusted cost, charged days, work type, letting date, and bid item (work item) title and quantities.

3.1.1 Project Work Type

Highway projects are classified into different project work types based on project activities and work descriptions. Table 3.1 shows different project work types and the number of historical projects assigned to each work type. Overlays is the most common project type in MDT, followed by Reconstruction and grading, Safety, Seal & cover, and Bridge construction, rehab, and removal. The top five work types account for 76% of all highway projects in MDT.

Table 3.1 Distribution of highway projects by project work type

Row	Project work type	Frequency	Percentage of total
1	Overlays	263	24%
2	Reconstruction, grading	178	16%
3	Safety	168	15%
4	Seal & cover	129	12%
5	Bridge construction, rehab, and removal	86	8%
6	Slides or slope stabilization	36	3%
7	Signals	29	3%
8	Guardrail	25	2%
9	Microsurfacing	21	2%
10	Miscellaneous	16	1%
11	Rehab (minor grade & overlay)	17	2%
12	Crack seal	17	2%
13	Signing	13	1%
14	Drainage	11	1%

15	Portland cement concrete pavement	8	1%
16	Sidewalk	22	2%
17	Environmental and wetland	6	1%
18	Fencing	10	1%
19	Bike and pedestrian	18	2%
20	Buildings (scales, rest areas)	6	1%
21	Rumble strips	6	1%
22	Lighting	2	0%
23	Scour projects	2	0%
24	Warm mix bit surf	1	0%
	Total count	**1090**	100%

3.1.1 Work Items

The frequency of bid items or work items in the database was analyzed to remove work items with a low frequency of occurrence to avoid complexity in model development. The research team already identified most significant controlling work items in MDT projects from a recently completed research project (Jeong & Alikhani, 2020). A controlling work item may consist of one or multiple work items and is defined as major activities associated with a relatively high amount of work quantity. Controlling work items influence the duration of a project and they are highly likely to fall on the critical path of the project schedule (Jeong & Alikhani, 2020). The detailed list of all controlling work items and their associated work items is available in Jeong & Alikhani (2020). In this research, controlling items that appeared in less than 50 projects from the entire database of 1,090 projects were eliminated in order to reduce the number of variables and increase the robustness of the model. Table 3.2 shows the frequency of controlling work items in the database after removing uncommon controlling items. Common controlling work items (occurred in more than 50 projects in the past) were used as input variables.

Table 3.2 Frequency analysis of controlling work items that appeared in at least 50 projects

Row	Controlling Work Item	Frequency	% of projects
1	Mobilization	1090	100%
2	Traffic control	1086	100%
3	Remove existing structures	868	80%
4	Pavement marking	839	77%
5	Emulsified asphalt	718	66%
6	Cover	700	64%
7	Signs	697	64%
8	Temporary activities	596	55%
9	Crushed aggregate course	500	46%
10	Base preparations (soil stabilization)	460	42%

11	Guard rail	460	42%
12	Milling and pulverizing	457	42%
13	Asphalt cement	367	34%
14	Plant mix surfacing	354	32%
15	Seeding	340	31%
16	Topsoil-salvaging and placing	321	29%
17	Rumble strips	308	28%
18	Commercial mix	272	25%
19	Excavation-unclassified	272	25%
20	Farm fence	270	25%
21	Drainage pipe (<=24 in)	264	24%
22	Bridge deck	246	23%
23	Riprap	234	22%
24	Special borrow	219	20%
25	Sidewalk	181	17%
26	Curb and gutter	176	16%
27	Embankment in place	158	15%
28	Drainage pipe (> 24 in)	155	14%
29	Deck grooving (after curing)	125	12%
30	Reinforcing steel	113	10%
31	Bridge deck repair	103	9%
32	Piling	93	8%

3.1.2 Project Duration Performance

Two attributes related to project duration exist in the database: project bid days and total charged days. Project bid days indicate the estimated project time before the beginning of the construction and the total charged days reflect the actual duration of the project. Figure 3.1 shows a histogram of projects by different ranges of charged days. 75% of projects took less than 100 working days, indicating that most MDT projects take less than a year to complete.

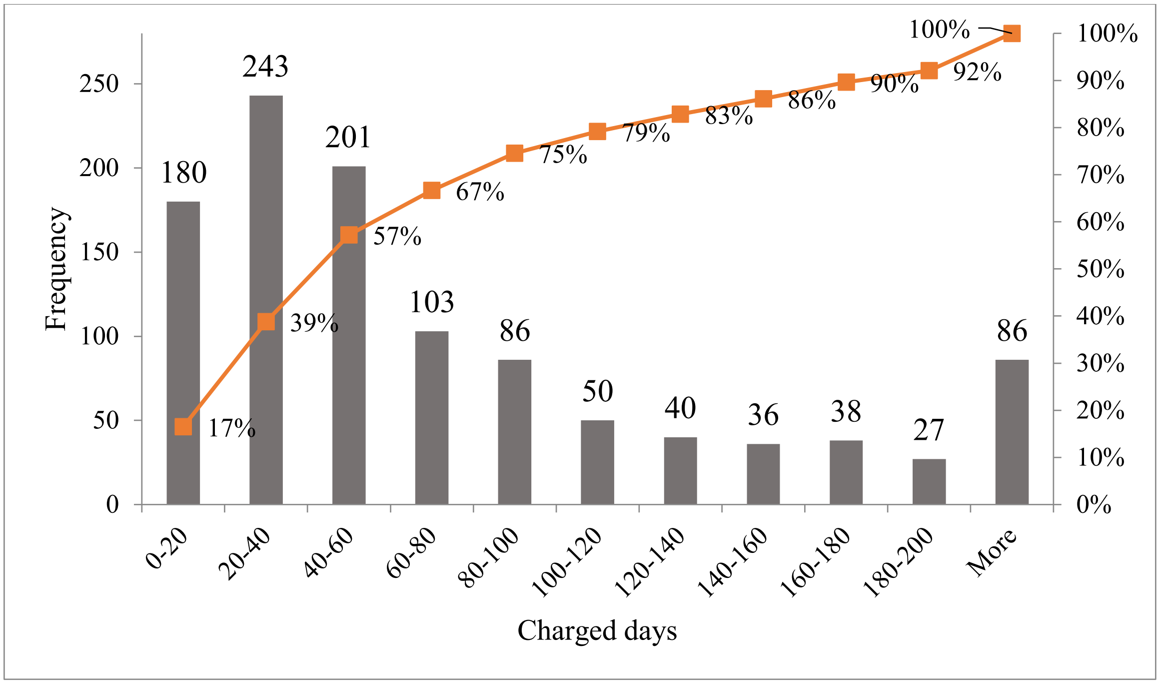


Figure 3.1 Frequency of projects with different charged days

The project bid days are compared to the project charged days to evaluate the accuracy of project time estimation before construction. Figure 3.2 illustrates the average percentage of the difference between estimated duration and actual duration for different ranges of project bid time. The percentage is calculated using Equation 1.

$$\textit{Percentage of time difference} = \frac{\textit{Charged days} - \textit{Bid days}}{\textit{Bid days}} * 100 \qquad \text{(Equation 1)}$$

For example, according to Figure 3.1, 17% of projects took less than 20 working days. Figure 3.2 shows the difference between estimated and actual project time for these projects is -23.7%, meaning such projects were finished 23.7% sooner than estimated. The percentage rises with the increase of project bid days. In projects that take 120 to 140 days, there is an average of 21.8% delay in finishing the project within estimated time. Figure 3.2 indicates that many projects have not been completed within the estimated time.

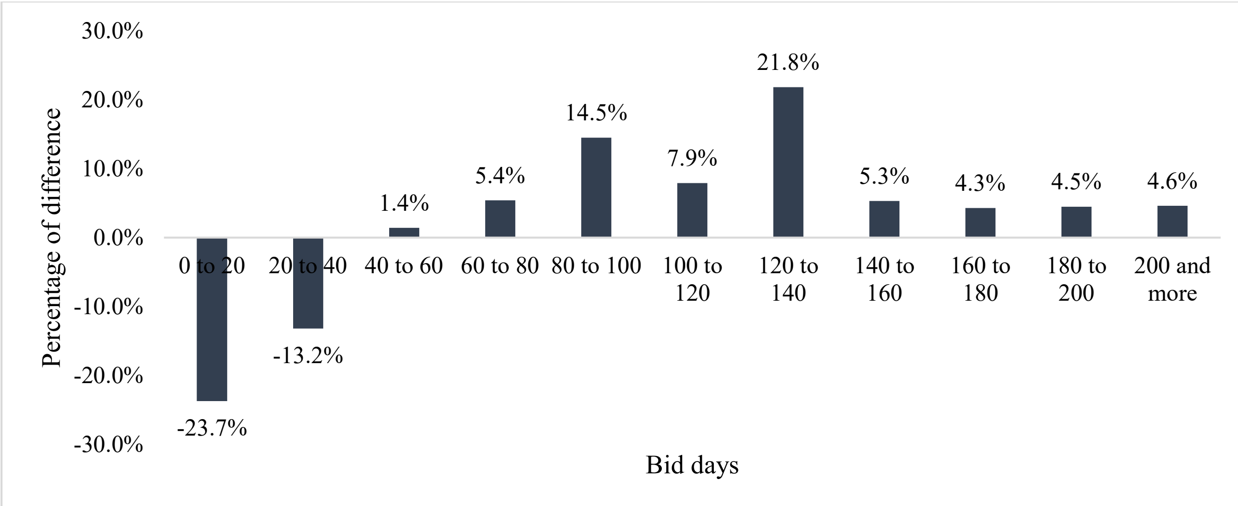


Figure 3.2 Average percentage of the difference between bid days and charged days

3.1.3 Project Cost Performance

Figure 3.3 shows a histogram of project cost in scale of million dollars. All historical project costs are adjusted to the base year of 2018 using the National Highway Construction Cost Index (NHCCI) in order to neutralize the impact of inflation and consider the time value of money. Jeong et al. (2017) identified multidimensional HCCIs for highway projects in MDT ranging from 2010 to 2014, which are used in this study. For the rest of the years (2008-2010 and 2010 to 2016), the NHCCI of FHWA was used (FHWA, 2022). Table 3.3 illustrates the Cost Indexes for different years used in this study.

Table 3.3 Cost Indexes for different years used in this study

Year	Cost Index	Resource
2008	115.06	FHWA (2020)
2009	100.25	FHWA (2020)
2010	100	Jeong et al. (2017)
2011	110.46	Jeong et al. (2017)
2012	111.12	Jeong et al. (2017)
2013	113.06	Jeong et al. (2017)
2014	115.46	Jeong et al. (2017)
2015	118.01	FHWA (2020)
2016	115.38	FHWA (2020)
2017	116.34	FHWA (2020)
2018	124.1	FHWA (2020)
2019	133.45	FHWA (2020)
2020	133.4	FHWA (2020)
2021	139.66	FHWA (2020)

Shrestha et al. (2017) divided MDT projects in terms of dollar value into three ranges of small size (up to \$3.5M), medium size (\$3.5M to \$10.5M), and large size (\$10.5M to \$50M). According to Figure 3.3, 67% of the projects took less than \$3.5M, indicating that small size projects are very common in MDT.

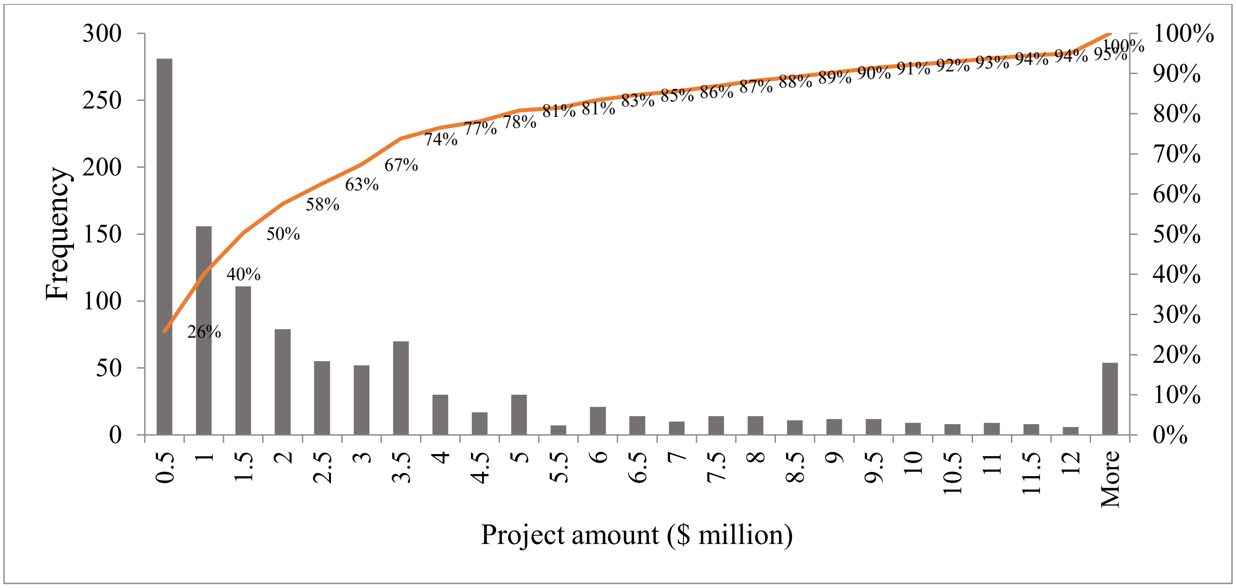


Figure 3.3 Frequency of projects in terms of cost

3.2 Preliminary Analysis on Correlations Between Variables

Correlations between project characteristics were explored to identify the importance of variables in estimating the project duration. Pearson correlation coefficient is typically used to measure the strength of a linear association between two variables. The Pearson coefficient can take a value between -1 to +1, where the value of 0 indicates no correlation and the values of 1 and -1 show the highest correlation. Positive values indicate a positive relationship, whereas negative values indicate a negative relationship. A positive association means that when the value of one variable rises, so does the value of the other. A negative relationship means when one variable decreases, another increases. Table 3.4 illustrates the Pearson correlation coefficients for project features including project location, project starting year, and project cost, project type, and controlling work items with project time. According to the table, the project cost has the highest correlation with project time followed by controlling work items. No relationship was detected between project location and project time. Correlation between project type and project time is insignificant. Also, project starting year does not influence the project time.

Table 3.4 Pearson correlation coefficient between project features and project time

Project feature	Pearson Coefficient
Location	0
Starting year	-0.035
Project cost	0.83
Controlling work items	0.72
Project type	0.07

The results of the preliminary analysis on correlations between project features and project duration are used at the model development stage to pick the most correlated variables and eliminate insignificant variables.

4 AI Model Development

This chapter describes the AI model development process for project duration estimation. To increase the model's performance, hyperparameter tuning, model regularization, and feature selection were used. The model was trained and tested by splitting the entire dataset into the training and testing datasets. A regression model was also developed for comparison with the AI model and as a companion to the AI model. The statistical metrics of the models as well as practical value of the models are described in the chapter.

4.1 Regression Analysis with Artificial Neural Networks

Artificial Neural Networks (ANNs) is an algorithm that learns the relationships between features and the target. In this research, feature variables including numeric variables (such as quantity of work item) and categorical variables are preliminarily used as input variables. The target is a numerical variable (project duration) as the output variable. To handle categorical variables, they are transformed to binary values. Figure 4.1 shows the architecture of the ANN model. In the input layer, the ANN model receives several numerical input variables and passes them through hidden layers, where weight matrices and non-linear functions are applied to the input values and added together to form a single number as the output value.

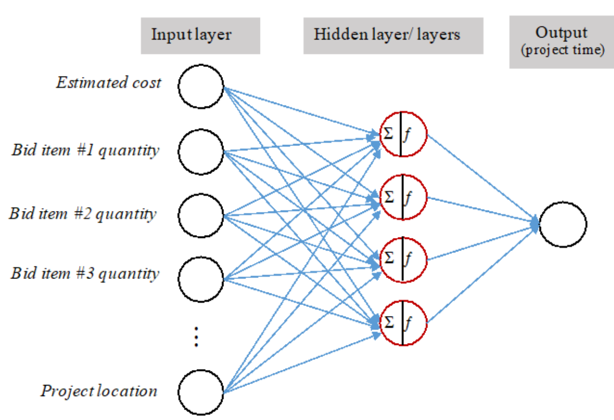


Figure 4.1 Architecture of the ANN model

4.2 Database Overview

Data attributes include six feature categories of project characteristics including five independent variables and one dependent variable. Independent variables include controlling work item quantities (total of 32 controlling work items), project work type (total of 24 types), project location (urban/rural), starting year, and construction cost. Independent variables are used to predict the dependent variable of project duration. Independent variables include numerical (such as project cost) and categorical (such as project type) variables. Categorical variables are transferred to numerical variables using one-hot encoding technique to shape the input matrix with the size of 1,090*60 (total number of projects* total number of columns). Target matrix has the size of 1,090*1 (the bid days of all 1,090 projects). Columns of the input matrix correspond to project features depicted in Figure 4.2. Initial 32 variables correspond to work item quantities, the next 24 variables account for project types, the next two variables correspond to project location, followed by starting year and estimated construction cost.

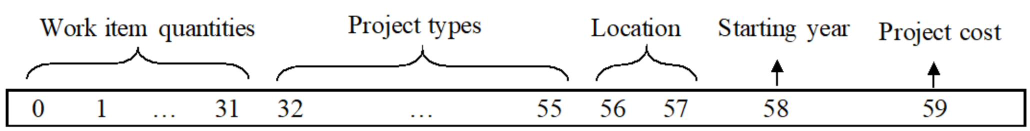


Figure 4.2 Feature arrangement in the feature matrix

4.3 Model Development, Training, and Testing

The current dataset includes values of different types with different ranges. For example, estimated construction cost is a number reflecting the estimated monetary value, while work item quantity is a number indicating the size of a particular work item. To handle the variety of values, each column is normalized separately and scaled into the range of [0,1]. The whole dataset is divided into 80% for training and 20% for testing. To test the model, two metrics of R-squared and Mean Squared Error (MSE) are used. MSE measures the average squared difference between the estimated values and the actual values. R-squared is a goodness-of-fit measure for a regression model that quantifies the proportion of variation explained by an independent variable in a regression model for a dependent variable. It is preferable to have a lower MSE and a higher R-squared. The final ANN model includes one input layer, multiple hidden layers with activation function of Sigmoid, and one output layer.

4.4 Comparison of the ANN Model with Regression Model

A linear regression model was also developed with the same dataset to compare with the ANN model. Table 4.1 summarizes the comparison results of both models in terms of two metrics. Slight differences indicate the similar performance of both ANN and Linear Regression models in predicting project duration.

Table 4.1 Comparison of linear regression and ANN model using whole features

Model	MSE	R-squared
ANN	0.0023	0.82
Linear Regression	0.0033	0.80

4.5 Feature Selection

In developing predictive models, feature selection is the process of selecting the most influential input variables. It is desired to limit the number of input variables in order to lower the computational cost of modeling and, in certain situations, increase the model's performance. To select important features, a feature score is calculated based on F-test. F-test is a statistical test which compares the importance of a model's improvement when new variables are included. Figure 4.3 illustrates the importance scores of all features. Refer to Figure 4.2 to figure out which feature corresponds to which project attribute. For example, feature # 59: estimated construction cost shows the highest importance in determining project duration.

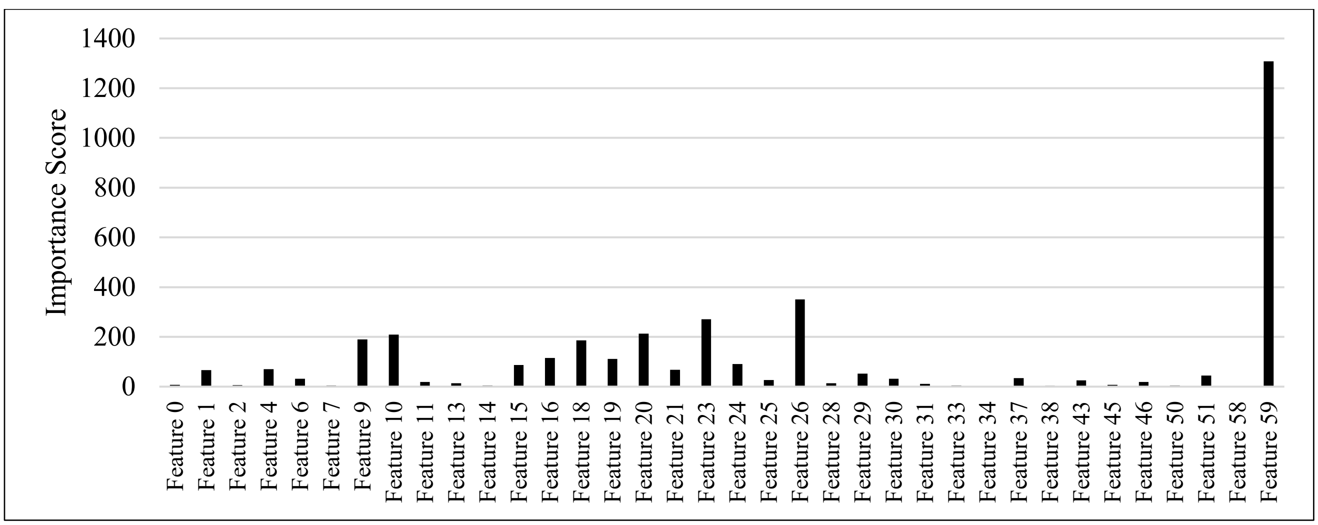


Figure 4.3 Feature importance scores

According to Figure 4.3, some features have a higher importance score than others. Feature selection includes selecting such important features and removing insignificant features. To identify the best number of features to select for the final model, different model configurations with a different number of features were developed and validated using k-fold cross validation with k=10 with the metric of MSE. Figure 4.4 shows the means of MSEs resulting from the k-fold cross validation for different numbers of features. The lowest MSE is obtained when the number of features is 26. Because of the randomness of partitioning the dataset into training and testing datasets during the cross-validation procedure, the minimum value of MSE differs somewhat from the model's MSE. For the final model, the top 26 features according to their feature score were selected.

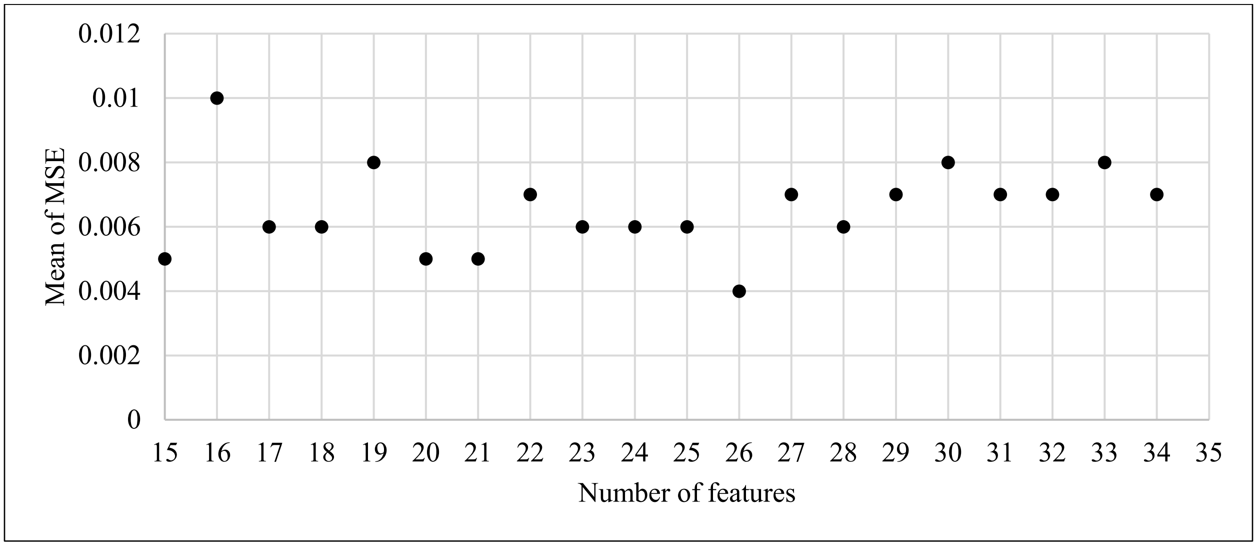


Figure 4.4 Mean of MSE in a 10-fold cross validation for different number of features

Among the top 26 features, the first and most significant feature is the estimated construction cost, and the other features include 25 controlling work items. Table 4.2 shows 25 significant controlling work items whose quantities are highly influential on project duration. Pearson coefficient was used to determine the correlation between each controlling item and the project duration. Among all work items, the “Traffic Control” work item has the highest association with project duration with the Person coefficient of 0.63. The Traffic Control work item includes the pay items of TRAFFIC CONTROL DEVICES CB with the item code of 618030005 and 999618000, TRAFFIC CONTROL with the item code of 999618030 and 999618035), and TRAFFIC CONTROL-FIXED with the item code of 618030015 (MDT, 2021). In the database, the pay item of TRAFFIC CONTROL DEVICES CB accounts for 90% of traffic control related pay items, that is the number of devices used for maintaining the traffic during construction (MDT, 2020).

Table 4.2 Significant controlling work items with highest impact on project duration

Row	Controlling Work Items	Correlation with time (Pearson Coefficient)
1	Traffic control	0.63
2	Drainage pipe (<= 24 in)	0.53
3	Crushed aggregate course	0.52
4	Excavation-unclassified	0.51
5	Mobilization	0.50
6	Seeding	0.50
7	Special borrow	0.44
8	Reinforcing steel	0.39
9	Plant mix surfacing	0.38
10	Asphalt cement	0.38
11	Drainage pipe (> 24 in)	0.36
12	Deck grooving (after curing)	0.36
13	Riprap	0.35
14	Piling	0.27
15	Sidewalk	0.24
16	Curb and gutter	0.24
17	Guard rail	0.19
18	Bridge deck	0.19
19	Commercial mix	0.17
20	Signs	0.16
21	Embankment in place	0.14
22	Emulsified asphalt	0.11
23	Cover	0.11
24	Rumble strips	0.10
25	Milling and pulverizing	0.10

The results of feature selection are compatible with correlation analysis in the last chapter. Estimated construction cost was discovered as the most significant variable in both correlation analysis and feature selection analysis, while project location and starting year were identified as insignificant in influecint project duration.

4.6 Final Results

After feature selection and correlation analysis, the number of features reduced from 60 to 26. The ANN and linear regression models were developed based upon the selected features. Table 4.3 summarizes the final MSE and R-squared of the models. In terms of project duration estimation, the result demonstrates that the ANN outperforms Regression by a little margin. Because the differences between the two models are minor, both will be incorporated in the Microsoft Excel tool, which will allow users to consider the estimates of both models.

Table 4.3 Comparison of linear regression and ANN model using selected features

Model	MSE	R-squared
ANN	0.0022	0.72
Linear Regression	0.0034	0.75

5. Tool Development

This chapter describes the development process of a Microsoft Excel based Tool for implementation and provides a comprehensive guideline on using the tool. The tool helps MDT engineers to enter the quantities of major activities and the estimated construction cost to receive an early estimation of project duration using the two models developed in this research. Real examples are used to show how the tool works and how to obtain and interpret the results.

5.1 Tool Development Process

The ANN and the regression models developed in this research were transferred and embedded into a Microsoft Excel tool. Figure 5.1 represents the equation of the linear regression model. The vector of input variables incorporates the quantities of 25 major work items plus the estimated construction cost (vector size of 1*26). The unit impact of input variables is then eliminated by normalizing the input vector, converting values to numbers between 0 and 1 using Equation 5.1.

$$\text{Normalized} = \frac{(X - X_{\min})}{(X_{max} - X_{min})} \qquad \text{Equation 5.1}$$

The normalized input vector is then multiplied by a matrix of coefficients (size of 26*1) and added to an intercept value to yield the output, which is the estimated project duration in this research. The coefficient values and the intercept value obtained from the regression model are transferred to an Microsoft Excel sheet in the tool (a hidden sheet) to interact with the input variables in the background to calculate the estimated project duration as the output.

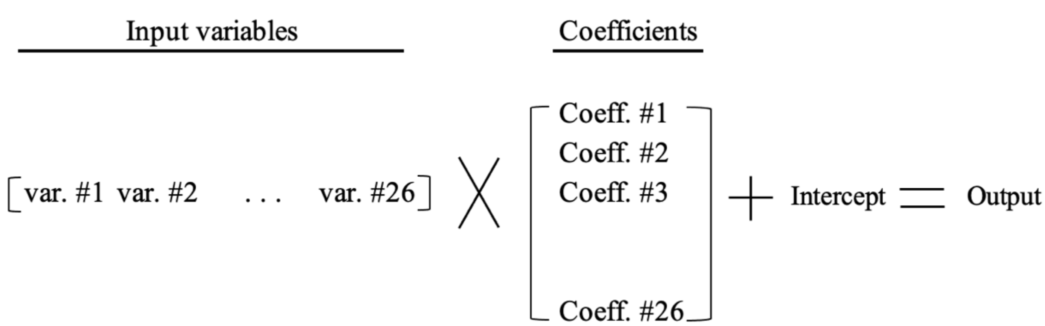


Figure 5.1 Linear regression equation presentation

The ANN model includes an input layer (size of 1*26), three hidden layers with activation functions and an output layer (size of 1*1). The hidden layers include a weight matrix (similar to coefficients matrix in the regression model) and a bias matrix (similar to intercepts matrix in the regression model) and an activation function. The following describes how to obtain the network's output, which represents the estimated project duration.

Output of layer 1: *Relu* [normalized_input_vector $_{(1*26)}$ * $W^1_{(26*500)}$ + $B^1_{(1*500)}$]

Output of layer 2: *Relu* [layer1_output $_{(1*500)}$ * $W^2_{(500*100)}$ + $B^2_{(1*100)}$]

Output of layer 3: *Sigmoid* [layer2_output $_{(1*100)}$ * $W^3_{(100*50)}$ + $B^3_{(1*50)}$]

Output of network: layer3_output $_{(1*50)}$ * $W^4_{(50*1)}$ + $B^4_{(1*1)}$

Where W is the weights and B is the bias matrix. The subscripts in parentheses indicate the size of matrixes and superscripts indicate the hidden layer numbers. Relu is a function that outputs the input directly if it is positive, otherwise, it outputs zero. Sigmoid function is obtained from Equation 5.2, which yields numbers between 0 and 1.

$$\text{Sigmoid} = \frac{1}{1+e^{-x}} \qquad \text{Equation 5.2}$$

The weights and bias matrixes as well as the activation functions are transferred to the Microsoft Excel tool (in hidden sheets) to execute interactions in the background and compute the network's output automatically given the input vector.

5.2 Tool Overview

The MS Excel Tool was named AI-PDET (Artificial Intelligence based Project Duration Estimation Tool). As Figure 5.2 shows, the AI-PDET includes three sheets; I) Home page, which provides a general guideline on using the tool. Figure 1.1 shows a screenshot of the home page. Clicking the “Start Analysis” button opens up the next worksheet. II) Project duration estimation work sheet, that takes input variables and predicts the duration of the project with two different methods of the ANN model and the regression model. III) sample projects sheet, which includes the information of three real cases to provide examples of using the tool.

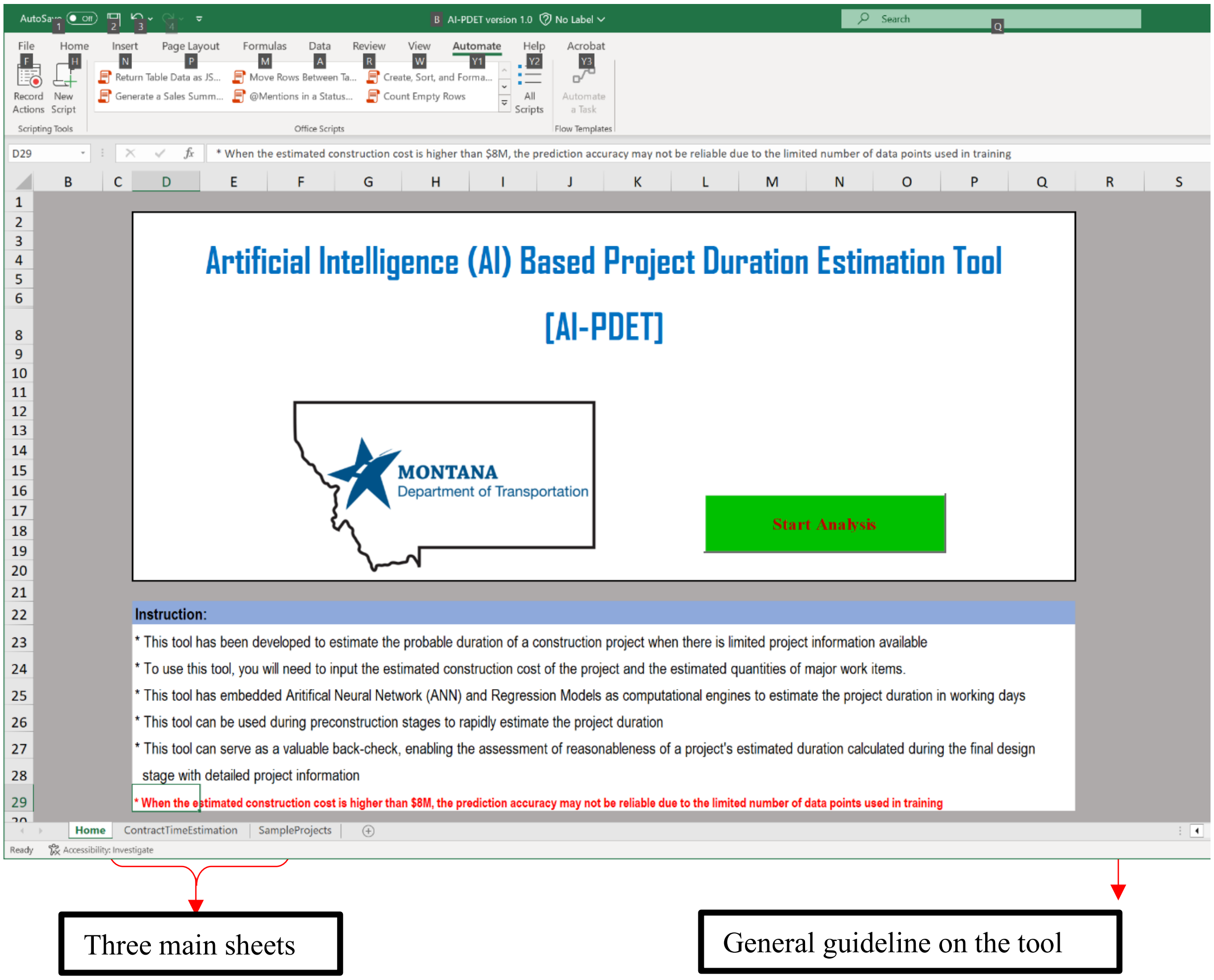


Figure 5.2 Screenshot of AI-PDET home page

5.3 Project Duration Estimation

Figure 5.3 illustrates the second page of the tool in which a user enters input variables in the input section and then, the output section shows the two estimated project durations predicted from ANN and regression models. The input section takes the estimated quantity of each of the 25 major work items and the estimated construction cost of the project in the yellow cells. Each major work item may include one or multiple pay items. To learn more about pay items and their associated major work items, refer to Appendix A of Jeong & Alikhani (2020). After the input section is completed, AI-PDET automatically presents the predicted project durations in the output section.

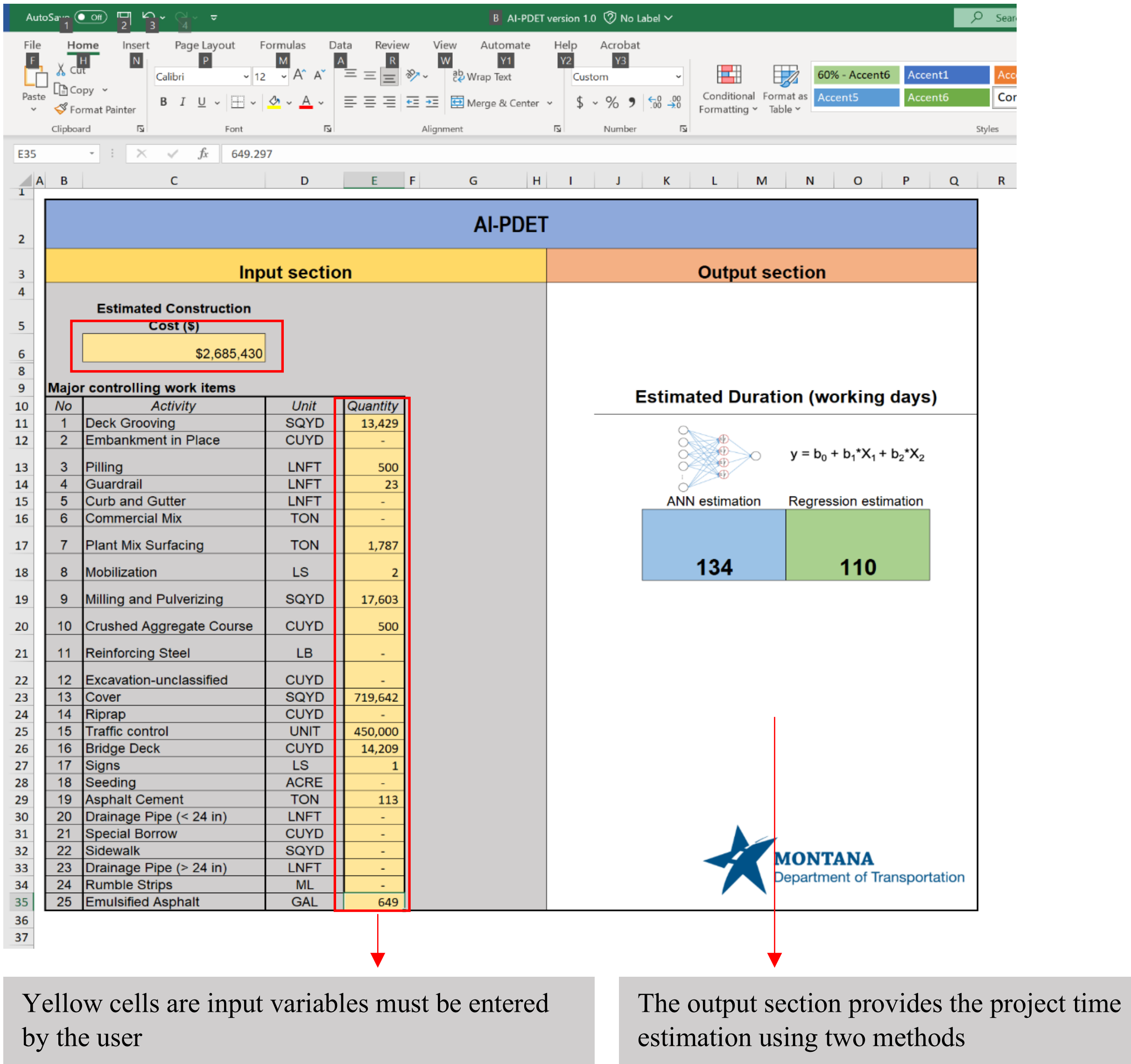


Figure 5.3 Screenshot of the AI-PDET Input and Output Page

5.4 Sample projects

The third page of the AI-PDET includes three cases of real highway projects in MDT (Figure 5.4). The first project is categorized as "Miscellaneous," the second as "Bike and pedestrian," and the third is a "Bridge construction" work type. All projects started in 2019. The actual charged days of each project are provided in the sheet to help the user to compare it with the predicted project durations with the AI-PDET. The quantities of major work items with their proper units are provided for each project. When the user takes the provided input values and enter them in the second sheet of the AI-PDET, the user can see the predicted project durations on the right hand side and compare them with the real duration of the project.

Project characteristics	*Case1*	*Case2*	*Case3*
Project number	9617133000	9149077000	8085164000
Project ID	IM 90-9(133)528	TA 41(77)	
Project type	MISCELLANEOUS	BIKE AND PEDESTRIAN	BRIDGE CONSTRUCTION,RE
Project location	Bighorn County	Ravalli County	Cascade County
Project begin year	2019	2019	2019
Engineers' estimate ($)	$ 2,682,663	$ 519,287	$ 5,980,982
Project time (working days)	75	42	162
Resource	MDT website	MDT website	MDT website
Activity quantities			
Deck Grooving	0	0	0
Embankment in Place	0	0	31.56
Pilling	0	0	0
Guardrail	5	0	0
Curb and Gutter	0	0	0
Commercial Mix	398	0	1006.23
Plant Mix Surfacing	0	0	0
Mobilization	1	1	1
Milling and Pulverizing	0	0	6322.2
Crushed Aggregate Course	540.15	30.5	0
Reinforcing Steel	0	0	30444
Excavation-unclassified	0	507.9	0
Cover	5188.7	0	7441
Riprap	0	0	0
Traffic control	19916	1	295704
Bridge Deck	0	0	0
Signs	0	0	360
Seeding	2.9	0	0
Asphalt Cement	0	0	0
Drainage Pipe (< 24 in)	0	0	0
Special Borrow	0	0	0
Sidewalk	0	1231	0
Drainage Pipe (> 24 in)	0	0	0
Rumble Strips	0	0	0
Emulsified Asphalt	10.44	0	0

Home | ContractTimeEstimation | SampleProjects | +

Figure 5.3 Screenshot of the sample projects sheet of AI-PDET

5.5 Results Interpretation and Model Limitations

The AI-PDET is a top-down tool that can estimate a project's duration when a limited amount of project information is available during the preconstruction stages. The estimated project duration using this tool is helpful in the early stages of the project delivery process for project programming and budgeting purposes. It can also be used to check the reasonableness of the project duration estimate derived from detailed project scheduling activities, once more detailed information on activity quantities, production rates, and activity sequencing become available in later design stages.

6. Maintenance of AI-PDET and Database Update

This research used a data-driven approach to leverage historical data to predict the project duration of a new construction project in the early construction phases. As time goes on, new projects are carried out, containing new experiences. The data from recent projects can be added to the older ones to enrich the database. Data-driven prediction models can be trained on the updated database to obtain new knowledge from recent projects to enhance predictions in the future. Two data-driven models developed in this research include the ANN and the regression model trained on data from 2008 to 2019.

This chapter explains how new project data can be used to update the models and the AI-PDET to make the tool more relevant to recent projects. In this research, coding was conducted in the Google Colab environment using Python language to create, train, and evaluate models. This chapter explains different coding parts and actions that need to be taken to transfer models' attributes to AI-PDET. The input of the code includes the excel dataset of projects, and the code's output is the attributes of the models, such as weights, biases, coefficients, and intercepts. The models' attributes need to be transferred to the AI-PDET to finalize the updating process.

6.1 Updating Process for AI-PDET

There are three main steps to update AI-PDET (Figure 6.1). Step 1 involves updating the project's historical data and, if needed, updating the activity dictionary and the cost indexes. Step 2 includes running the developed code in this research that receives updated files of Step 1 to train the ANN and the regression models and output models' attributes. In Step 3, the models' attributes are transferred into the AI-PDET.

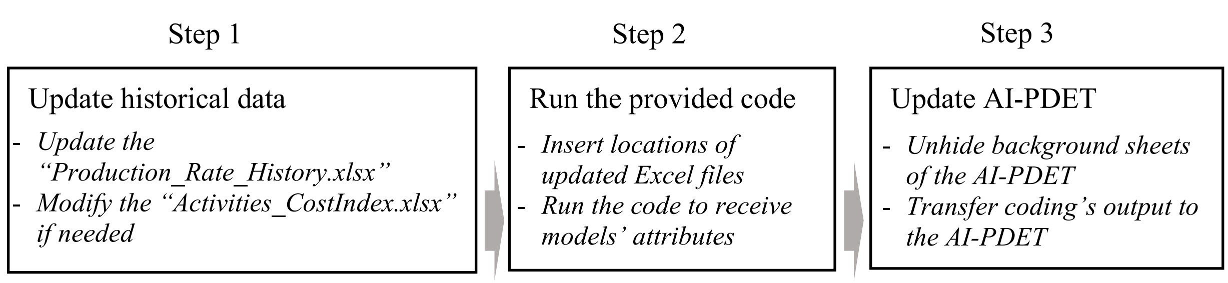
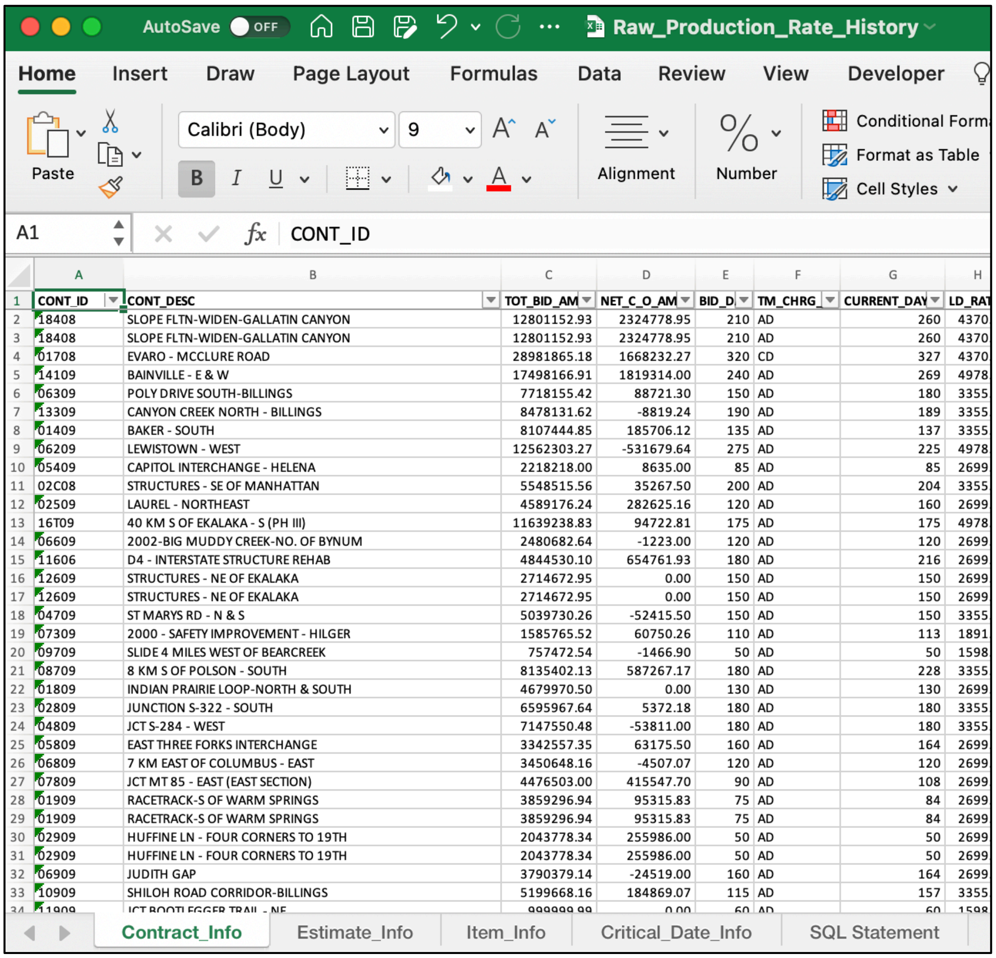


Figure 6.1 Process of updating the AI-PDET

6.2 Step 1: Update Historical Data

Two Excel files must be updated as inputs of the coding section. The first Excel file is the dataset of past projects. The original dataset obtained from MDT in this research is named "Production_Rate_History.xlsx", which is assumed as the main historical dataset (Figure 6.2). For the purpose of this research, it is required to update and use two sheets in this file called "Contract_info" and "Item_Info". Attributes used from the "Contract_info" include project number, project cost, adjusted charge days, and project year. The "Item_Info" sheet is used to obtain the information of work item quantities.



	A	B	C	D	E	F	G	H
1	CONT_ID	CONT_DESC	TOT_BID_AM	NET_C_O_AM	BID_D	TM_CHRG_	CURRENT_DAY	LD_RAT
2	18408	SLOPE FLTN-WIDEN-GALLATIN CANYON	12801152.93	2324778.95	210	AD	260	4370
3	18408	SLOPE FLTN-WIDEN-GALLATIN CANYON	12801152.93	2324778.95	210	AD	260	4370
4	01708	EVARO - MCCLURE ROAD	28981865.18	1668232.27	320	CD	327	4370
5	14109	BAINVILLE - E & W	17498166.91	1819314.00	240	AD	269	4978
6	06309	POLY DRIVE SOUTH-BILLINGS	7718155.42	88721.30	150	AD	180	3355
7	13309	CANYON CREEK NORTH - BILLINGS	8478131.62	-8819.24	190	AD	189	3355
8	01409	BAKER - SOUTH	8107444.85	185706.12	135	AD	137	3355
9	06209	LEWISTOWN - WEST	12562303.27	-531679.64	275	AD	225	4978
10	05409	CAPITOL INTERCHANGE - HELENA	2218218.00	8635.00	85	AD	85	2699
11	02C08	STRUCTURES - SE OF MANHATTAN	5548515.56	35267.50	200	AD	204	3355
12	02509	LAUREL - NORTHEAST	4589176.24	282625.16	120	AD	160	2699
13	16T09	40 KM S OF EKALAKA - S (PH III)	11639238.83	94722.81	175	AD	175	4978
14	06609	2002-BIG MUDDY CREEK-NO. OF BYNUM	2480682.64	-1223.00	120	AD	120	2699
15	11606	D4 - INTERSTATE STRUCTURE REHAB	4844530.10	654761.93	180	AD	216	2699
16	12609	STRUCTURES - NE OF EKALAKA	2714672.95	0.00	150	AD	150	2699
17	12609	STRUCTURES - NE OF EKALAKA	2714672.95	0.00	150	AD	150	2699
18	04709	ST MARYS RD - N & S	5039730.26	-52415.50	150	AD	150	3355
19	07309	2000 - SAFETY IMPROVEMENT - HILGER	1585765.52	60750.26	110	AD	113	1891
20	09709	SLIDE 4 MILES WEST OF BEARCREEK	757472.54	-1466.90	50	AD	50	1598
21	08709	8 KM S OF POLSON - SOUTH	8135402.13	587267.17	180	AD	228	3355
22	01809	INDIAN PRAIRIE LOOP-NORTH & SOUTH	4679970.50	0.00	130	AD	130	2699
23	02809	JUNCTION S-322 - SOUTH	6595967.64	5372.18	180	AD	180	3355
24	04809	JCT S-284 - WEST	7147550.48	-53811.00	180	AD	180	3355
25	05809	EAST THREE FORKS INTERCHANGE	3342557.35	63175.50	160	AD	164	2699
26	06809	7 KM EAST OF COLUMBUS - EAST	3450648.16	-4507.07	120	AD	120	2699
27	07809	JCT MT 85 - EAST (EAST SECTION)	4476503.00	415547.70	90	AD	108	2699
28	01909	RACETRACK-S OF WARM SPRINGS	3859296.94	95315.83	75	AD	84	2699
29	01909	RACETRACK-S OF WARM SPRINGS	3859296.94	95315.83	75	AD	84	2699
30	02909	HUFFINE LN - FOUR CORNERS TO 19TH	2043778.34	255986.00	50	AD	50	2699
31	02909	HUFFINE LN - FOUR CORNERS TO 19TH	2043778.34	255986.00	50	AD	50	2699
32	06909	JUDITH GAP	3790379.14	-24519.00	160	AD	164	2699
33	10909	SHILOH ROAD CORRIDOR-BILLINGS	5199668.16	184869.07	115	AD	157	3355
34	11909	JCT BOOTLEGGER TRAIL - NE	999999.99	0.00	60	AD	60	1598

Figure 6.2 Screenshot of the "Production_Rate_History.xlsx" Excel file

The second imported Excel file called "Activities_CostIndex.xlsx", developed in this research, including the list of work items and their associated pay items and cost indexes. Figure 6.3 illustrates a screenshot of this Excel file that includes two sheets. Two columns on the "Controlling" page list the "pay item description" and the related "controlling activities". Refer to the project's final report for further information on how pay items are merged and tied to activities.

This page can be modified by placing the name of the controlling activity in front of the pay items if more pay items need to be linked to controlling activities. The year and the year's cost index are listed on the "CostIdex" sheet. In this study, the ratio of indexes of different years is employed to adjust project cost to the base year of 2021 as the current year of research. For future uses, there is a need to add the cost indexes for upcoming years to this file. For more information about the cost indexes used in this research, refer to the final report. Both Excel files can get updated over time as more information on new projects and cost indexes becomes available.

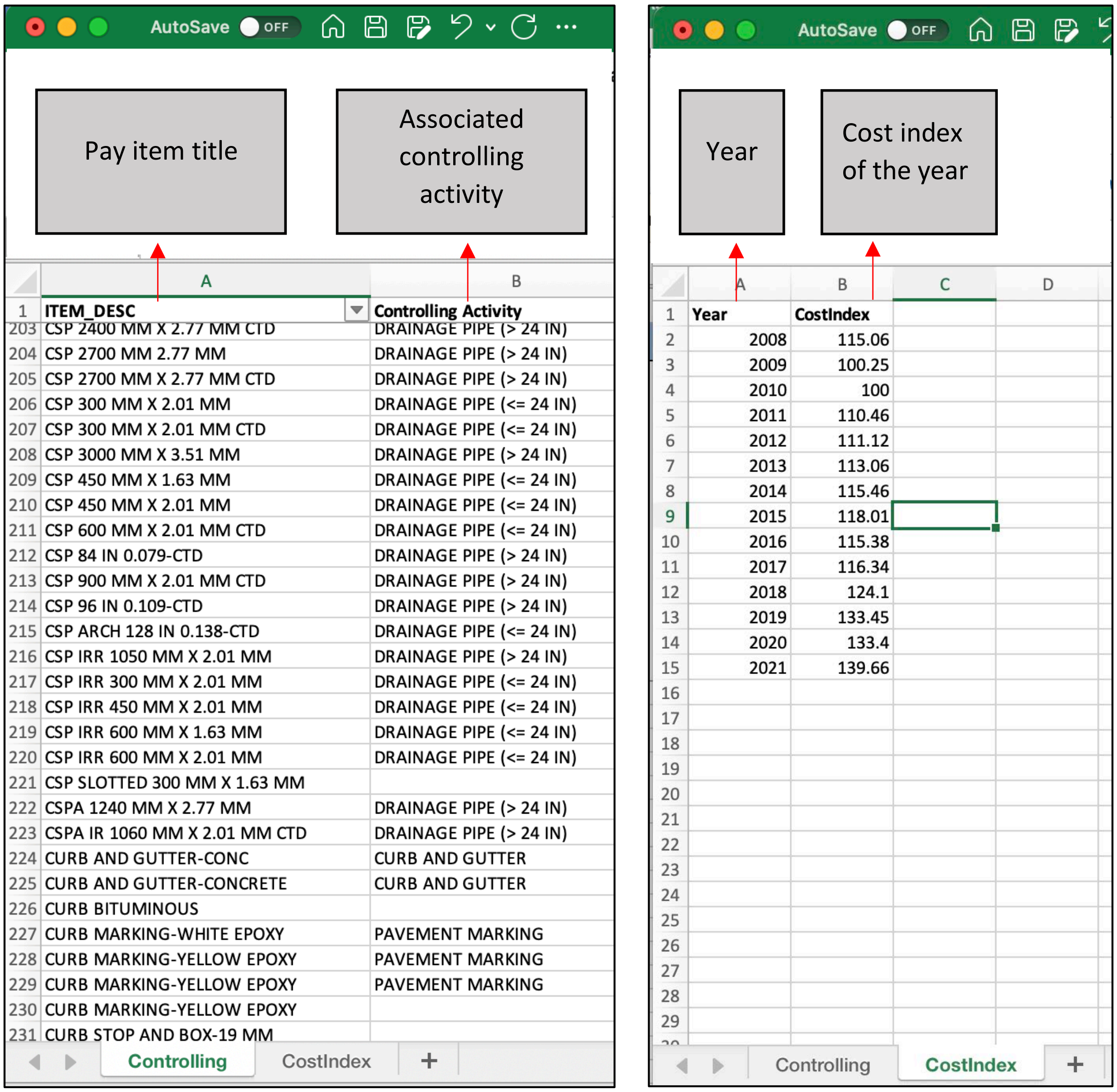


Figure 6.3 Screenshot of two sheets in the “Activities_CostIndex.xlsx”

6.3 Step 2: Run the Provided Code

The developed code in this research takes the MS Excel files of the last step to train and evaluate the ANN and regression models and produce the models’ attributes required to update the AI-

PDET. The code consists of six parts, five of which use input data to train and test models, while the sixth part outputs the properties of the models in MS Excel format. The following explains each section of the code and discusses how to update the AI-PDET using the coding outputs.

6.3.1 Part 1: Import Required Python Libraries and Read Raw Data

Figure 6.4 shows the first part of the code. Required python libraries are imported to use the features of pre-defined functions to facilitate the computation, model development, and evaluation. Two updated excel files from the last step are imported. The location of the files has to be changed in order to allow the code to read the files from the local computer or a virtual drive.

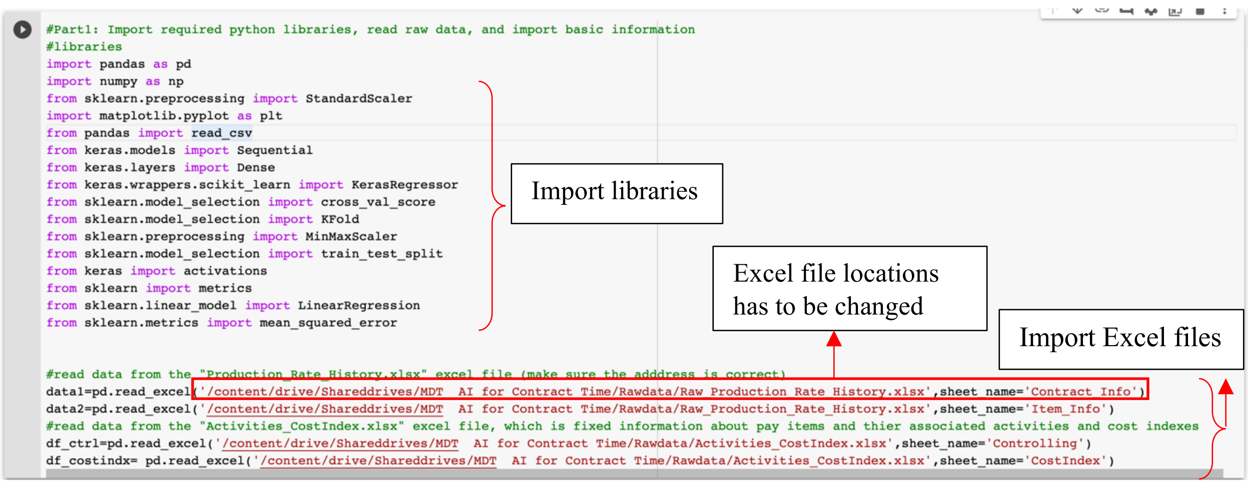


Figure 6.4 Screenshot of Part 1 of the code

6.3.2 Part 2: Define Activities, Project Numbers, Project Quantities, Project Costs, and Project Times

Part 2 of the code obtains information from the imported files (Figure 6.5). First, a list of activities is defined, including the name of 25 significant activities and their associated pay items obtained through the feature selection process in Task 3. Then, project information is obtained from the imported Excel files, including:

- Project number: extracted from "PRJ_NBR" column of the "Contract_Info" sheet from the "Production_Rate_History.xlsx" Excel file.
- Project quantities: for each project, the quantity of each pay item is available in "Item_Info" sheet of the "Production_Rate_History.xlsx" Excel file. The pay item quantities are used to compute the quantity of the 25 important activities for each project.
- Project year: extracted from "WORK_BEG" column of the "Contract_Info" sheet.

- Project cost: the cost of each project is the summation of "TOT_BID_AMT" and "NET_C_O_AMT" of each project available in the "Contract_Info" sheet. The costs are then adjusted using the cost indexes imported from the "Activities_CostIndex.xlsx" Excel file.
- Project time: if the "ADJUST_CHARGE_DAYS" column of the "Contract_Info" sheet has a value, it is considered as the project time. Otherwise, the "CHARGE_DAYS" column is used as equal to project time.

```
#Part 2: define activities, project numbers, project quantities, project costs, and project times
#define list of activities
ActivitiesNew=['DECK GROOVING (after curing)','EMBANKMENT IN PLACE','PILING','GUARD RAIL','CURB AND GUTTER','COMMERCIAL MIX',
            'PLANT MIX SURFACING','MOBILIZATION','MILLING AND PULVERIZING','CRUSHED AGGREGATE COURSE','REINFORCING STEEL',
            'EXCAVATION-UNCLASSIFIED','COVER','RIPRAP','TRAFFIC CONTROL','BRIDGE DECK','SIGNS','SEEDING','ASPHALT CEMENT',
            'DRAINAGE PIPE (<= 24 IN)','SPECIAL BORROW','SIDEWALK','DRAINAGE PIPE (> 24 IN)','RUMBLE STRIPS','EMULSIFIED ASPHALT']

#define a dictionary of pay items and their activities
controlling=list(set(df_ctrl['Controlling']))
payitem_actvty={}
for i in range(len(df_ctrl['ITEM_DESC'])):
  if df_ctrl.iloc[i,1] in ActivitiesNew:
    payitem_actvty[df_ctrl.iloc[i,0]]=[df_ctrl.iloc[i,1]]

#define project number list
proj_nbr=list(set(data1['PRJ_NBR']))
print('Number of projects:',len(proj_nbr) )

#find the quantities of each item for each project
projquant={}

for j in proj_nbr:
        projquant[j] = [0] * len(ActivitiesNew)
for i in range(len(data2['PRJ_NBR'])):
  if data2['ITEM_DESC'][i] in payitem_actvty.keys():

  #update values of quantities for each project
    projquant[data2['PRJ_NBR'][i]][ActivitiesNew.index(payitem_actvty[data2['ITEM_DESC'][i]][0])] += data2['QTY_PD_TO_DT'][i]

#find project starting year
proj_year={}
for i in range(len(data1['PRJ_NBR'])):
  proj_year[data1['PRJ_NBR'][i]]=str(data1['WORK_BEG'][i])[0:4]

#find project costs and adjust them to the base year of 2021
#project cost is the "total bid amount" plus "net-c-o-amt" amount
proj_cost={}
for i in range(len(data1['PRJ_NBR'])):
  proj_cost[data1['PRJ_NBR'][i]]=data1['TOT_BID_AMT'][i]+data1['NET_C_O_AMT'][i]
proj_cost_adj={}
for i in proj_cost.keys():
  if int(proj_year[i]) in list(df_costindx['Year']):
    proj_cost_adj[i]=proj_cost[i]*int((df_costindx.loc[df_costindx.Year==2021,'CostIndex']))/int((df_costindx.loc[df_costindx.Year==i


#find project time that is 'adjusted charge days' if available, otherwise is 'charge days
proj_time={}
for i in range(len(data1['PRJ_NBR'])):
  if np.isnan(data1['ADJUST_CHARGE_DAYS'][i]):
    proj_time[data1['PRJ_NBR'][i]]=data1['CHARGE_DAYS'][i]
  else:
    proj_time[data1['PRJ_NBR'][i]]=data1['ADJUST_CHARGE_DAYS'][i]
```

Extract project information from the imported Excel files

Define list of 25 important activities

Figure 6.5 Screenshot of Part 2 of the code

6.3.3 Part 3: Create the Dataset, Normalize Data, and Split the Dataset into Train and Test

Part 3 of the code stacks project quantities and project adjusted costs to be used as the input matrix (X) and saves the project durations in a vector as the output (y). The input and output are then normalized to have values between 0 and 1 (X_normalized and y_normalized). Then, normalized values are split to training (80%) and testing (20%) datasets (Figure 6).

```python
#Part 3: create the dataset, normalize data, and split dataset into train and test
#create the dataset
#define flattening function
def flatten(t):
    return [item for sublist in t for item in sublist]

#create the dateset
X=[]
for i in range(len(proj_nbr)):
  x1=[]
  x1.append(projquant[proj_nbr[i]])
  x1.append([proj_cost_adj[proj_nbr[i]]])
  X.append(flatten(x1))

y=[]
for i in range(len(proj_nbr)):
  y1=[]
  y1.append(proj_time[proj_nbr[i]])
  y.append(y1)

##normalization
#normalize X
scaler = MinMaxScaler()
scaler.fit(X)
X_normalized = scaler.transform(X)
#normalize y
scaler = MinMaxScaler()
scaler.fit(y)
y_normalized=scaler.transform(y)

#split the dataset into 80% train and 20% test
X_train, X_test, y_train, y_test = train_test_split(X_normalized, y_normalized, test_size = 0.2, random_state = 0)
y_train=np.array(y_train)
y_test=np.array(y_test)
```

Create the dataset

Normalization

Split into train and test

Figure 6.6 Screenshot of Part 3 of the code

6.3.4 Part 4: Develop a new ANN Model

Part 4 of the code defines the ANN layers, activation functions, training epochs, and trains the model using the training dataset, then computes the model's MSE on the test dataset (Figure 6.7). While the number of layers and activation functions cannot be modified to keep the AI-PDET consistent, some hyperparameters can be changed, such as the number of training epochs and the method of optimization.

```
#Part 4: develop the ANN model and evaluate it
#define the ANN layers, hyperparameters, and training
model = Sequential()
model.add(Dense(500, input_dim=len(X_normalized[0]), activation= "relu"))
model.add(Dense(100, activation= "relu"))
model.add(Dense(50, activation= "sigmoid"))
model.add(Dense(1))
model.compile(loss= "mean_squared_error" , optimizer="adam", metrics=["mean_squared_error"])
model.fit(X_train, y_train, epochs=200)

#evaluate the model
pred = model.predict(X_test)

#measure MSE error.
score = metrics.mean_squared_error(pred,y_test)
print("Final score (MSE): {}".format(score))
```

ANN model development

Model evaluation

Figure 6.7 Screenshot of Part 4 of the code

6.3.5 Part 5: Develop a new Regression Model

Part 5 of the code executes the regression model and exports the coefficients and the intercept. The evaluation measurements, such as the regression score and the MSE are reported (Figure 6.8).

```
#Part 5: develop the regression model and evaluate it
regr = LinearRegression()
regr.fit(X_train, y_train)
print('regression score is:',regr.score(X_test, y_test))

y_pred=regr.predict(X_test)

print('regression MSE is',mean_squared_error(y_test, y_pred, multioutput='raw_values'))
```

```
regression score is: 0.7152690126304917
regression MSE is [0.00398994]
regression bias is: [0.05326561]
```

Figure 6.8 Screenshot of Part 5 of the code

6.4 Step 3: update the AI-PDET

In step three, the attributes of the models are extracted into Excel files using the code, and the files are then transferred into the AI-PDET.

6.4.1 Part 6: Extract the Attributes of the new ANN and Regression Models

Part 6 of the code extracts both the ANN and the regression attributes and saves them into Excel files (Figure 6.9). A matrix of weights and biases and an activation function are available for each layer of the ANN model. Running this part of the code will automatically save weights matrices (W_0 to W_3) and biases matrices (B_0 to B_1). Additionally, it produces the minimum and maximum values of the input and output variables that are necessary for the normalization process. The coefficients of the regression model are exported, and the regression intercept is reported.

```
#Part 6: extract the ANN's model weights and biases into excel files
c=[]
w=[]
for layer in model.layers:
  c.append(layer.get_config())
  w.append(layer.get_weights())
#extract weights and biases
pd.DataFrame(w[0][0]).to_excel('w0.xlsx')
pd.DataFrame(w[0][1]).to_excel('b0.xlsx')
pd.DataFrame(w[1][0]).to_excel('w1.xlsx')
pd.DataFrame(w[1][1]).to_excel('b1.xlsx')
pd.DataFrame(w[2][0]).to_excel('w2.xlsx')
pd.DataFrame(w[2][1]).to_excel('b2.xlsx')
pd.DataFrame(w[3][0]).to_excel('w3.xlsx')
pd.DataFrame(w[3][1]).to_excel('b3.xlsx')
#exctract min and max to reverse normalization
min1=np.min(X,axis=0)
max1=np.max(X,axis=0)
pd.DataFrame(min1).to_excel('minx.xlsx')
pd.DataFrame(max1).to_excel('maxx.xlsx')
print('minimum of y:', np.min(y))
print('maximum of y:', np.max(y))
#extract coefficients and intercepts of the regression into excel files
pd.DataFrame(regr.coef_).to_excel('regressioncoeff.xlsx')
print('regression bias is:',regr.intercept_)
```

ANN model 's attributes

Normalization vectors

Regression attributes

```
minimum of y: 1.0
maximum of y: 549.0
```

```
regression bias is: [0.05326561]
```

Figure 6.9 Screenshot of Part 6 of the Code

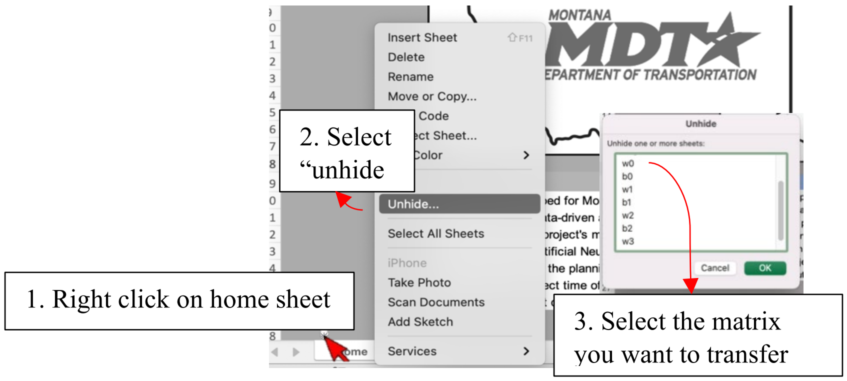
6.5 Transfer Matrices to AI-PDET

The AI-PDET includes hidden sheets with the same name as the models' attribute matrices. Three steps are needed to move the matrices to the proper location in the AI-PDET.

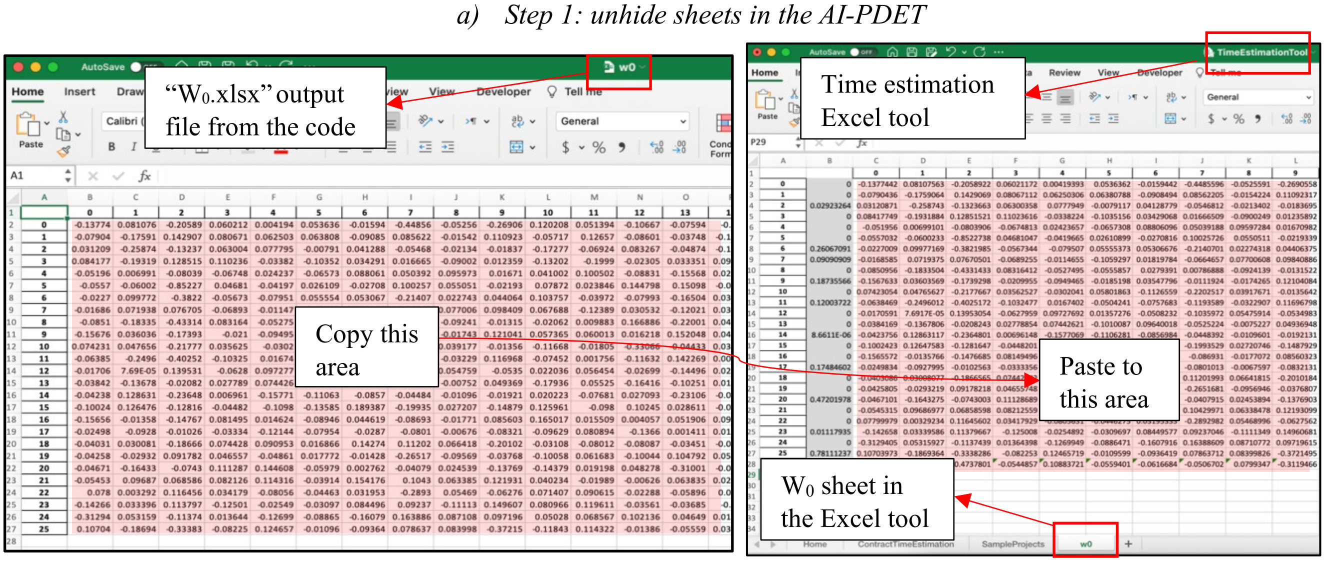
I) Step 1: unhide sheets in the AI-PDET; To unhide a sheet, right click on the home page, then select "unhide", then select the matrix you want to transfer (Figure 6.10.a).

II) Step 2: transfer the ANN models' weights and biases from the code's output files to the AI-PDET (Figure 6.10.b).

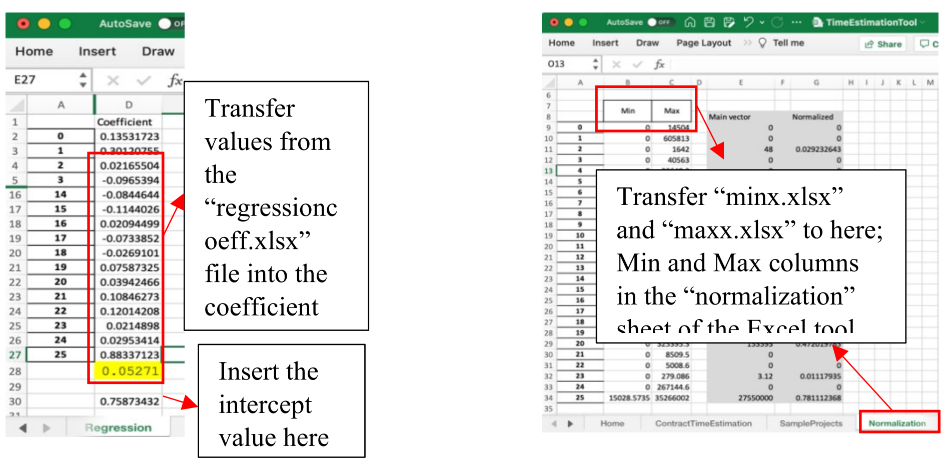
III) Step 3: transfer the regression model's attributes and the normalization values to the AI-PDET (Figure 10.c).



a) Step 1: unhide sheets in the AI-PDET



b) Step 2: transfer ANN models' attributes from the code's output files into AI-PDET



c) Step 3: transfer the regression attributes and the normalization values to the AI-PDET

Figure 6.10 Steps to transfer values from the code to the AI-PDET

6.6 Summary

This chapter explained how to update the AI-PDET using new project data. The coding was fully explained to make users understand how the code works and how to use it to take new data and update the database and the key attributes of the prediction models. To simplify the process, the code generates model attributes in MS Excel format. The updated AI-PDET would help users to obtain project duration predictions that are more relevant to recent projects.

References

Abdel-Raheem, M., Torres Cantu, C., & Wang, X. (2020). Dynamic Contract Time Determination System for Highway Projects. *Transportation Research Record*, *2674*(5), 381–392. https://doi.org/10.1177/0361198120915896

Al-saadi, A. M., Kh. Zamiem, S., A. Al-Jumaili, L. A., JameelJubair, M., & Abdalla Al-Hashemi, H. (2017). Estimating the Optimum Duration of Road Projects Using Neural Network Model. *International Journal of Engineering and Technology*, *9*(5), 3458–3469. https://doi.org/10.21817/IJET/2017/V9I5/170905007

Alikhani, H., Le, C., & Jeong, H. D. (2020). *A Deep Learning Algorithms to Generate Activity Sequences Using Historical As-built Schedule Data*. 2–6. https://doi.org/10.3311/ccc2020-039

Cheng, M.-Y., Chang, Y.-H., & Korir, D. (2019). Novel Approach to Estimating Schedule to Completion in Construction Projects Using Sequence and Nonsequence Learning. *Journal of Construction Engineering and Management*, *145*(11), 04019072. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001697

Connor, J. O. (2004). *Development of Improved Information for Estimating Construction Time : a Report* (Vol. 7, Issue 22). www.ntis.gov

Daradkah, M., Al, E., & Qusef, J. A. (2018). Top-Down vs. Bottom-Up in Project Management: A Practical Model. *Proceedings of the Fourth International Conference on Engineering & MIS 2018 - ICEMIS '18*. https://doi.org/10.1145/3234698

FHWA. (2002). *FHWA Guide for Construction Contract Time Determination Procedures - Contract Administration - Construction - Federal Highway Administration*. https://www.fhwa.dot.gov/construction/contracts/t508015.cfm

Fraser, J. (2017). *30 road projects deferred as Montana Department of Transportation faces budget shortfall | State & Regional | helenair.com*. https://helenair.com/news/state-andregional/30-road-projects-deferred-as-montana-department-of-transportation-faces-budgetshortfall/article_d0e8dfa2-e9ad-5f6f-9561-19b9fd501ac0.html

Gondy, C., & Hildreth, J. (2007). *Contract Time Determination Guidelines*. http://www.virginiadot.org/business/resources/const/0708_contracttimedeterminationguidelines.pdf

Gransberg, D. D., David, H., Karaca, J. I., & Gardner, B. (2017). *TOP-DOWN CONSTRUCTION COST ESTIMATING MODEL USING AN ARTIFICIAL NEURAL NETWORK Final Report prepared for THE STATE OF MONTANA DEPARTMENT OF TRANSPORTATION in cooperation with*.

Hancher, D. E., McFarland, W. D., & Alabay, R. T. (1992). *Construction contract time determination*. *1262-1F*.

Hildreth, J. C. (2005). *A Review of State DOT Methods for Determining Contract Times State DOT Methods for Determining Contract Times*. *0105*.

Jeong, H. D., & Alikhani, H. (2020). *Activity Sequencing Logics Using Daily Work Report Data*.

Jeong, H. D., Le, C., & Devaguptapu, V. (2019). *EFFECTIVE PRODUCTION RATE ESTIMATION USING CONSTRUCTION DAILY WORK REPORT DATA THE STATE OF MONTANA DEPARTMENT OF TRANSPORTATION in cooperation with*.

Jeong, H. D., Oberlender, G., Atreya, S., Research, G., Venkatesh, A., Graduate, A., Assistant, R., Loafman, P., Preconstruction, O., Manager, P., & Schmitt, B. (2008). *Final Report DEVELOPMENT OF AN IMPROVED SYSTEM FOR CONTRACT TIME DETERMINATION (Phase I & II)*. https://shareok.org/handle/11244/299485

Jeong, H. D., Shane, J., Scheibe, K., Nilakanta, S., & Alikhani, H. (2019). *Optimizing Maintenance Equipment Life-Cycle for Local Agencies*.

Jeong, H. D., Gransberg, D. D., & Shrestha, K. J. (2017). Advanced Methodology to Determine Highway Construction Cost Index (HCCI). In *FHWA/MT-17-006/8232-001*. https://trid.trb.org/view/1474077

Jeong, H. S., Atreya, S., Oberlender, G. D., & Chung, B. Y. (2009). Automated contract time determination system for highway projects. *Automation in Construction*, *18*(7), 957–965. https://doi.org/10.1016/j.autcon.2009.04.004

Jiang, Y., & Wu, H. (2004). Determination of INDOT Highway Construction Production Rates and Estimation of Contract Times. *JTRP Technical Reports*. https://doi.org/10.5703/1288284313455

Karaca, I., Gransberg, D. D., & Jeong, H. D. (2020). Improving the Accuracy of Early Cost Estimates on Transportation Infrastructure Projects. *Journal of Management in Engineering*, *36*(5), 04020063. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000819

MDT. (2020). *STANDARD AND SUPPLEMENTAL SPECIFICATIONS FOR ROAD AND BRIDGE CONSTRUCTION 2020 EDITION V1.2*.

MDT. (2021). *MDT BID ITEMS*. Https://Mdt.Mt.Gov/Other/Webdata/External/Contractplans/Reports/ENG-ITEM-LIST-2020.Pdf. https://mdt.mt.gov/other/webdata/external/contractplans/reports/ENG-ITEM-LIST-2020.pdf

Mensah, I., Adjei-Kumi, T., & Nani, G. (2016). Duration determination for rural roads using the principal component analysis and artificial neural network. *Engineering, Construction and Architectural Management*, *23*(5), 638–656. https://doi.org/10.1108/ECAM-09-2015-0148

Nevett, G., Goodrum, P. M., & Littlejohn, R. L. (2020). Understanding the Effect of Bid Quantities, Project Characteristics, and Project Locations on the Duration of Road Transportation Construction Projects during Early Stages. *Journals.Sagepub.Com*, 036119812095314. https://doi.org/10.1177/0361198120953142

Ohio DOT. (2020). *Contract Time Determination Tool*. http://www.dot.state.oh.us/Divisions/ConstructionMgt/Admin/Documents/Forms/DispForm.aspx?ID=554

Petruseva, S., Zileska-Pancovska, V., & Car-Pušić, D. (2019). Implementation of process-based and data-driven models for early prediction of construction time. *Advances in Civil Engineering*, *2019*. https://doi.org/10.1155/2019/7405863

Shrestha, K. J., Le, C., Jeong, H. D., & Le, T. (2019). *Mining Daily Work Report Data for Detecting Patterns of Construction Sequences*. 578–583. https://doi.org/10.3311/ccc2019-079

Shrestha, K. J., Jeong, H. D., & Gransberg, D. D. (2017). Multidimensional Highway Construction Cost Indexes Using Dynamic Item Basket. *Journal of Construction Engineering and Management*, *143*(8), 04017036. https://doi.org/10.1061/(asce)co.1943-7862.0001307

Taylor, T. R. B., Paul, P. E., Goodrum, M., Brockman, P. E. M., Bishop, E. I. T. B., Shan, E. I. T. Y., Roy, E. I. T., Sturgill, E., & Hout, P. E. K. (2013). *Updating the Kentucky Contract Time Determination System*.

Taylor, T. R. B., Sturgill, R. E., & Li, Y. (2017). Practices for Establishing Contract Completion Dates for Highway Products. In *Practices for Establishing Contract Completion Dates for Highway Products*. Transportation Research Board. https://doi.org/10.17226/24705

Texas DOT. (2018). *Contract Time Determination guidance*.

Zhai, D., Shan, Y., Sturgill, R. E., Taylor, T. R. B., & Goodrum, P. M. (2016). Using Parametric Modeling to Estimate Highway Construction Contract Time. *Transportation Research Record: Journal of the Transportation Research Board*, *2573*(1), 1–9. https://doi.org/10.3141/2573-01