CONSTRUCTION DELAY PREDICTION

This poster summarizes my research on using machine learning methods to predict construction delays efficiently, leveraging public datasets for accuracy.

BACKGROUND & PROBLEM:

- COST OVERUNS
- LABOR INEFFIENCIES
- CONTRACTURAL DISPUTES
- PROJECT SCHEDULE DISRUPTIONS

DELAYS COMMONLY ARISE FROM WEATHER, LABOR SHORTAGES, PLANNING ISSUES, AND MATERIAL AVAILABILITY.
GOAL: USE MACHINE LEARNING TO PREDICT THE LIKELIHOOD OF PROJECT DELAY SO PROJECT MANAGERS CAN INTERVENE EARLIER.

DATASET:

- PROJECT MANAGEMENT RECORDS
- TASK AND FORM-LEVEL DATA

FINAL MERGED DATASET:

- 300 PROJECTS
- ENGINEERED FEATURES

CLASS DISTRIBUTION:

- 71% ON TIME
- 29% DELAYED

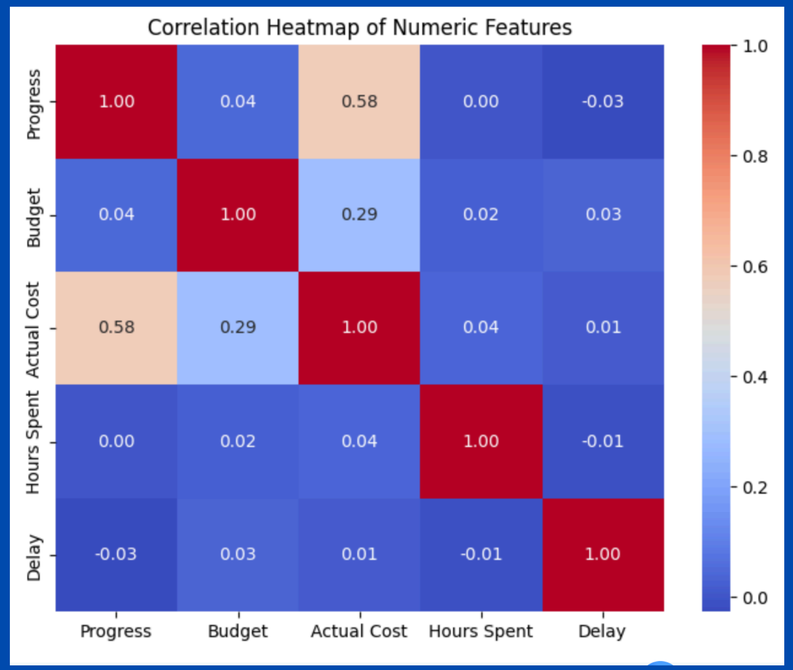

FIGURE 1: CORRELATION HEATMAP
NUMERIC FEATURES (BUDGET, PROGRESS, HOURS SPENT) SHOW NEAR-ZERO CORRELATION WITH DELAT. THIS SUGGETS NUMERICAL FIELDS ALONE DO NOT STRONGLY PREDICT DELAY.

METHODS:

THE MODELING APPROCH USED LOGISTIC REGRESSION, SLECTED FOR ITS INTERPRETABILITY AND STRONG PERFORMACE ON STRCUTURED TABULAR DATA.

PREPROCESSING:

- STANDARDIZING INCONSISTENT COLUMNS
- MERGING DATASETS USING PROJECT ID
- CLEANING MISSING VALUES
- CONVERTING DATE FORMATS
- AGGREGATING TASK/ FORM METRICS INTO PROJECT LEVEL FEATURES
- ONE-HOT ENCODING CATEGORICAL VARIBALES
- SCALING NUMERICAL FEATURES

KEY INPUTS:

- BUDGET, ACTUAL COST
- HOURS SPENT
- PROJECT STATUS (ENCODED)
- TASK COMPLETION METRICS
- PROGRESS INDICATORS

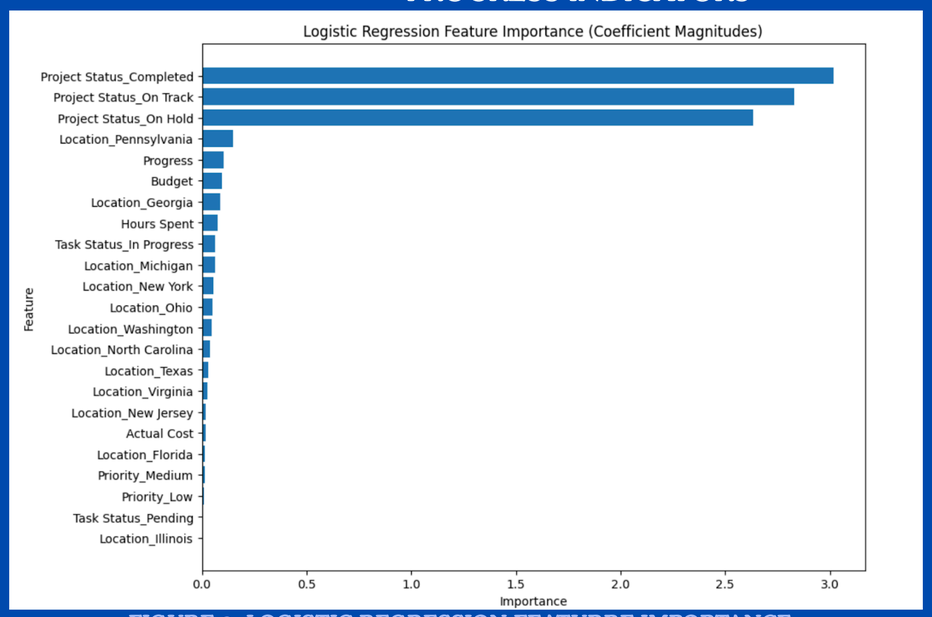

FIGURE 2: LOGISTIC REGRESSION FEATURRE IMPORTANCE
PROJECT STATUS CATEGORIES DOMINATE MODEL INFLUENCE. THESE FEATURES LIKELY ENCODE DELAY RELATED INFORMATION, RAISING THE POSSIBILITY OF LABEL LEAKAGE.

RESULTS:

BASELINE CLASSIFIER:

- PREDICTS ALL PROJECTS AS "ON TIME:
- ACCURACY: 71.7%
- F1 SCORE FOR DELAYED CLASS=0

LOGISTIC REGRESSION MODEL:

- TRAINING ACCURACY: 1.00
- TESTING ACCURACY: 1.00
- F1 SCORE: 1.00

THE MODEL CORRECTLY CLASSIFIED ALL DELAYED AND ON-TIME PROJECTS; PERFORMANCE FAR EXCEEDS THE BASELINE

INTERPREPTATION: CERTAIN CATEGORICAL PROJECT-STATUS FEATURES PROVIDE HIGHLY PREDICTIVE SIGNALS THAT SEPERATE DELATED AND ON TIME PROJECTS PERFECTLY

ERROR ANALYSIS:

- NUMERIC VARIABLES HAD WEAK CORRELATION WITH DELAY
- CATEGORICAL PROJECT STATUS FEATURES OVERWHELMINGLY INFLUENCED PREDICTIONS

THE EXTREMELY HIGH ACCURACY SUGGETS:

- LABEL LEAKAGE
- DATASET IS SMALL AND HIGHLY STRUCTURED
- LIMITED REAL WORK VARIABILITY

THESE PATTERNS LIKELY MADE THE CLASSIFICATION TASK EASIER THAN IN REAL CONDITIONS

CONCLUSION:

THIS PROJECT DEMONSTRATES THAT LOGISTIC REGRESSION CAN PERFORM EXTREMELY WELL ON INTEGRATED PROJECT /TASK/FORM DATA FOR CONSTRUCTION DELAY PREDICTION. HOWEVER, PERFECT ACCURACY SUGGESTS THE PRESENSE OF STRONG CATEGORICAL INDICATORS TIED DIRECTLY TO DELAY STATUS.

FUTURE WORK:

- REMOVE OR RESTRUCTURE PROJECT STATUS FEATURES TO PREVENT LEAKAGE
- EXPAND DATASET WITH MORE DIVERSA REAL WORLD PROJECTS
- EXPERIMENT WITH RANDOM FOREST, GRADIENT BOOSTING, AND NEURAL MODELS
- ADD TEMPORAL FEATURES (TASK DURATIONS, DELAYS OVER TIME)

REFERENCES:

- ABIOYE ET AL. (2023). AI IN THE CONSTRUCTION INDUSTRY: A SYSTEMATIC REVIEW.
- AGHIMIEN ET AL. (2023). MACHINE LEARNING–BASED FRAMEWORK FOR CONSTRUCTION DELAY MITIGATION.
- ALICE TECHNOLOGIES. AI CONSTRUCTION SCHEDULE OPTIMIZATION.
- ELLIS, G. (2025). THE RISE OF AI IN CONSTRUCTION. AUTODESK.
- NGUYEN, N. (2025). AI USE CASES IN CONSTRUCTION. SMARTDEV.
- AHMADILMANASHRAF. (2025). PROJECT MANAGEMENT DATASET EXAMPLE [KAGGLE].
- MILLER, C. (2021). CONSTRUCTION AND PROJECT MANAGEMENT EXAMPLE DATA [KAGGLE].

HTTPS://GITHUB.COM/EMMAKUSHNER/CONSTRUCTION-DELAY-PREDICTION-