**1. Abstract**

This project explores the challenges and complexities of the construction permitting process in Austin, Texas and employs a data-driven approach to enhance efficiency and transparency. It uses a comprehensive dataset from the City of Austin, containing over two million records with several features related to construction permits like Permit Type, Permit Class, Applied and Issued Dates. This study applies several machine learning algorithms to gain deeper insights and optimize the permitting process. Algorithms like Random Forest, Gradient Boosting, Logistic Regression, Decision Trees, Isolation Forest and Exponential Smoothing were employed to predict the permit issuance time, classify permits based on descriptions, detect anomalies to indicate potential errors and forecast the demand of permit applications. The findings reveal significant insights into the distribution and timings of permits, the efficiency of the permit issuance process, and the potential for improving operations through data analytics. This project aims to facilitate a smoother permit process that could enhance urban planning, accelerate economic development, and improve compliance with construction standards, thus benefiting city planners, builders and the community.

**2. Introduction**

The construction industry is a crucial pillar of the US economy which drives a substantial employment and contribute to the development and enhancement of the necessary infrastructure in the country. According to Statista, a statistics and data visualization portal, the construction industry has contributed around 1.2 trillion U.S. Dollars in the year 2023 to the Gross Domestic Product (GDP) of the United States, which is around 4.4% of the USA GDP as shown in Figure 2-1 [1].

|  |
| --- |
| Figure 2-1: Share of value added to the GDP of the USA in 2023 by Industry [1] |

Previously, the construction industry contributed 1.09 trillion U.S. Dollars to the GDP which is lower than 2023. Figure 2-2 shows the trend of the amount contributed by the Construction Industry over the years 2000 to 2023 to the GDP of the United States [2].

|  |
| --- |
| Figure 2-2: Value added to the GDP of the United States by Construction Industry from 2000 to 2023 [2] |

Construction permits, critical to the operation of construction industry, are the regulatory documents issued by local government agencies to authorize the legal commencement of construction projects. Several types of development-related permits are issued for various construction projects like building permits for new constructions and interior remodeling or additions, electrical permits for auxiliary power, new construction, repairs and upgrades and home builder loop, plumbing permits for auxiliary water, septic tank and sewer tap, irrigation, repairs and new constructions, and driveway and sidewalk permit [3]. Construction permits are an important control mechanism to ensure that all the construction projects comply with planning policies, zoning laws and building codes for safety and environmental protection [4]. As cities get revamped constantly with the latest infrastructure and technological advancements, managing the construction permits becomes a critical factor to maintain systematic development and an orderly architectural standard during the construction process. Construction permitting is a 4 to 6 step process as shown in Figure 2-3 [5].

|  |
| --- |
| Figure 2-3: Construction Permit Process [5] |

Proper construction regulation using permits is necessary not only for the safety of the public by protecting them from faulty buildings, but also for the health of the construction sector and the economy as a whole. Property rights can be strengthened by having a well-functioning building permitting and inspection system in place [6]. Construction permitting also leads to safe and responsible construction practices. Building permits need a review of construction plans by qualified professionals to ensure building codes and safety standards are met, which in turn helps in identifying potential hazards and that necessary precautions are taken to minimize them. Building codes are put in place to protect safety and welfare of public by maintaining standards for construction design, materials and practices, and can result in legal consequences, demolition or fines if not adhered to [4]. Property values can be increased for homeowners and businesses if the construction project is properly permitted contributing to the overall desirability and quality of the neighborhood, while unpermitted constructions can potentially diminish the value of the property. Building permits also help in retaining records with vital information which acts as a source for future city planning.

Despite their significance, the construction permitting process can be challenging, complex and time-consuming for both the applicants and cities, as it involves multiple approval layers to address concerns from environmental impact to structural safety. Inefficiencies or delays in this process lead to significant financial costs for developers and builders leading to changes in the overall market dynamics of the construction industry [7]. Another challenge that cities, especially small municipal departments, can experience is the handling, processing and managing of the permit data due to the lack of the latest technologies and infrastructure [6]. KXAN News has highlighted that large cities like Austin usually tend to receive vast numbers of applications which has led to large backlogs and longer processing times [8]. A few of the primary factors contributing to permitting challenges include complex processes, resource constraints and poor data governance. One of the common challenges is the need for a central source and single POC for information. Local, State and Federal agencies have different timelines and requirements which can complicate the process [9]. In addition to addressing these challenges, the construction permitting process should be accelerated to encourage economic development in the city, reduce costs for the stakeholders or tenants involved, permanently increase the revenue of local governments and attract investment from other cities or countries [7].

To understand and enhance the construction permitting process, this research project aims to use data analytics and machine learning techniques on the ‘Issued Construction Permits’ dataset from the City of Austin [10]. This study helps in improving the methodologies of permit processes and facilitates data-driven decisions regarding the issuance of construction permits by constructing research questions. This study leverages a comprehensive dataset of more than a million records consisting of a diverse array of attributes like permit types, descriptions, contractor and applicant details, issue and expiration dates, descriptions of the projects etc. [11]. This project focuses on providing insights into the critical aspects of the construction permitting process like predicting the valuation of projects, the duration for issuing permits, identifying anomalies in the process, classifying permits based on descriptions and predicting future demand for permits.

The project aims to find answers to below research questions

1. How much time would it take to issue a permit based on the project details and permit type?

2. Are there any anomalies in the issuance process of permits indicating errors or fraud?

3. Classification of permits based on the project description and other details.

4. Forecast the future demand for construction permits based on historical data and trends.

This project aligns with the current need for efficiency and transparency of the government processes as analyzing the permit process and predicting the future trends not only helps construction companies and city planners but also benefits the citizens by speeding up the permitting process and ensuring compliance with urban standards. This research work is also timely as the need to provide data-driven solutions to optimize the operations of government departments and to improve public services has increased in several cities across the U.S.

We used machine learning algorithms and techniques like random forest and gradient boosting to predict the permit approval times by training and testing the data collected from the City of Austin’s website using attributes like permit type, project specifics etc. [12]. We used a real-time anomaly detection (RAD) technique called the Isolation Forest algorithm to identify potential anomalies in the permit issuance process which might indicate fraud or error [13]. We also used machine learning classification techniques like Naïve Bayes, Logistic Regression and Decision Tree to classify permits based on the project details. [14]. Also, the Exponential Smoothing algorithm was used to forecast the future demand for permit applications.

The report is organized as follows: Section 3 consists of the Literature Review and related research work; Section 4 consists of data description, pre-processing, descriptive statistics, and exploratory data analysis; Section 5 explains various methodologies and algorithms used in the project; Section 6 elaborates on the results and findings, followed by the Conclusion and Future Scope in Section 7.

**3. Literature Review**

Our literature review explores the traditional methods like manual inspection of construction permits, and duration of permits in the first section and identifies the need for an alternative to conventional methods. So, we also discuss the implementation of modern techniques such as machine learning (ML) models and artificial intelligence in the construction industry to predict accuracy in the second section.

**3.1 Traditional Approaches**

Time management in construction projects revealed that there are frequent delays in the permit process, leading to extended duration of projects and high costs [15]. Control managers are particularly concerned about the public who don't have full details about the development project during the permit process. This period indicated a growing awareness of the importance of professional supervision to ensure compliance with local regulations. Building permit processes revealed that obtaining permissions took excessively long, which caused a major obstacle to the construction of new homes. Permit applications can be processed more quickly by using Total Quality Management models [16]. The processing of building permits is sometimes affected by issues such as prolonged review times, a huge amount of paperwork and a lack of transparency. The general statement “that instead, there should be digitalization through which all issues concerning building permits would be resolved” suggests that there may have been an alternative. This indicates that such issues are best handled through further digitalization as there is no other feasible solution. By reducing the amount of human paper processing and utilizing computers to speed up the approval of building permits, modern technology can improve project implementation's overall efficiency [17].

Certain tasks involved in the process of getting approval for building permits could be done by a computer. To do this work, researchers talked to people who were involved in the process, looked at tasks to see if they could be mechanized and then sought after standard subroutines and common algorithms [18]. Globally recognized spatial data models are CityGML and IFC. Underlining the importance of these models in improving transparency and efficiency within the construction planning process. By enabling improved data integration and standardization, these models make it easier to process building permits, resulting in more effective systems with minimal chances of mistakes during implementation [19] [20].

Subsequently, several countries have started to implement e-government efforts targeted at digitizing administrative procedures, including construction permit approval. The argument was that to fully utilize the potential of digitization, all stakeholders, governments, businesses, civil society organizations, etc. must go digital, without introducing media breaks or changing formats [21]. Construction information flow can be modified, and process efficiency can be increased with the use of building information modeling (BIM). Because stakeholders participate at all levels, many people work together on digital building models under BIM, from design to the operational phase, which means no reentry and few errors. The construction sector saw a major transition in the 2010s from closed BIM to open BIM in development projects. In closed BIM, data sharing was restricted to software programs, which decreased its efficiency.

Meanwhile, open BIM allows information to be shared between different applications, promoting flexibility as well as accessibility on all levels [22]. Building permits in the USA are analyzed using data from web search queries, aiming to highlight the delay in the release of building permit data, which is a crucial factor of economic activity in the construction sector. By using Google Trends, specific terms like "new construction" and "new home construction” give strong predictive information. In foreseeing and forecasting building permits, the models that contain search queries do more effectively than other benchmark models. As this method makes use of freely accessible internet information, it is both economical and efficient. The research fills a gap in the existing literature by concentrating on predicting building permits directly rather than broader housing market conditions. The importance of using internet data for predicting in the construction business is discussed in the dataset details, forecasting models, analyzed methodologies and results [23].

**3.2 Modern Technologies for Predicting Building Permits**

Apart from traditional methodologies, ML technologies have gained more popularity in estimating building permits. Initially, machine learning was used for cost estimation, safety management and resource allocation during construction projects. For example, ML has been used to predict project costs by analyzing historical information about material prices, manpower and equipment hire charges. An exhaustive overview of machine learning can be divided into three main types: supervised, unsupervised, and reinforcement learning. Understanding various algorithms like K-nearest neighbors (KNN), support vector regression (SVR), gradient boosting trees (GBT) and artificial neural networks (ANN) is important for researchers to predict construction duration. Random Forest (RF) algorithms are used to develop a strategy to predict project duration based on requests for inspection. ML is very useful when predicting construction deadlines and accuracy improvement is necessary for the models [24].

In the early 2000s, the utilization of various linear regression (MLR) models became prominent. Factors influencing construction duration were identified and it was concluded that MLR afforded better predictions than the former models [25]. By the mid-2000s, neural networks (NNs) were in progress. In Ghana, they compared how well regression analysis and Artificial Neural Networks predicted bridge construction durations. ANN turned out to be somewhat less accurate, although still important in their case​​. Artificial Neural Networks in construction time models supported findings with their increased accuracy and dependability [26]. Case-based reasoning models help forecast the performance of new projects using historical cases. CBR application, which is supposed to estimate how long multi-housing projects may take with high accuracy [27].

There have been applications in this study of linear regression analysis methods like the logistic regression method, including the use of neural networks showing better forecast performance using statistical techniques. Artificial Intelligence methods have been employed, such as neural networks that provide superior forecasts compared with artificial intelligence models articulated. AI outcomes demonstrated the supremacy of artificial intelligence models over neural network forecasts, among other findings. The usefulness of ensemble methods, which leverage different machine learning techniques to increase prediction quality, recently attracted attention from several studies. Artificial neural networks together with support vector machines (SVM) for predicting construction costs and durations. Consequently, they achieved much higher accuracy levels than before [28]. The popularity of digital city models in the third dimension, together with their relevance in urban management applications like energy consumption forecasting, facility management, and emergency response system development, emphasized 3D urban modelling’s blossom to achieve smart growth, marking it out for helping different urban management areas through elaborate display and scrutiny of data [29]. The use of neural networks to forecast the performance of design-build projects in Singapore in the early 2000s. It improved various findings conceptually by indicating how neural networks have the capacity to work out intricate patterns in construction project information. They eventually came up with the conclusion that neural networks could predict project results with great precision by merely analyzing previous project information, hence suggesting a novel alternative [30].

Ultimately, construction managers use Support Vector Machines (SVM) to forecast road project costs with accuracy; they analyzed historical data from 70 completed jobs and identified 12 key factors that have a major impact on costs. Construction cost estimation is difficult because of a lack of information and a high level of uncertainty, particularly in the beginning stages of construction. Many factors like the project's size, location, design specifications, market condition and unexpected issues during construction all add to its complexity. It is important to have a skilled team as they have thorough knowledge and can understand the scope of the project. Other key requirements are creative contracting methods, which consider the local economy's impacts, industry capacity, material price fluctuations, and other design and construction requirements. The study makes use of several forecasting techniques to create an accurate cost estimation model, which includes random (Box-Jenkins models) and predictable (regression methods and econometric models) methods. Past information and related expenses are some of the factors that make historical data so important. The main goal is to address the common issue of cost overruns in road construction projects. The research obtains an efficient prediction accuracy of 95%, highlighting the effectiveness of SVM in parametric cost estimation. This will help the construction industry to manage projects more reliably and effectively [31].

Subsequently, the critical issue of cost risk and time overruns in civil engineering construction is highlighted by the most recent changes in the sand process in Tamil Nadu. To evaluate the importance of cost prediction, 324 research publications from the Scopus dataset between 1990 and 2015 are analyzed. It addresses several topics like other countries conducting this research, frequently used predictive tools (like Artificial Neural Networks, regression models, and time-series models), these tools usability, complexity, and accuracy. This study provides information for further research areas in construction cost prediction (CCP) by discovering trends and gaps in existing models, thus giving the importance of decision-making and predictive knowledge and telling us the importance of cost prediction tools in the construction industry [32]. It was a milestone to combine regression analysis with artificial neural networks (ANN), thus enabling forecasting schedules and quality performance for construction projects. Her study was geared towards establishing as well as comparing multiple linear regression (MLR) and ANN models. In her approach’s first phase, she would sift through certain write-ups in order to identify very important elements that could influence how well a given project performs, after which predictive models would be built around such aspects [33].

Recently, the construction industry has witnessed significant technological advancements using machine learning techniques and data models to predict project durations and speed up the permit process. The researchers carried out a study on construction projects in India to determine how machine learning could help in predicting delays. The result showed that when it comes to identifying risk factors or predicting time, a data-driven approach using machine learning algorithms leads to more accurate results [34].

Eventually, in a different paper, they used a combination of support vector machines and Monte Carlo methods for developing a hybrid forecast model that incorporated risks associated with time estimation and carried out an evaluation of the performance of different ML algorithms, such as linear regression [35]. ArcGIS Pro is used with a forest-based model to forecast building permits in Calgary, Canada. Here, they analyzed nine independent variables, including units under construction, the number of homeowners, and urbanization status. He worked with supervised machine learning strategies, and to achieve this, he trained the model using the previous building permit records while validating it using some parts of this data set. The process involved running the model with 500 trees and conducting ten runs for validation to avoid overfitting. Evaluating the model’s performance by using R-squared values, mean squared error and out-of-bag errors was then made with visualization of findings being made possible through box plots and prediction maps [36].

In conclusion, the transformation from traditional to machine learning-based techniques in construction duration estimation has been a major step forward in construction management. Presently, the highest level of this development is achieved by combining multiple algorithms into ensemble models that provide the most precise and dependable predictions possible. Future research should continue to expand datasets and explore new algorithms that further improve prediction capabilities as well as address particular challenges faced when constructing tall buildings. Ongoing progress and the use of machine learning in construction time estimation will result in better project management and outcomes in the building sector. ML’s role in project management and delays’ elimination through optimization will become very important as the construction industry continues its digital transformation. Further research can take advantage of various strengths offered by different machine learning models to create stronger prediction systems that can be scaled and generalized for a variety of construction projects due to their dynamic nature. In addition, ML is increasingly being combined with other emerging technologies, which will change the outlook of the field as it provides effective tools for builders to complete their designs in a timely and cost-efficient manner.

**4. Data Description and Analysis**

The ‘Issued Construction Permits’ dataset includes relevant details like issue date, location, latitude, longitude, project description, council district, square footage of the project, etc., for different types of permits issued by the City of Austin, such as driveway/sidewalk, plumbing, electrical, building, and mechanical permits.

This section is divided into two sections. The first one is the data description, which gives details about the dataset and its attributes, followed by data pre-processing. The second part is the data analysis, where the data is explored, and the descriptive questions of this research are answered.

**4.1 Data Description**

This dataset is sourced from the City of Austin and contains 68 columns and more than a million records where a row or record represents a permit. Below are the key columns and their descriptions. Appendix section contains the list of all the columns along their descriptions.

**Permit Type:** This field contains raw values representing the type of permits issued by the City of Austin in an abbreviated format like PP for Plumbing Permit, EP for Electrical Permit, MP for Mechanical Permit, DS for Driveway/Sidewalk and BP for Building Permits.

**Permit Type Desc:** This field is the description of the permit type and contain values like Plumbing Permit, Electrical Permit, Mechanical Permit, Driveway/Sidewalk and Building Permit.

**Permit Num:** This is a unique number or series assigned to each permit.

**Permit Class Mapped:** This field denotes the type of construction for which the permit is applied for, Residential or Commercial. This field is derived from the 'Permit Class’ column where the Permit Class Mapped will be Residential if the 'Permit Class' starts with "R", otherwise Commercial.

**Permit Class:** In this it shows the category of permit class whether it belongs to commercial, residential or other.

**Work Class:** The class of work in the permit is described and it specifies whether the work is new construction, remodel, repair, etc.

**Project Name:** This field provides the names of the construction project by unique name.

**Description:** This description gives the information about the project in detail.

**Applied Date:** This describes the date when the permit application was submitted or applied.

**Issued Date:** Date when the permit was issued is described here. This marks the approval of the permit.

**Day Issued:** Specific day the permit was issued. It provides the exact issuance date.

**Calendar Year Issued:** Calendar year the permit was issued.

**Fiscal Year Issued:** This field indicatesthe fiscal year in which each permit was issued.

**Status Current:** This column provides information about permit current status which indicates whether the permit is in-progress, expired, completed, etc.

**Status Date:** Date when the permit status was last updated and provides the latest update on the permit’s status.

**Expires Date:** Provides information of when the project permit will expire.

**Completed Date:** Gives the information about the project completed date and tracking project completion times.

**Original Address 1:** Original street address of the property which provides location information.

**Original City:** Original city of the property and indicates the city where the project is located.

**Original Zip:** Gives us information of the property location which specifies the postal code.

**Jurisdiction:** County overseeing the permit which provides information on the regulatory authority.

**Project ID:** Each and every project will be assigned an ID which can differentiate each project from the others.

**Latitude:** Latitude coordinate of the property which provides geographical positioning,

**Longitude:** Longitude coordinate of the property which specifies geographical positioning,

**Location:** Physical location of the property which combines latitude and longitude.

**4.2 Data Pre-Processing**

|  |
| --- |
| Figure 4-1: Heatmap representing missing values in the dataset |

The raw dataset contains over 1M records, which is huge and could become a bottleneck for the analysis later. Hence, only data from the year 2018 was considered for this research which has around 433K rows. The dataset does not contain duplicate values, as each record represents a permit. So, once the data is filtered based on the year the permit was issued, we check for missing values in the columns. Figure 4-1 represents a heatmap of the missing values showing a high proportion of missing data across various columns like Valuation fields, Applicant’s data, etc. The black area represents complete data, while the lighter band represents missing data.

To handle the missing values, a threshold is set to remove columns with more than 70% missing values. Also, columns which hold no relevance to the research, like 'Property Legal Description', 'Contractor Company Name', and 'Condominium', etc., are also removed to streamline the dataset and focus the research on the relevant information. Lastly, the rows with null values are handled by removing them from the dataset to ensure it is complete and ready for further analysis and modelling. As the dataset has over a million records, deleting few records from it wouldn’t impact the results of the prediction.

A new column 'Time to Issue' is created to calculate the duration or time it took for a permit to be issued. For this calculation, the existing date fields are converted to a DateTime format to handle datatype mismatch errors. From the below Figure 4-2, which is showing the distribution of 'Time to Issue' field, it can be inferred that there are outliers like negative values and records with more than 1000 days which seems to be inaccurate as the average expected time to issue a permit is around 6-8 months. These outliers are handled by removing the rows with negative values and duration greater than 1000 days. 1000 is taken as the threshold for this research considering the Covid-19 impact on the construction industry. Once the dataset is cleansed and pre-processed, there are 26 columns and around 360K rows in total.

|  |
| --- |
| Figure 4-2: Distribution of ‘Time to Issue’ field |

**4.3 Data Analysis**

This section has 2 sub sections. The first section presents the findings of the exploratory analysis followed by the descriptive data analysis in second section to answer few descriptive questions of this research.

**4.3.1 Exploratory Data Analysis**

As part of the Exploratory Data Analysis, frequency distribution graphs are created for various attributes like Calendar Year Issued, Fiscal Year Issued, Zip Code, Latitude, Longitude and Time to Issue.

* **Distribution by Calendar Year Issued:** The below histogram, Figure 4-3 shows that the permits issued are distributed almost evenly till 2022 followed by a decline in the year 2023.

|  |
| --- |
| Figure 4-3: Distribution of permits by Calendar Year Issued |

* **Distribution by Zip:** Figure 4-4 shows the Top 10 Zip codes with highest number of permits. It can be seen that the Zipcode 78744 has the highest number of permits issued from 2018 with nearly 24000 followed by 22000 permits issued for 78704 zipcode. Zipcodes 78723 and 78745 have nearly same number of permits issued at 17000.

|  |
| --- |
| Figure 4-4: Distribution of permits by Zip code |

* **Distribution by Time to Issue:** Figure 4-5 shows that the majority of the permits are issued within 50 days, with a rapid decline in the permits issued as the time to issue increases.

|  |
| --- |
| Figure 4-5: Distribution of permits by Time to Issue |

* **Distribution by Day Issued:** Based on the histogram in Figure 4-6, it can be seen that most of the permits are issued on weekdays, with the greatest number of permits issued on Tuesdays and relatively few permits issued on weekends.

|  |
| --- |
| Figure 4-6: Distribution of permits by Day Issued |

**4.3.2 Descriptive Data Analysis**

This section contains descriptive data analysis performed on the pre-processed dataset and focuses on answering a few descriptive questions of this research.

* **Distribution of Issued Permits by Permit Class:** The pie chart in Figure 4-7, shows that the majority of permits are issued for residential projects with 282K permits while only 79K permits were issued for Commercial projects. Hence, it can be inferred that the residential projects account for a significantly larger number of permits compared to commercial projects.

|  |
| --- |
| Figure 4-7: Distribution of Issued Permits by Permit Class |

* **Trend of Applied and Issued Permits by Calendar Year:** Figure 4-8 shows a line graph depicting the trend of permits applied and issued from the year 2018 to 2024. Both the applications and issuances show an increasing trend from 2018 to 2019, followed by a slight dip in 2020. This drop in count was recovered during 2021 but was followed by a decrease in the count over the later years, indicating a potential decline in permit activity in Austin. Also, the volume of permit applications and issuances has been consistently similar across the years.

|  |
| --- |
| Figure 4-8: Trend of Applied and Issued Permits by Calendar Year |

* **Distribution of Issued permits by Geographical Location:** The visualization in Figure 4-9, uses a geographical map with Austin as base location and shows the distribution of permits by Location. This is an interactive map, which means zooming in on an area shows the permit distribution on a detailed level as it uses Latitude and Longitude information. The colors on the map change according to the volume of the permits in that area. Orange signifies higher relative volume while yellow signifies medium volume and green signifies lower volume of permits. The teardrop symbol shows the map at the lowest level i.e., address if a residence/commercial space for which the permit is applied. [Interactive HTML File](file:///D:\Summer%202024\DSCI%205260%20-%20Business%20Process%20Analytics\Code\map_with_counts.html).

|  |
| --- |
| Figure 4-9: Distribution of Issued Permits by Location |

* **Trend of Permits Issued and Average Time Taken to Issue by Permit Type:** Figure 4-10 is a combination chart showing the count of permits for each permit type and the average time taken to issue a permit. It can be seen that Driveway/Sidewalks have a low count at 15K but have the highest average time to issue at 240 days, while the Electrical permit shows a reverse trend where it has the highest number of permits at 110K but takes less average time to issue a permit at 80 days. Building and Mechanical permits have a similar volume of permits at 70K, but the building permits are issued quicker in 75 days when compared to Mechanical at 130 days. Plumbing permits have the second highest count of permits at 100K and take a moderate issuance time of 105 days. This information highlights significant variations in the volume of permits and the issuance time across various permit categories, meaning that high volumes may not always mean that some licenses will be processed fast, whilst others may face considerable delays even if few people apply for them.

|  |
| --- |
| Figure 4-10: Distribution of Issued Permits by Location |

* **Distribution of Issued Permits across Current Status:** The chart in Figure 4-11 shows the distribution of permits across various statuses. The chart shows that the majority of the permits, around 357K, are in the ‘Final’ state, followed by 4.5K in the ‘Withdrawn’ state and 217 in the ‘Active’ state. Few permits are in Pending, Void and Aborted status, each with less than 100 permits. The chart also shows other statuses like Cancelled - Contractor Required, Inactive Pending Revision, Cancelled - New Permit Required, Pending Permit, On Hold, and “Expired’ with less than ten permits in each. Between all other categories and ‘final’, it is clear how efficient and effective a process of permission approval can be. Consequently, low counts for classifications like ‘Withdrawn’ or ‘Pending’ imply that few issues or delays arise during this process, with the majority of these licenses passing onto completion without any major challenges encountered. Most of these permits are closed, thus emphasizing efficient distribution.

|  |
| --- |
| Figure 4-11: Distribution of Issued Permits across Permit Status |

* **Top 5 cities by Permit Count:** Figure 4-12 shows the Top 5 cities in Austin with higher permit counts. Austin city has the highest number of issued permits by a significant margin of 359K, followed by other cities with lower counts like Bee Cave, with 1209 permits in second place. Lakeway City is in third place with 647 permits, closely followed by West Lake Hills at 561 and Rollingwood at 264 permits.

|  |
| --- |
| Figure 4-12: Top 5 cities by Permit Count |

**5. Methodology**

This section introduces and then explains various classification and regression algorithms and techniques used in our research to predict the issue time, classify the permit, anomaly detection, etc. The methodology will be implemented on the pre-processed dataset for better results and accuracy. When data is large the classification error tends to be less [37].

**5.1 Text Classification Techniques**

Although, there are several different classification techniques, selecting an appropriate algorithm for the dataset is crucial. As we can see from Figure 5-1, the text classification techniques are mainly divided into statistical and machine-learning approaches. The machine learning approach is again divided into three types: supervised learning, unsupervised learning and semi-supervised learning algorithms. Depending upon the dataset, there are different kinds of supervised learning algorithms like parametric-based and non-parametric classifiers. Logistic Regression and Naïve Bayes come under parametric supervised machine learning algorithms while, SVM, Decision Trees, Rule Induction, KNN, and Neural Networks come under non-parametric supervised machine learning algorithms. Fuzzy c-means, K-means clustering and Hierarchical clustering are a few unsupervised algorithms and Co-training, Self-training, Transductive SVM and Graph-based methods are a few semi-supervised learning algorithms [37].

|  |
| --- |
| Figure 5-1: Text Classification Techniques |

**5.2 Logistic Regression**

Logistic Regression is a statistical method for developing machine learning models used for binary classifications. It is a data analysis technique that is used to find the relationship between two data factors and is the best technique for calculating probabilities. It is mostly used to predict the resultant of a dependent variable. The outcome can be Success or Failure, Approved or Denied or Yes or No [38]. Logistic Regression, which is also defined as a model is a technique used to describe a binary result, like yes or no by looking at the previous data observations. It predicts the outcome of the variable by studying the connection between one or more variables. For example, Logistic Regression may be employed to foresee if a high school student will gain admission to a college. These binary results facilitate cut choices between two options [39].

A Logistic Regression model can be considered as input factors. For example, when it comes to college admissions, logistic regression model could consider some aspects like a student’s GPA, SAT score. By using the data related to input factors the model then assesses some of the new cases based on their likelihood of belonging to one of the two possible outcomes. Due to its easy interpretation and effectiveness in handling classification tasks, Logistic Regression is widely used across various industries like marketing, finance, and healthcare. For an example, it can help to predict the scenarios such as disease risk, credit default rates [39].

Logistic Regression Model in the Permit Analysis:In this study, logistic regression plays a crucial role in addressing important research inquiries and goals.

* Forecasting Permit Approval:Using the data regarding permit applications and their results in the logistic regression can check the approval status of given applications. This forecast considers various factors like the permit type, application date, etc.
* Identifying Key Factors:Logistic regression helps pinpoint and predict the variables that significantly influence the probability of permit approval. This insight is important for shaping policies and enhancing permit issuance procedures.
* Spotting Anomalies:Through a comparison between actual outcomes and the predicted probabilities, logistic regression results in identifying anomalies within the permit approval process. These findings may also have some signal issues, such as fraud or procedural errors.

Below are the steps followed for our analysis.

* Step 1: Training the Model: Initially, while creating the model, we will train the regression model by using some part of the data which mainly has the permit application records. The goal of the model is mainly to understand how the predictor variables relate in which the permit is approved or not.
* Step 2: Understanding Results: Coefficients and odd ratios: In the process of regression, we will get coefficients for each of the predictor variables that can be modified into some of the odd ratios. These ratios show how much the odds of approval change with a one unit increase in predictor variable.
* Step 3: Making Predictions and Decisions: Estimating the probability: The logistic regression model will provide probability scores of each permit application. These scores represent us how likely an application is   approved and can guide us to decision-making on which applications the examination is required.

Logistic Regression is a mathematical modelling technique that is suitable for characterizing the connection between one or more independent variables and a dependent variable when the outcome is discrete in nature. Failure and success are represented as numbers 0 and 1, which indicate the likelihood of outcomes in the order [40].

Where:

Ln: Natural log of the odds

X1: independent variable

P: probability of occurrence

C0: constant

C1: coefficient

The likelihood of occurrence in this form spans from 0 to 1, while the odds natural logarithm varies from -∞ to +∞.

In this project, the target variable is permitting status. The binary value of 1 denotes the issuance of the permit, and 0 indicates that the permit has not been issued. Based on this result, the logistic regression model would forecast the likelihood of a permit being granted by considering multiple independent variables, including the type of permission, the project description, and the applicant's details. The valuation of future building projects can be estimated by using some of the comparable qualities based on the past project valuation data by using some of techniques like regression. Project descriptions can also be processed by using text data methods such as natural language processing. Thus, feature engineering defines converting unprocessed data into modelling-ready features.

The performance of the machine learning model is evaluated to determine how effectively it will classify the new data. This process is called model evaluation. Metrics like Mean Absolute Error frequently help to monitor the accuracy of predictions for regression analysis of permit processing time [41]. Analysis measures include precision, recall, accuracy, F1-Score, and area under ROC (receiver operating characteristic curve) for classifying various tasks. To reduce the chance of overfitting and provide a more accurate estimate of generalization skills, cross-validation is used.

**5.3 Random Forest**

Random forest is a powerful and largely used machine learning technique, well known for its efficiency and accuracy in classification and regression tasks. Random Forest works on the idea of decision trees, where many trees are combined to obtain result which has more prediction accuracy, never becomes overfitting and handles noise and outliers.

**Principle of Operation:**

In 2001, Leo Breiman defined random forests (RF) as a collection of tree-structured classifiers {h (x, \ ), k=1, …} where {} are independent and similarly distributed random vectors. Each tree in the forest has a unit vote for the most popular class at input x [42]. Basically, an RF combines many tree-structured classifiers. Each tree is built using a training sample set and a random variable , with being independent and identically distributed across trees. This results in a classifier h (x, )) where x is the input vector. After generating k such classifiers, they form a sequence ({ (x), (x), …., (x)}. The final classification result is determined by majority vote, described by the decision function:

Where H(x) = the combined model

= an individual decision tree

Y = the output variable

I(.) = the indicator function.

**Characters of Random Forest:**

The margin function in RF measures how much the average number of votes for the correct class exceeds that for any incorrect class, defined as:

mg (X, Y) = I ( (X) = Y) – I ( (X) = j)

A larger margin value indicates higher accuracy and confidence in classification. The generalization error \(PE^\*\) is defined as:

PE\* = (mg (X Y = < 0)

Breiman proved that RF does not overfit and converges to a limiting generalization error as the number of trees increases. The upper bound of this error depends on the strength of the individual trees and the correlation between them, with the formula:

PE\* <= ρ (1-) /

s = the strength of the classifiers, and p is their average correlation.

**Out-of-Bag Estimation:**

During RF construction, trees are planted on new training sets created by random feature selection and bagging methods. Bagging enhances accuracy and provides out-of-bag (OOB) data, which is not included in the training set of any tree. Given a training set T with N samples, approximately 36.8% of the samples are OOB. OOB approximation uses these samples to assess the performance of the RF, providing an impartial estimate of generalization error. For each tree, the error is calculated using its OOB data, and the average error across all trees calculates the overall model performance. OOB estimation is as accurate as using a separate test set and more efficient than cross-validation, eliminating the need for a distinct validation set while continuously providing insights into the model's strength and correlation.

**Architectural components of Random Forest Algorithm:** There are several architectural components that makes Random Forest algorithm widely versatile for different kinds of tasks like outlier detection, classification and regression. Below are few architectural components of a Random Forest algorithm.

* Ensemble Learning: This architectural component of Random Forest uses multiple models with different unique advantages are combined to improve overall robustness and performance. By aggregating the predictions of multiple different decision trees, Random Forest can achieve better generalization and results compared to single decision trees.
* Bootstrapping: Bootstrapping is a training strategy which involves sampling data points with replacement from the original dataset. This results different subsets of data at each decision tree, which helps in variability in the training process.
* Random Feature Selection: Random Forest algorithm uses random feature selection during the training of each tree, random subset of features is chosen to ensure the model focus on different aspects of data.
* Decision Making and Voting: During prediction, random forest takes predications from each of different decision trees and aggregates them to give final prediction. For classification analysis the final prediction is determined by the mode most frequent prediction across all the trees and in regression analysis the average of individual tress prediction is taken.

**5.4 Isolation Forest**

The dataset used for this project has details about building permits at different timeframes. Due to this, we have different sets of data at each time frame. In some cases, the data will get deviated from the desired outcome, where in such cases, anomaly detection helps to identify new patterns which will dissect further to dive deeper into the analysis. Also, for the data like contractors to help analyze their strategy to know how they are performing in different time frames. By leveraging this concept, we can identify and effectively distinguish between normal and abnormal behavior which helps to mitigate unknown risks and sometimes these hidden patterns trigger new trends.

Anomaly Detection is the backbone of data analysis to identify patterns or events that deviate significantly from the norm in a dataset. Isolation random forest algorithm is an anomaly detection technique which works well with high dimensional problems which have a large number of irrelevant attributes. Isolation Forest is one technique we will try to use for this dataset. Isolation Random Forest performs favorably to ORCA, a near-linear time complexity distance-based method, LOF and random forests in terms of AUC and processing time, especially in large data sets [43]. Isolation Forest operates by isolating anomalies within a dataset through recursive partitioning.

Isolation forest is an anomaly detection technique famous for its efficiency and accuracy and uses binary partitioning. It quickly identifies anomalies in less computational time unlike traditional methods. It basically depends on proximity measures where it will randomly select features and split the data until individual data points are isolated. The isolation process is responsible for creating partitions which help to separate anomalies from normal observations. The working of isolation forest spans different stages from random partitioning to the classification of anomalies and the working algorithm is as follows:

* Random Partitioning: It operates on randomly selected data points from the dataset at random thresholds which helps to create isolation trees.
* Recursive Isolation: Each isolated data point is a subset of data that aims to separate normal observations from anomalies until data get smaller portions.
* Anomaly Identification: It identifies data points that require a few steps to isolate as they typically deviate from normal data points.
* Isolation Path: The tree measures the number of splits to isolate which measures the isolation score to measure the anomaly.
* Classification: The mean separation distance is calculated between all isolated trees from each data point which yields a standard deviation from the data where this standard deviation is useful to distinguish between normal data and abnormal data.

Isolation Forest is more accurate and yields better results compared to other types in terms of time and space complexity.

**5.5 Artificial Neural Network**

Artificial Neural Networks (ANNs) are computer models based on the structure and workings of the human brain. They consist of many simple processing units, or neurons, which are thickly interconnected. It is through these networks that we learn from and store experimental data. The main objective of an ANN is to gain high performance on tasks such as classification, prediction, or control by adaptation connection weights between its neurons [44].

**Neural Network Modeling:** Multi-Layer Perceptron’s (MLPs) is a well-known ANN used for modeling many processes. MLPs are feed-forward networks composed of multiple layers like input layer, output layer and one or more hidden layers as shown in Figure 5-2.

|  |
| --- |
| Figure 5-2: Topology of a three-layered feed-forward neural network [44] |

Each layer comprises neurons that convert input signals into output signals through weighted connections as shown in below Figure 5-3.

|  |
| --- |
| Figure 5-3: Schematic Diagram of a neuron [44] |

The output of each neuron is deceived by applying an activation function to the inner product of its input vector and its weight vector, subtracting a bias term. A sigmoid function is a common activation function used in MLPs, which can be defined as:

The smooth gradient and interest in using this function makes it helpful for training networks using gradient-based optimization methods. The sigmoid activation function is specifically useful for its bounded, non-decreasing response, making it capable of mapping any nonlinear process.

**6. Findings**

We have used several machine learning algorithms like Linear Regression, Random Forest and Gradient Boosting to predict the time taken to issue a permit. Techniques like Logistic Regression, Naïve Bayes and Decision Tree were used to classify a project into a specific type of permit based on the description. While Isolation Forest has been used to detect the anomalies in the dataset, the Exponential Smoothing technique was used to forecast the permit applications in the future.

**6.1 Prediction of ‘Time to Issue’ a permit**

The Random Forest Regressor demonstrates superior performance compared to the Gradient Boosting Regressor in several key metrics. The Mean Absolute Error (MAE) for the Random Forest Regressor is significantly lower at 43.90 compared to 67.31 for the Gradient Boosting Regressor. This indicates that the predictions made by the Random Forest Regressor are, on average, closer to the actual values. Additionally, the R2 score, which measures the proportion of variance explained by the model, is higher for the Random Forest Regressor (0.66) compared to the Gradient Boosting Regressor (0.44). This means that the Random Forest Regressor is better at capturing the underlying patterns in the data.

The output also predicts the sample Permit with the following characteristics: Permit Type EP, Permit Class Sign Permit, Work Class Wall, original city Austin, latitude 30.27, longitude -97.75, applied date March 28, 2019. For this sample, it was obtained that the Random Forest Regressor predicts 4.39 days for Time to Issue, and the Gradient Boosting Regressor predicts much longer at 44.85 days.

|  |
| --- |
| Figure 6-1: Results of ‘Time to Issue’ prediction model |

Finally, the Random Forest Regressor does better in the measurement of prediction accuracy and proportion of variance explained in the data compared to the Gradient Boosting Regressor, which can be derived from a lower Mean Absolute Error and higher R2 score. Furthermore, the sample prediction for processing time when using a permit also points out the significant difference in performance between the two models, in which the Random Forest Regressor gives a much more reasonable estimate than the Gradient Boosting Regressor.

In conclusion, the lower Mean Absolute Error and higher R2 score achieved by the Random Forest Regressor, along with its more accurate sample prediction, suggest that it is the more reliable model for this prediction.

**6.2 Classification of a Construction Project**

**Logistic Regression:** The model as shown in Figure 6-2 displays balanced performance with reasonably good precision, recall, and F1 scores for all permit types. It is not overfitting the data and generalizes well over unseen data points, as the performance metrics are consistent, and the accuracy is relatively high at 63.5%.

**Naive Bayes:** This model as shown in Figure 6-3 tends to overfit, more prominently for the DS permit type, on which it appears to have perfect precision, recall, and F1 score (all 1.00). It performs well enough in accuracy (0.63) and shows high precision for some permit types (EP, MP, PP), but its high scores for DS suggest it may not generalize well to new data.

**Decision Tree:** As per Figure 6-4, the Decision tree model shows signs of overfitting, especially with the DS permit type, where it perfectly scores, too. The overall accuracy is the lowest, being 0.56, and it has significantly lower precision and recall to the BP permit type, meaning that generalizing new data does not turn out well.

|  |
| --- |
| A screenshot of a computer  Description automatically generated  Figure 6-2: Results of Logistic Regression Model |
| A screenshot of a computer  Description automatically generated  Figure 6-3: Results of Naïve Bayes Algorithm |
| A screenshot of a computer  Description automatically generated  Figure 6-4: Results of Decision Tree model |

The Logistic Regression model is the most reliable among the three, however, it performs well and does not present any significant issues concerning overfitting. Both Naive Bayes and Decision Tree show overfitting phenomena, especially for the DS permit type, which turns out to be perfect. However, this directly implies that the models are more likely to perform poorly with data that are outside the norm. In that sense, Logistic Regression is the model of choice for predicting permit types in this scenario.

**6.3 Anomaly Detection in the Permit Issuance process**

We have used the Isolation Forest technique to identify the characteristics of anomalies that are few or many to detect anomalies in the dataset. We found 3k+ anomalies in the dataset. If the time to issue duration is longer or shorter and if the longitude and latitude coordinates are out of the permit location then we consider those applications as anomalies.

Below visualization, Figure 6-5 shows the output of Isolation Forest. The green color represents normal data and the red color represents anomalies that were detected by using location, apply date, issue date, complete date and permit type features.

|  |
| --- |
| Figure 6-5: Scatter Plot of Anomalies |

In the below Figure 6-6, we can see that the time to issue days are 3 days,995 days, and 855 days which are longer and shorter than expected.

|  |
| --- |
| A screenshot of a computer  Description automatically generated  Figure 6-6: Detailed Anomalies Data |

And, we wanted to predict the mean-time issue of construction permits for these Anomalies and compare it with normal data. So, we have given a condition for isolation forest where it is to detect anomalies of construction permits whose time to issue a permit is greater than 800 days. Upon training the model assumes different contamination values, which is a key parameter used by the isolation forest algorithm as it directly impacts the specificity and sensitivity of the anomaly detection. So, by adjusting this parameter we can compare, and identify between detecting too many anomalies for tailored analysis. We assumed the data has 1%, 5%, and 10% anomalies so the contamination of the data having anomalies was set at 0.01,0.05,0.1 to train the algorithm. The results were visualized using scatter plots and box plots to identify outliers.

|  |
| --- |
| A screen shot of a graph  Description automatically generated  Figure 6-7: Scatterplot of Anomalies Detected by Isolation Forest with Contamination = 0.01 |

From the above scatterplot Figure 6-7, we can see permit class on the x-axis and the time to issue on the y-axis and contamination is taken at 0.01 which yielded 21 anomalies across the data.

|  |
| --- |
| A screenshot of a graph  Description automatically generated  Figure 6-8: Box Plot of Time to issue permit distribution for Anomalous Data Points with Contamination = 0.01 |

For the above box plot in Figure 6-8, we have taken time to issue a permit on the y-axis and we identified the outliers where 0 on the x-axis indicates normal data and 1 indicates anomalous data and the contamination was set at 0.01.

|  |
| --- |
| A screenshot of a computer  Description automatically generated    Figure 6-9: Output generated by Isolation Forest with contamination = 0.01 |

As we can see from Figure 6-9, it shows the number of anomalies detected at 0.01 contamination is displayed as Excel and where contamination=0.01 field displays anomalies as 1 otherwise 0. We predicted the meantime to issue a normal permit at contamination 0.01 is 890.601684, whereas the predicted mean time to issue an anomalous permit is 983.5238037.

From the below scatterplot Figure 6-10, we can see permit class on the x-axis and the time to issue on the y-axis and contamination is taken at 0.05, which gives us 102 anomalies across the data.

For the below box plot in Figure 6-11, we have taken time to issue a permit on the y-axis and we identified the outliers, where 0 on the x-axis indicates normal data and 1 indicates anomalous data and here the contamination was set at 0.05.

|  |
| --- |
| A screen shot of a graph  Description automatically generated  Figure 6-10: Scatterplot of Anomalies Detected by Isolation Forest with Contamination = 0.05 |
| A graph showing a number of blue rectangular objects  Description automatically generated with medium confidence  Figure 6-11: Box Plot of Time to issue permit distribution for Anomalous Data Points with Contamination = 0.05 |

As we can see from Figure 6-12, it shows the number of anomalies detected at 0.05 contamination are displayed as excel, where contamination=0.05 field displays anomalies as 1 otherwise 0. And we predicted the mean time to issue a normal permit at contamination 0.05 is 889.447, whereas the predicted mean time to issue an anomalous permit is 931.509.

|  |
| --- |
| A screenshot of a computer  Description automatically generated    Figure 6-12: Output generated by Isolation Forest with contamination = 0.05 |
| A screen shot of a graph  Description automatically generated  Figure 6-13: Scatterplot of Anomalies Detected by Isolation Forest with Contamination = 0.1 | |

From the above scatterplot in Figure 6-13, we can see permit class on the x-axis and time to issue on the y-axis and contamination is taken at 0.1 which gives us 203 anomalies across the data.

|  |
| --- |
| A graph of a diagram  Description automatically generated with medium confidence  Figure 6-14: Box Plot of Time to issue permit distribution for Anomalous Data Points with Contamination = 0.1 |

From the above box plots we have taken time to issue a permit on the axis and we identified the outliers where 0 on the x-axis indicates normal data 1 indicates anomalous data and the contamination was set at 0.1.

|  |
| --- |
| A screenshot of a computer  Description automatically generated    Figure 6-15: Output generated by Isolation Forest with contamination = 0.1 |

As we can see from Figure 6-15, it shows the number of anomalies detected at 0.1 contamination is displayed as Excel and where contamination=0.1 field displays anomalies as 1 otherwise 0. We predicted the meantime to issue a normal permit at contamination 0.1 is 889.830, whereas the predicted mean time to issue an anomalous permit is 907.128.

**6.4 Demand Forecast of Permit Applications**

Exponential Smoothing technique was used to forecast the permit applications volume for the next 48 months. The below graph Figure 6-16, constitutes the time series data of construction permit applications for the years 2008 to 2023 and predicts the future permit applications up-to the year 2025. This model is applied to the monthly application counts and uses additive seasonality assuming that the changes are consistent over time.

The blue line represents the actual number of permit applications from the year 2008 to 2023. We can see the trend is a fluctuating over the years wherein there was a sudden increase in the year 2019 and afterwards there is a decline till 2023.

The orange dashed line shows the forecasted number of construction permit application for the next 48 months till the year 2025. It shows periodic fluctuations, designating an expected seasonal pattern.

Specific forecasted values are interpreted at notable points with numbers like 286, 313, 253, 236 and 187.

|  |
| --- |
| A graph showing the growth of a stock market  Description automatically generated  Figure 6-16: Demand forecast chart of permit applications |

**8. Conclusion**

The aim of this project is to study issued construction permits from the city of Austin to help improve the methodologies of permit processes and facilitate data-driven decisions regarding the issuance of construction permits. The study leverages a comprehensive dataset of more than a million records consisting of a diverse array of attributes like permit types, descriptions, contractor and applicant details, issue and expiration dates, descriptions of the projects etc. We use this data to provide insights into important factors like predicting the duration for issuing permits, identifying anomalies in the process, classifying permits based on descriptions and predicting future demand for permits.

The exploratory data analysis of Austin's permit issue revealed that permits were uniformly distributed by calendar year until 2023 before they started to fall. The highest permit counts were observed in Zip code 78744, and most permits were given within 50 days, mostly on weekdays, with a peak on Tuesdays. Residential permits greatly outnumbered commercial permits. The pattern indicated a rise in permits until 2019, a drop in 2020, and a continued fall after 2021. Geographically, large permit volumes were focused on specific places. Electrical permits were issued the fastest, while Driveway/Sidewalk permits took the longest. Most of the permits were processed swiftly, with Austin leading the way among cities in terms of permit count.

The project involved building different machine learning algorithms to find the factors involved in construction permit approvals. We used models like Random Forest and Gradient Boosting Regressor to predict permit issuance time. We compared all the model’s outputs, Random Forest is the most reliable model due to its low error rate and high predictive accuracy i.e., MAE of 43.90 days and a high R² score of 0.66, accurately predicting an issuance time of 4.39 days. Then we checked on the Classification of Permits Based on Project Details with Logistic Regression, Naive Bayes and Decision Tree, for which the Logistic Regression model is most effective with an accuracy of 64%.

To detect anomalies in the permit issuance process, the Isolation Forest method is used, identifying over 3,000 anomalies. Estimated the average time to issue licenses and compared anomalies to our data. By adjusting contamination levels (0.01, 0.05, 0.1), we detected different no. of anomalies. At 0.01 contamination, 21 anomalies had a mean issue time of 983.52 days (normal: 890.60 days); at 0.05, 102 anomalies had 931.51 days (normal: 889.45 days) and at 0.1, 203 anomalies had 907.13 days (normal: 889.83 days), illustrating the impact of contamination on anomaly detection. In forecasting the demand for permit applications, we used data from 2008 – 2023 and predicted future applications up to 2025, the blue line represents the actual number of permit applications and the orange line indicates the expected number of construction permit applications for the next 48 months. Some projections are interpreted at significant points with numbers like 286,313,253,236 and 187.

The future scope of the project is broad, with a focus on expanding analysis and combining more datasets to create a greater understanding of construction permit processes. Geographic expansion to other cities and areas, along with the implementation of new permit types like spatial analysis, will expand the study and improve the depth of outcomes. Integrating real estate market data and financial indicators will enhance the study by providing important context for project evaluations and market trend forecasts. Improved machine learning methods for anomaly detection and forecasting will increase the reliability and precision of detecting potential fraud and predicting future permit demand. Policy impact analysis will assess the consequences of new regulations on permit issuance time and project estimates, guiding future policy decisions. Creating interactive dashboards will give real-time insights to city planners and stakeholders, making data more accessible and useful, while constant integration of feedback from users ensures that the models continue to improve and are relevant.

In conclusion, our project answers the current need for efficiency and flexibility in government procedures. Evaluating the permit system and predicting future trends helps construction businesses, city planners, and everyone else by speeding up the permitting process and ensuring that it meets municipal standards. The study provides data-driven solutions to improve the functioning of government departments and public services throughout several regions of the United States.

# **References**

|  |  |
| --- | --- |
| [1] | Statista Research Department, "Statista," 2024. [Online]. Available: https://www.statista.com/statistics/248004/percentage-added-to-the-us-gdp-by-industry/. |
| [2] | Statista Research Department, "Statista," 2024. [Online]. Available: https://www.statista.com/statistics/785445/value-added-by-us-construction/. |
| [3] | Development Services Department, "AustinTexas.gov," City of Austin, [Online]. Available: https://www.austintexas.gov/page/types-permits. |
| [4] | Hudson Valley Permits, "LinkedIn," Hudson Valley Permits, [Online]. Available: https://www.linkedin.com/pulse/crucial-role-building-permits-construction-ensuring-c8cqc. |
| [5] | [Online]. Available: https://www.phoenix.gov/pddsite/MediaAssets/dsd\_image\_resflow.gif. |
| [6] | World Bank Group, "Subnational Studies," [Online]. Available: https://subnational.doingbusiness.org/en/data/exploretopics/dealing-with-construction-permits/good-practices. |
| [7] | PricewaterhouseCoopers LLP, "The Economic Impact of Accelerating Permit Processeson Local Development and Government Revenues," American Institute of Architects, National Economic Consulting, 2017. |
| [8] | T. Rahman, "KXAN News," June 2021. [Online]. Available: https://www.kxan.com/news/local/austin/amid-its-housing-shortage-austin-is-trying-to-keep-up-with-permits-to-build-more/. |
| [9] | KPMG, "Permitting: Streamlining delivery of today’s infrastructure opportunity," KPMG, 2023. |
| [10] | City of Austin, "Issued Construction Permits," City of Austin, May 2024. [Online]. Available: https://catalog.data.gov/dataset/issued-construction-permits. |
| [11] | City of Austin, "Open Data Network," May 2024. [Online]. Available: https://www.opendatanetwork.com/dataset/data.austintexas.gov/3syk-w9eu. |
| [12] | L. Florez-Perez, Z. Song and J. C. Cortissoz, "Predicting construction productivity with machine learning approaches," *39th International Symposium on Automation and Robotics in Construction (ISARC 2022),* vol. 2022, pp. 107-114, 2022. |
| [13] | B. Wu, F. Zhang, Y. Wang, M. Hu and X. Bai, "Anomaly Detection Algorithm for Urban Infrastructure," *Sustainability,* vol. 16, no. 8, p. 3335, 2024. |
| [14] | W. S. Alaloul and A. H. Qureshi, "Material Classification via Machine Learning Techniques: Construction Projects Progress Monitoring," in *Deep Learning Applications*, 2021. |
| [15] | D. W. Chan and . M. M. Kumaraswamy , "Forecasting construction durations for public housing projects: a Hong Kong perspective," *Building and Environment,* vol. 34, no. 5, pp. 633-646, 1999. |
| [16] | S. A. Amirbeigi, "A study of Optimizing the processing time for a building permit," 2012. |
| [17] | S. Malsane, J. Matthews, . S. Lockley, . P. . E. Love and . D. Greenwood, "Development of an object model for automated compliance checking," *Automation in Construction,* vol. 49, pp. 51-58, 2015. |
| [18] | F. Jalaei and A. Jrade, "Integrating building information modeling (BIM) and LEED system at the conceptual design stage of sustainable buildings," *Sustainable Cities and Society,* vol. 18, pp. 95-107, 2015. |
| [19] | F. Noardo, . F. Biljecki, . G. Agugiaro, K. A. Ohori, . C. Ellul, . L. Harrie and J. Stoter, "The GeoBIM benchmark 2019," *Transactions in GIS,* 2019. |
| [20] | S. M. İlal and H. M. Günaydın , "Computer representation of building codes for automated compliance checking," *Automation in Construction,* vol. 82, pp. 43-58, 2017. |
| [21] | M. Kassem and B. Succar, "Macro BIM adoption: Comparative market analysis," *Automation in Construction,* vol. 81, pp. 286-299, 2017. |
| [22] | A. Borrmann, . M. König, . C. Koch and . J. Beetz, Building Information Modeling - Technology Foundations and Industry Practice, Springer, 2018. |
| [23] | D. Coble and P. Pincheira , "Forecasting building permits with Google Trends," *Empirical Economics,* vol. 61, p. 3315–3345, 2021. |
| [24] | Z.-H. Zhou, "Open-environment machine learning," *National Science Review,* vol. 9, no. 8, 2022. |
| [25] | G. J. Hoffman, A. E. Thal and T. S. Webb, "Estimating Performance Time for Construction Projects," *Journal of Management in Engineering ,* vol. 23, no. 4, 2000. |
| [26] | I. Mensah, G. Nani and T. Adjei-Kumi, "Development of a Model for Estimating the Duration of Bridge Construction Projects in Ghana," *International Journal of Construction Engineering and Management,* vol. 5, no. 2, 2016. |
| [27] | . R. Jin, . S. Han, . C. Hyun and . Y. Cha, "Application of Case-Based Reasoning for Estimating Preliminary Duration of Building Projects," *Journal of Construction Engineering and Management,* vol. 142, no. 2, 2016. |
| [28] | I. Peško, V. Mučenski, M. Šešlija, N. Radović, A. Vujkov, D. Bibić and . M. Krklješ, "Estimation of Costs and Durations of Construction of Urban Roads Using ANN and SVM," *Artificial Neural Networks and Fuzzy Neural Networks for Solving Civil Engineering Problems,* vol. 2017, 2017. |
| [29] | F. Biljecki, . H. Ledoux and . J. Stoter, "Generating 3D city models without elevation data," *Computers, Environment and Urban Systems,* vol. 64, pp. 1-18, 2017. |
| [30] | F. Y. Y. Ling and . M. Liu, "Using neural network to predict performance of design-build projects in Singapore," *Building and Environment,* vol. 39, no. 10, pp. 1263-1274, 2004. |
| [31] | N. I. El-Sawalhi , "Support Vector Machine Cost Estimation Model for Road Projects," *Journal of Civil Engineering and Architecture,* vol. 9, pp. 1115-1125 , 2015. |
| [32] | . P. Velumani and . N. V. N. Nampoothiri, "Analysis of construction cost prediction studies – global perspective," *International Review of Applied Sciences and Engineering,* vol. 10, no. 3, p. 275–281, 2019. |
| [33] | A. T , "Application of Regression Analysis and Artificial Neural Network in Construction Project," *International Journal of Scientific Research & Engineering Trends,* vol. 5, no. 3, 2019. |
| [34] | . A. Gondia, . A. Siam , . W. El-Dakhakhni and . A. H. Nassar, "Machine Learning Algorithms for Construction Projects Delay Risk Prediction," *Journal of Construction Engineering and Management,* vol. 146, no. 1, 2019. |
| [35] | J. P. Fitzsimmons, R. Lu, Y. Hong and I. Brilakis, "Construction schedule risk analysis – a hybrid machine learning approach," *Journal of Information Technology in Construction,* vol. 27, pp. 70-93, 2022. |
| [36] | M. Jahanbani, "ECCE Blog: Prediction of Urban Building Permits: Forest-based Classification and Regression," 2024. [Online]. Available: https://ecce.esri.ca/ucalgary-blog/2024/03/29/prediction-of-urban-building-permits/. |
| [37] | T. M. and S. M. , "TEXT CLASSIFICATION TECHNIQUES: A LITERATURE REVIEW," *Interdisciplinary Journal of Information, Knowledge, and Management,* vol. 13, pp. 117-135, 2018. |
| [38] | P. Ranganathan, C. S. Pramesh and R. Aggarwal, "Common pitfalls in statistical analysis," *Perspectives in Clinical Research,* vol. 8, no. 3, pp. 148-151, 2017. |
| [39] | K. Yasar, "TechTarget," 2024. [Online]. Available: https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression. |
| [40] | MedCalc, "MedCalc," MedCalc, [Online]. Available: https://www.medcalc.org/manual/logistic-regression.php. |
| [41] | . Y. L. Joyce and Z. Shiyun , "RPubs," December 2019. [Online]. Available: https://rpubs.com/jyqliu/phillyresidentialhousingactivity. |
| [42] | . L. Yanli, . J. Zhang and Y. Wang, "New Machine Learning Algorithm: Random Forest," in *International Conference on Information Computing and Applications*, 2012. |
| [43] | F. T. Liu, . K. M. Ting and Z.-H. Zhou, "Isolation Forest," in *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy, 2008. |
| [44] | D. K. Panda and R. K. Bhoi, "Artificial Neural Network Prediction of Material Removal Rate in Electro Discharge Machining," *Materials and Manufacturing Processes,* vol. 20, no. 4, pp. 645-672, 2005. |
| [45] | I. WILLIAMSON, S. ENEMARK, J. WALLACE and A. RAJABIFARD, "Land Administration for sustainable development," *Facing the Challenges – Building the Capacity,* 2010. |
| [46] | J. B. Pedro, F. Meijer and H. Visscher, "Comparison of building permit procedures in European Union countries," *COBRA,* pp. 356-375, 2011. |
| [47] | C. Eastman, J.-m. Lee, . Y.-s. Jeong and J.-k. Lee, "Automatic rule-based checking of building designs," *Automation in Construction,* vol. 18, no. 8, 2009. |
| [48] | M. I. JORDAN and . T. M. MITCHELL, "Machine learning: Trends, perspectives, and prospects," *SCIENCE,* vol. 349, no. 6245, pp. 255-260, 2015. |
| [49] | J. C. Ang, A. Mirzal, H. Haron and H. . N. A. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection," vol. 13, no. 5, pp. 971 - 989. |
| [50] | . P. Velumani and . N. V. N. Nampoothiri, "Analysis of construction cost prediction studies – global perspective," *International Review of Applied Sciences and Engineering,* vol. 10, no. 3, pp. 275-281. |
| [51] | Z. Ma and Y. Ren, "Integrated Application of BIM and GIS: An Overview," *Procedia Engineering,* vol. 196, pp. 1072-1079, 2017. |
| [52] | N. . B. R. Mandi, I. B. P. Adnyana and I. P. E. Gunapatniyatsunu, "Factors Affecting the Success of the Building Permit Licensing Service in Denpasar City, Bali Province," *International Journal of Technology (IJTech),* vol. 10, no. 1, 2019. |

**Appendix**

**Data Description**

**Condominium:** This column represents if the project is a condominium or not.

**TCAD ID:** In this Travis Central Appraisal District ID for the property is defined and links the permit to the official property record.

**Property Legal Description:** In this the legal description of the property is described and it will provide the formal property boundaries and details which officially states the limits and specifics of the property.

**Issued In the Last 30 Days:** Indicates if the permit was issued in the last 30 days or more than 30 days and monitoring current permit issuance rates.

**Issuance Method:** Method by which the permit was issued and specifies whether the permit was issued online, in- person, etc.

**Total Existing Bldg SQFT:** Total square footage of the existing building and provides baseline building size.

**Remodel Repair SQFT:** Square footage of the remodel or repair work which indicates the scope of renovation work.

**Total New Add SQFT:** Total square footage of new additions where it will provide the size of new construction.

**Total Valuation Remodel:** Total valuation of the remodel if the project is remodeling work which indicates the financial value of renovation activities.

**Total Job Valuation:** Total valuation of the job, including all work types, which represents the overall financial investment in the project.

**Number Of Floors:** This column represents the number of floors in the housing units.

**Housing Units:** Indicates number of residential capacities in housing units.

**Building Valuation:** The estimated monetary value as per the current market of that area which represents the financial value of the building construction.

**Building Valuation Remodel:** Valuation of the building remodel if the project is remodeling work which indicates the financial value of building renovations.

**Electrical Valuation:** Valuation of the electrical work which represents the financial value of electrical installations.

**Electrical Valuation Remodel:** Valuation of the electrical remodel if the project is remodeling work which indicates the financial value of electrical renovations.

**Mechanical Valuation:** Valuation of the mechanical work which represents the financial value of mechanical installations.

**Mechanical Valuation Remodel:** Valuation of the mechanical remodel if the project is remodeling work which indicates the financial value of mechanical renovations.

**Plumbing Valuation:** Valuation of the plumbing work which represents the financial value of plumbing installations, and it is model instrument.

**Plumbing Valuation Remodel:** Valuation of the plumbing remodel if the project is remodeling work which indicates the financial value of plumbing renovations.

**MedGas Valuation:** Valuation of the medical gas work where it will represent the financial value of medical gas installations.

**MedGas Valuation Remodel:** Valuation of the medical gas if the project is remodeling work where this will indicate the financial value of medical gas renovations.

**Council District:** The council district where the property is located. Indicates the local governance area.

**Link:** Provides more information about the project permit and all necessary information of the project beneficial.

**Master Permit Num:** Master permit number, if applicable and links related to permits under a master permit.

**Contractor Trade:** Trade Category of the contractor (e.g., electrical, plumbing) which specifies the contractor’s specialization, important for understanding contractor roles and describes the contractor's area of expertise.

**Contractor Company Name:** The name of the contractors’ company for which contractors works for and this identifies the company responsible for the construction work.

**Contractor Full Name:** Legal name of the contractor. Provides the name of the individual contractor.

**Contractor Phone:** This field contains the contact information of contractor.

**Contractor Address 1:** The street address of the contractor where the contractor is permanent and the primary location of the contractor’s business.

**Contractor Address 2:** Alternative address information of the contractor and provides supplementary location details for the contractor’s business.

**Contractor City:** This field represents the city of the contractor located which identifies the city of the contractor’s business.

**Contractor Zip:** The zip code of the contractor is located which specifies the postal code of the contractor’s business.

**Applicant Full Name:** Legal name of the applicant, name of the individual applying for the permit and also the name of the person obtaining the permit.

**Applicant Organization:** The industry of the applicant is affiliated with, and it will identify the company or organization the applicant represents.

**Applicant Phone:** This field contains the contact numbers of the applicants.

**Applicant Address 1:** Permanent address of the applicant which specifies the primary location of the applicant’s residence or business.

**Applicant Address 2:** Alternative address for the applicant which provides supplementary location details for the applicant’s residence or business.

**Applicant City:** The city of the applicant is located which identifies the city of the applicant’s residence or business.

**Applicant Zip:** The zip code of the applicant is located and specifies the postal code of the applicant’s residence or business.

**Certificate Of Occupancy:** This field denotes whether the certificate of occupancy has been issued or not and confirms whether the building is approved for occupancy.

**Total Lot SQFT:** This field denotes the total square footage of the lot and specifies the size of the lot in square feet.