1 Introduction

The motivation for this project came from conversations with friends who own and operate a general contracting company in Atlanta, Georgia. One concern they described was the time required for a project manager to prepare an accurate bid. A bid depends on projected cost, time, and risk. Permit timing is one source of risk because delays in permit issuance can affect project scheduling before construction begins.

The dataset used in this project is the City of Atlanta building permit dataset for 2019 through 2024. The project focuses on issued residential permits and uses permit issue time as a proxy for schedule risk. The goal is to classify projects as `normal` or `high_risk` from permit information available in the dataset.

2 Methods

The raw permit data was filtered to issued residential permits with usable dates, latitude, longitude, zoning, land-use information, record type, and description. Permits approved in 0 days were removed. The response variable is based on

$$d_i = \texttt{record\_status\_date}_i - \texttt{date\_opened}_i,$$

where d_i is the number of days between the date the permit was opened and the date it was issued.

The cutoff was set at the 33rd percentile of d_i, which was 46 days. The bottom third of issued permits are labeled `normal`; the remaining permits are labeled `high_risk`:

$$y_i = \begin{cases} -1, & 0 < d_i \leq 46, \quad \texttt{normal}, \\ 1, & 46 < d_i, \quad \texttt{high\_risk}. \end{cases}$$

The processed dataset contains 7,504 issued residential permits. The training set uses 2019 through 2022, and the test set uses 2023.

Quantity	Count
Total records	7,504
Training records	5,988
Test records	1,516
Normal records	2,499
High-risk records	5,005

Table 1: Processed dataset counts.

Variable	Summary
`latitude`	$[33.023, 34.127]$
`longitude`	$[-85.040, -84.150]$
`application_month`	1 through 12
`record_type`	5 categories
`zoning`	93 categories
`cdp_land_use`	20 categories

Table 2: Variables used in the models.

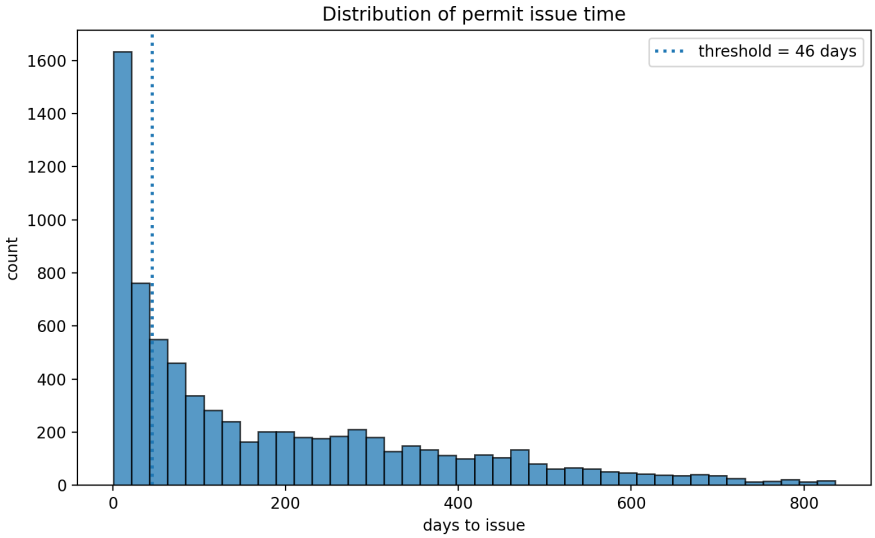

Figure 1: Distribution of permit issue time with the 46-day cutoff shown by the dotted vertical line.

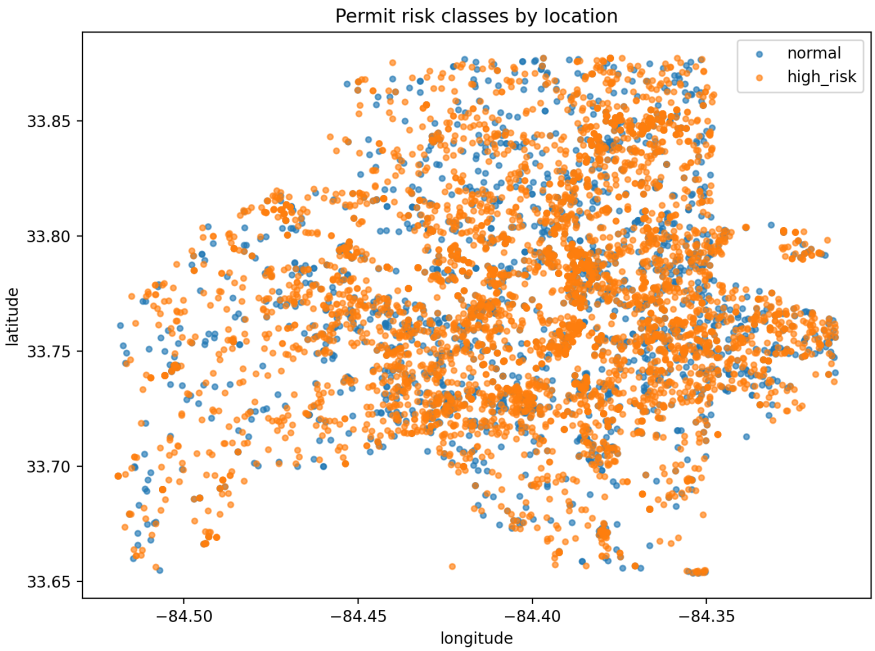

Figure 2: Spatial distribution of normal and high-risk permits after removing extreme location outliers from the plot.

2.1 Model 1: Kernel Support Vector Machine

The first model uses only location. For permit i, the input vector is

$$x_i = \begin{bmatrix} \texttt{latitude}_i \\ \texttt{longitude}_i \end{bmatrix} \in \mathbb{R}^2.$$

A support vector machine finds a decision function whose sign determines the predicted class. With a nonlinear feature map ϕ, the soft-margin optimization problem is

$$\min_{w,b,\xi} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i$$

subject to

$$y_i\left(w^T\phi(x_i) + b\right) \geq 1 - \xi_i, \qquad \xi_i \geq 0.$$

The model uses a radial basis function kernel,

$$K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|_2^2\right),$$

so the classifier can form nonlinear spatial decision boundaries without explicitly constructing $\phi(x)$. The prediction rule is

$$\hat{y} = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right).$$

The numeric inputs were standardized before fitting the SVM. The training matrix had shape 5988×2 after preprocessing.

2.2 Model 2: Feed-Forward Neural Network

The second model uses the structured permit variables:

`latitude`, `longitude`, `application_month`, `record_type`, `zoning`, `cdp_land_use`.

The numerical variables were standardized, and the categorical variables were one-hot encoded. After preprocessing, each project is represented by

$$x_i \in \mathbb{R}^{109}.$$

For a one-hidden-layer network with h neurons,

$$a_i^{(1)} = W_1^T x_i + b_1, \qquad z_i^{(1)} = \sigma(a_i^{(1)}),$$

where σ is either ReLU or hyperbolic tangent. The binary output is

$$\hat{p}_i = \frac{1}{1 + \exp\left(-W_2^T z_i^{(1)} - b_2\right)}.$$

The model is trained by minimizing binary cross-entropy loss,

$$\min_{W,b} -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i) \right] + \lambda \sum_{\ell} \|W_\ell\|_F^2.$$

Six neural network configurations were tested. The training data was balanced by random oversampling before fitting each network.

2.3 Model 3: Description Neural Network

The third model uses the same structured variables as Model 2 and adds the permit description. The description text was converted into term-frequency inverse-document-frequency features. For term j in permit description i, the text weight is

$$t_{ij} = \mathrm{tf}_{ij} \log\left(\frac{N}{n_j}\right),$$

where tf_{ij} is the term frequency, N is the number of descriptions, and n_j is the number of descriptions containing the term.

The text representation kept the 80 strongest single-word or two-word terms after removing common English stop words. The full input vector is

$$x_i = \begin{bmatrix} \text{structured variables}_i \\ \text{description terms}_i \end{bmatrix} \in \mathbb{R}^{189}.$$

The network used the strongest setup from the structured neural network comparison: one hidden layer with 24 ReLU neurons and learning rate 0.001.

3 Results

For all confusion matrices, `high_risk` is treated as the positive class. Sensitivity is the true positive rate for high-risk projects, and specificity is the true negative rate for normal projects.

3.1 Model 1

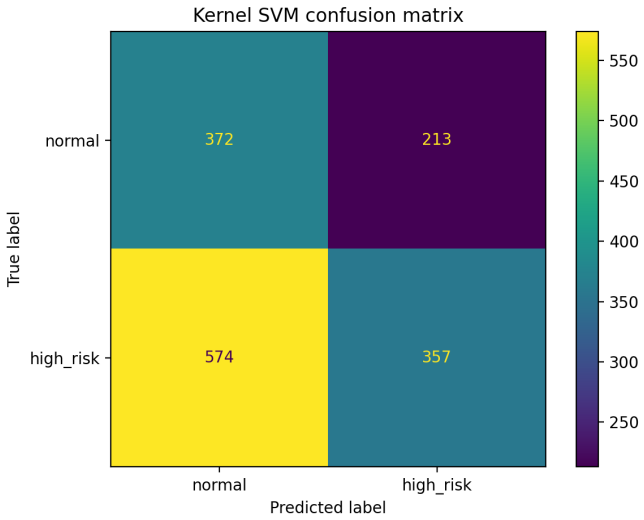

Figure 3: Confusion matrix for the kernel SVM.

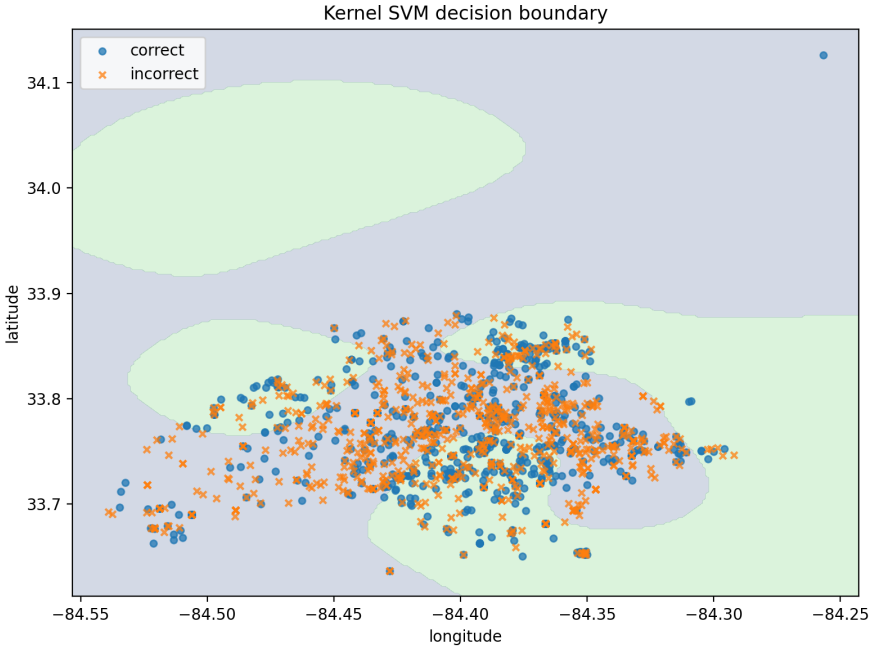

Figure 4: Kernel SVM decision boundary using longitude and latitude. Circles mark correct predictions, and crosses mark incorrect predictions.

Model	Accuracy	Precision	Sensitivity	Specificity
Kernel SVM	0.4809	0.6263	0.3835	0.6359

Table 3: Model 1 test metrics. The confusion matrix counts were $TN = 372$, $FP = 213$, $FN = 574$, and $TP = 357$.

3.2 Model 2

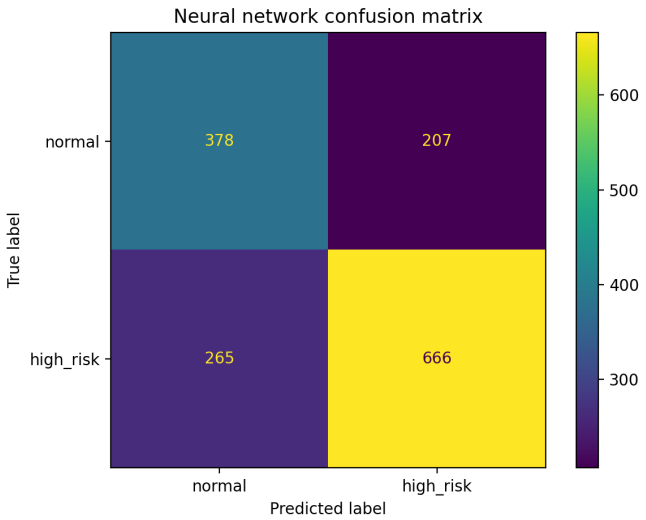

Figure 5: Confusion matrix for the highest-accuracy structured neural network.

Setup	Accuracy	Precision	Sensitivity	Specificity
$h = 12$, ReLU, $\eta = 0.001$	0.6867	0.7621	0.7121	0.6462
$h = 24$, ReLU, $\eta = 0.001$	0.6887	0.7629	0.7154	0.6462
$h = 24$, tanh, $\eta = 0.005$	0.6860	0.7588	0.7164	0.6376
$(16, 8)$, ReLU, $\eta = 0.001$	0.6840	0.7557	0.7175	0.6308
$(24, 12)$, ReLU, $\eta = 0.001$	0.6873	0.7605	0.7164	0.6410
$(32, 16)$, tanh, $\eta = 0.005$	0.6814	0.7523	0.7175	0.6239

Table 4: Model 2 neural network comparison.

The highest test accuracy in Model 2 was obtained by the one-hidden-layer network with 24 ReLU neurons. Its confusion matrix counts were $TN = 378$, $FP = 207$, $FN = 265$, and $TP = 666$.

3.3 Model 3

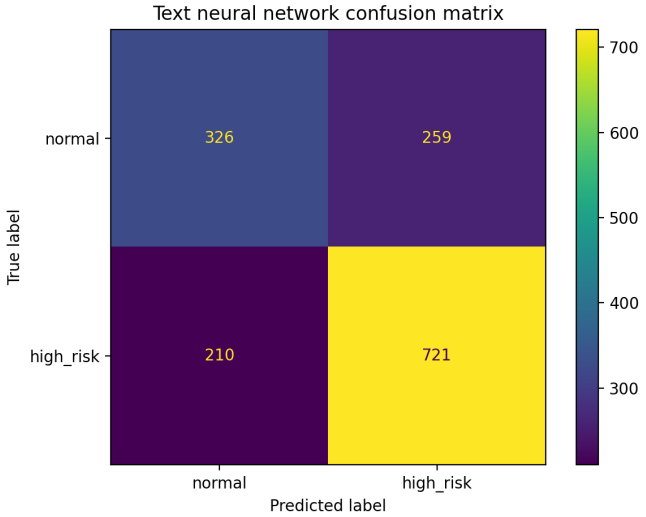

Figure 6: Confusion matrix for the description neural network.

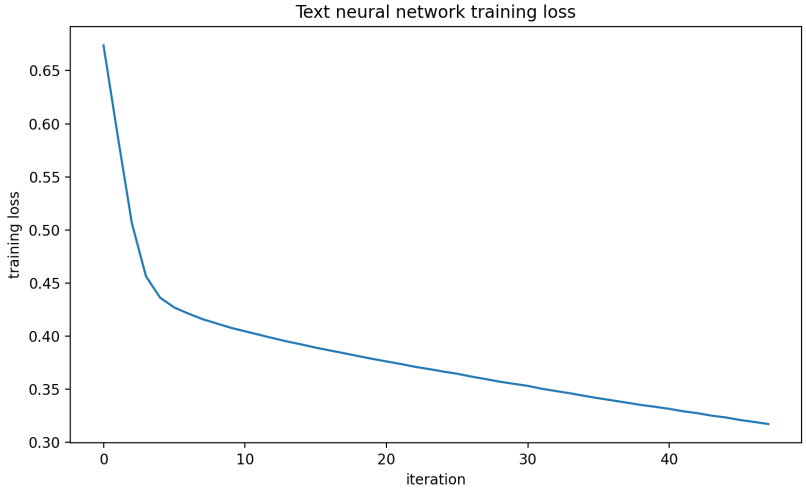

Figure 7: Training loss for the description neural network.

Model	Accuracy	Precision	Sensitivity	Specificity
Description neural network	0.6906	0.7357	0.7744	0.5573

Table 5: Model 3 test metrics. The confusion matrix counts were $TN = 326$, $FP = 259$, $FN = 210$, and $TP = 721$.

The highest-weighted description terms for high-risk projects included *new*, *addition*, *story*, *construction*, *garage*, *single*, *single family*, *porch*, and *rear*. These terms are consistent with projects that tend to involve larger construction scope.

4 Discussion

The location-only kernel SVM performed poorly compared with the neural networks. Its accuracy was below 50%, and its sensitivity was 0.3835, meaning it missed many high-risk permits. This suggests that latitude and longitude alone are not enough to classify permit-delay risk.

The structured neural networks produced much better overall performance. Their metrics were similar across configurations, with accuracy around 68% to 69%. The best structured network had accuracy 0.6887, precision 0.7629, sensitivity 0.7154, and specificity 0.6462.

Adding description text slightly increased accuracy to 0.6906 and increased sensitivity to 0.7744. The tradeoff was lower specificity, meaning the text model found more high-risk projects but also classified more normal projects as high risk. For a risk-filtering problem, this tradeoff can be useful because false negatives represent projects that were actually high risk but were not flagged.

5 Limitations

The main limitation is that the target is based only on permit issue time. A project may take longer for reasons not fully represented in the dataset, such as missing documents, applicant delays, staffing, review workload, or project-specific regulatory issues.

The cutoff at 46 days is also an assumption. It separates the fastest third from the rest, but it does not come from a business cost model. A contractor may want a different cutoff depending on schedule tolerance and the cost of incorrectly rejecting a project.

The model also uses only one train-test split. Since the test set is 2023, the results measure performance on one future year. Additional validation across multiple time periods would give a better estimate of model stability.

Finally, the description model uses a simple TF-IDF representation. It captures repeated words and short phrases, but it does not understand sentence meaning, project context, or permit-review requirements.

6 References

[1] Procore. Construction bidding process. https://www.procore.com/library/construction-bidding-process

[2] City of Atlanta. All Building Permits 2019-2024. https://dpcd-coaplangis.opendata.arcgis.com/datasets/655f985f43cc40b4bf2ab7bc73d2169b/about

[3] Scikit-learn developers. Support Vector Machines. https://scikit-learn.org/stable/modules/svm.html

[4] Scikit-learn developers. Neural network models. https://scikit-learn.org/stable/modules/neural_networks_supervised.html

[5] Scikit-learn developers. TF-IDF vectorizer. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html