Title: AI to Mitigate Delays and Cost Overruns in NYC Public Infrastructure Projects

Subtitle: A Machine Learning-Based Risk Forecasting Framework

- By Souleymane Doumbia
- CUNY School of Professional Studies, M.S. in Data Science Capstone

- Data 698 | Capstone Project Presentation Spring 2025

Introduction: Tackling Public Infrastructure Risks

Problem: NYC capital projects often exceed budget and schedule.

- This project leverages AI to predict and classify risk patterns in New York City capital project data
- **Objective**: Enable proactive decision-making by predicting project cost, time performance, and composite profiles

Introduction: Tackling Public Infrastructure Risks

Extent to Which NYC-Managed Capital Projects are Over Budget

($ in billions)

"50% Over Budget" and "$54.5B Increase"

Category	Number of Projects	Share of Total	Original Budget	Current Plan	Percent Change	Dollar Change
20% or more	2,029	39.6%	$ 29.9	$ 83.3	179.1%	$ 53.5
At least 10%, Less than 20%	199	3.9%	3.8	4.4	14.5%	0.6
Over 0%, Less than 10%	354	6.9%	10.4	10.9	4.4%	0.5
Subtotal	**2,852**	**50.4%**	**44.1**	**98.6**	**123.5%**	**54.5**
On Budget	1,377	26.9%	11.1	11.1	0.0%	- - -
Under Budget	1,071	20.9%	35.5	23.4	-33.9%	(12.0)
Incomplete Data	98	1.9%	- - -	2.7	N/A	2.7
Total	**5,128**	**100%**	**$ 90.7**	**$135.8**	**49.8%**	**$ 45.1**

Extent to Which NYC-Managed Capital Projects are Delayed

($ in billions)

"64.6% Delayed" and "$34B Increase"

Category by Months	Number of Projects	Share of Total	Original Budget	Current Plan	Percent Change	Dollar Change
36 months or more	2,560	49.9%	$ 41.9	$ 72.3	72.5%	$ 30.4
At least 12, Less than 36	686	13.4%	6.7	9.5	41.4%	2.8
At least 3, Less than 12	65	1.3%	2.0	2.8	38.9%	0.8
At least 0, Less than 3	23	0.4%	1.0	1.1	15.7%	0.2
Subtotal	**3,311**	**64.6%**	**50.6**	**84.6**	**67.1%**	**34.0**
On Time	1,435	28.0%	28.6	35.4	23.8%	6.8
Accelerated	359	7.0%	10.5	14.7	40.1%	4.2
Total	**5,128**	**100%**	**$ 90.7**	**$135.8**	**49.8%**	**$ 45.1**

Source: NYS Comptroller Office Audit report, 2025

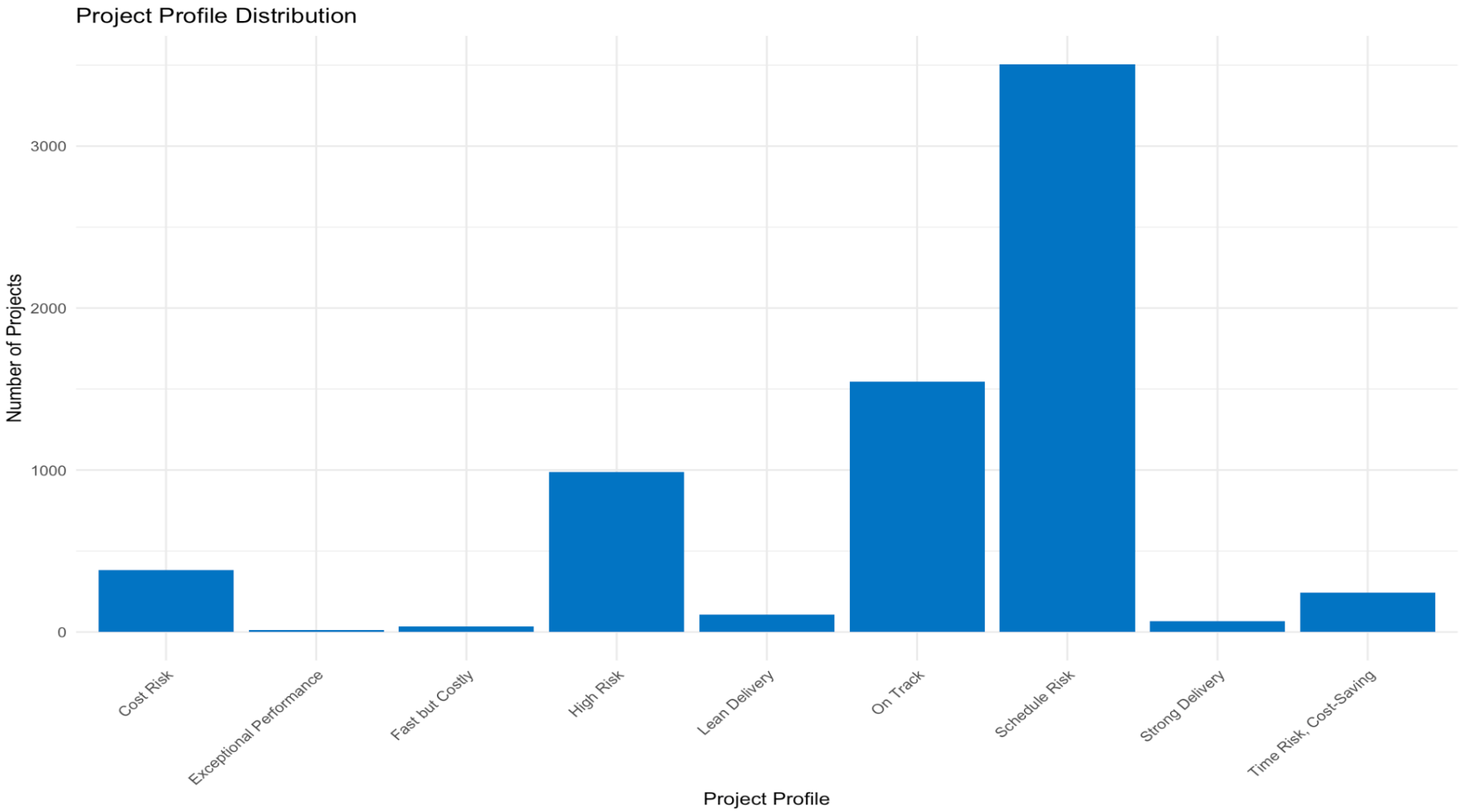
Data Display & Methodology

- **Data Sources:** Raw Datasets
  - NYC Capital Tracker
  - BLS Labor Statistics
  - NOAA Climate Data
  - NYS Comptroller Audit

- ↓ **Feature Engineering**

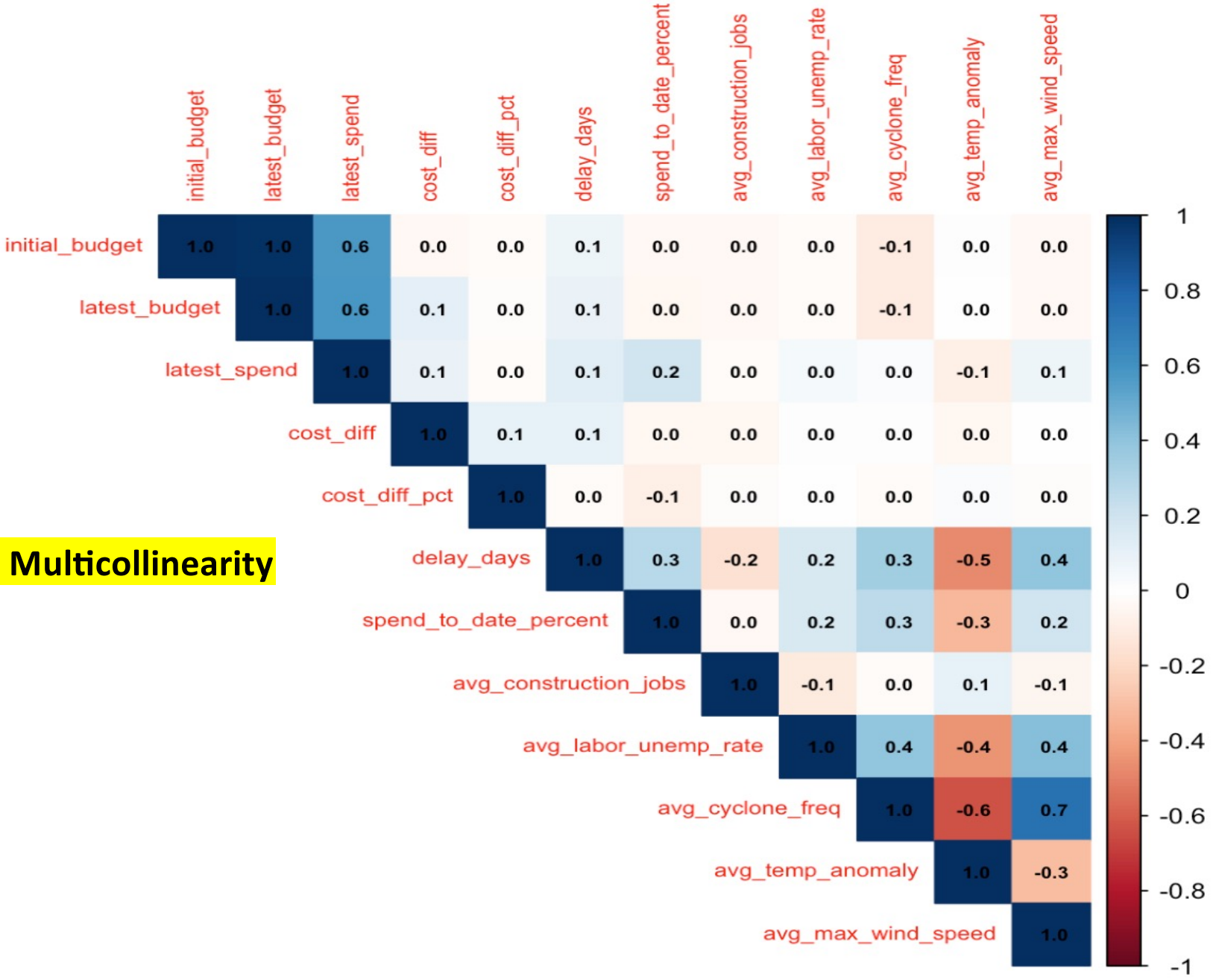
- **Derived Features :**
  - Budget Overrun Labels
  - Delay Category
  - Weather-Labor-Borough Overlays
  - Phase Duration Ratios
  - Project Type, Theme, Borough
  - Cost Performance Classification (Over/Under/On Budget)
  - Delay Classification (Early/Delayed/On Time)
  - Combined Risk Profile Classification (**project_profile: cost_class** & **delay_class**)

Data Display & Methodology



Data Display & Methodology

Limited Multicollinearity



Data Display & Methodology

Tasks
- Cost Performance Classification
- Delay Performance Classification
- Combined Risk Profile Modeling

Algorithms
- Random Forest
- XGBoost
- Decision Tree (CART)

Setup
- Multiclass labels
- Cross-validation (5-fold)
- Categorical and numeric feature handling

Output
- Project-level risk predictions
- Ranked variable importance (for later interpretation)

Modeling Results Summary

- We evaluated Random Forest, Decision Tree, and XGBoost on three prediction tasks

Model_Performance_Summary

Prediction Task	Model	Accuracy	Kappa	Balanced Accuracy (Avg)
Cost Classification	Random Forest	92.7%	0.802	~0.88
Cost Classification	Decision Tree	78.6%	0.299	~0.59
Cost Classification	XGBoost	94.1%	0.846	~0.90
Delay Classification	Random Forest	98.4%	0.964	~0.98
Delay Classification	Decision Tree	91.4%	0.800	~0.91
Delay Classification	XGBoost	98.7%	0.970	~0.99
Project Profile Classification	Random Forest	91.3%	0.867	~0.91
Project Profile Classification	Decision Tree	70.1%	0.501	~0.64
Project Profile Classification	XGBoost	94.2%	0.912	~0.93

Modeling Results Summary

- We evaluated Random Forest, Decision Tree, and XGBoost on three prediction tasks

Model_Performance_Summary

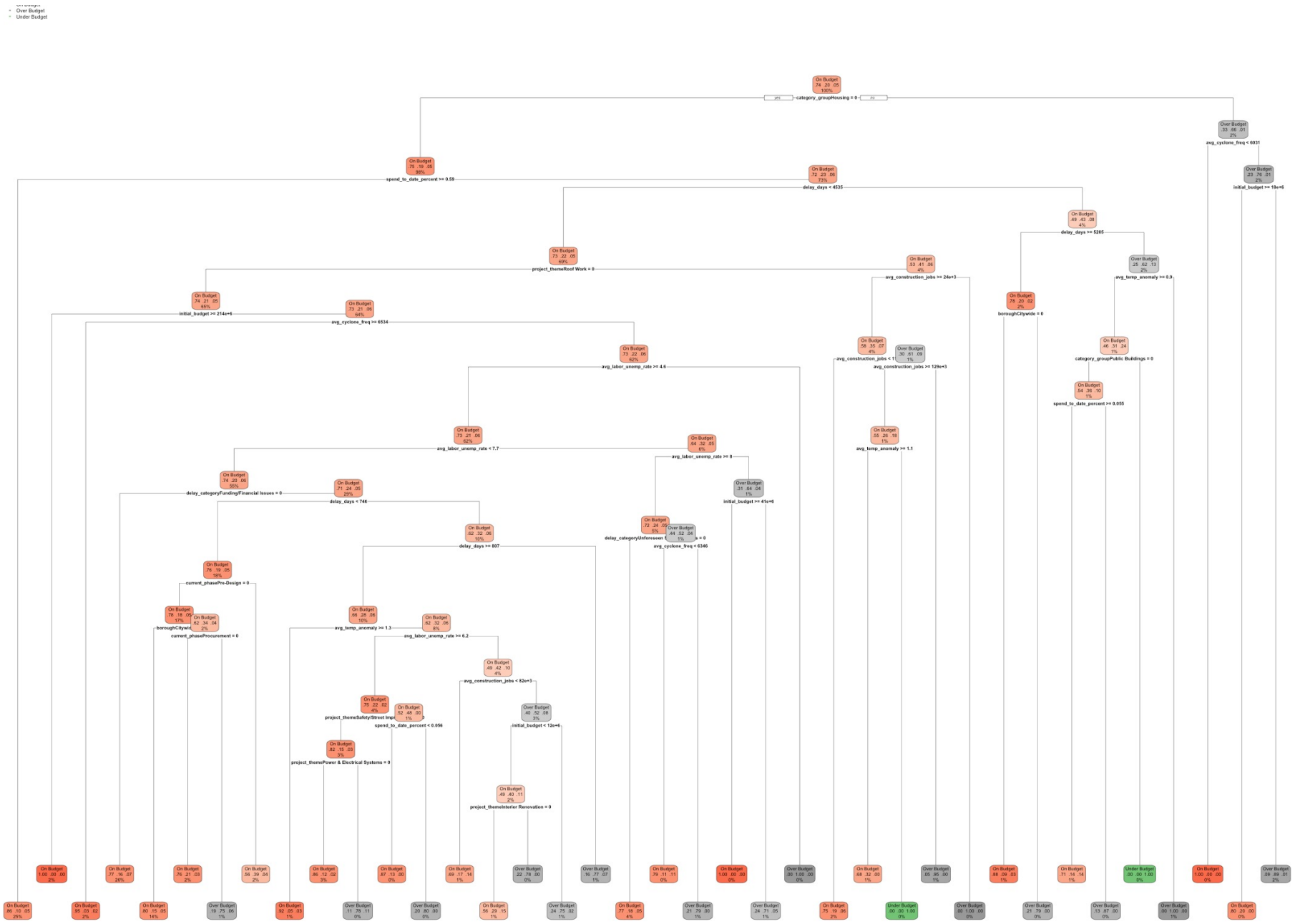
Prediction Task	Model	Accuracy	Kappa	Balanced Accuracy (Avg)
Cost Classification	Random Forest	92.7%	0.802	~0.88
Cost Classification	Decision Tree	78.6%	0.299	~0.59
Cost Classification	XGBoost	94.1%	0.846	~0.90
Delay Classification	Random Forest	98.4%	0.964	~0.98
Delay Classification	Decision Tree	91.4%	0.800	~0.91
Delay Classification	XGBoost	98.7%	0.970	~0.99
Project Profile Classification	Random Forest	91.3%	0.867	~0.91
Project Profile Classification	Decision Tree	70.1%	0.501	~0.64
Project Profile Classification	XGBoost	94.2%	0.912	~0.93

Modeling Results Summary

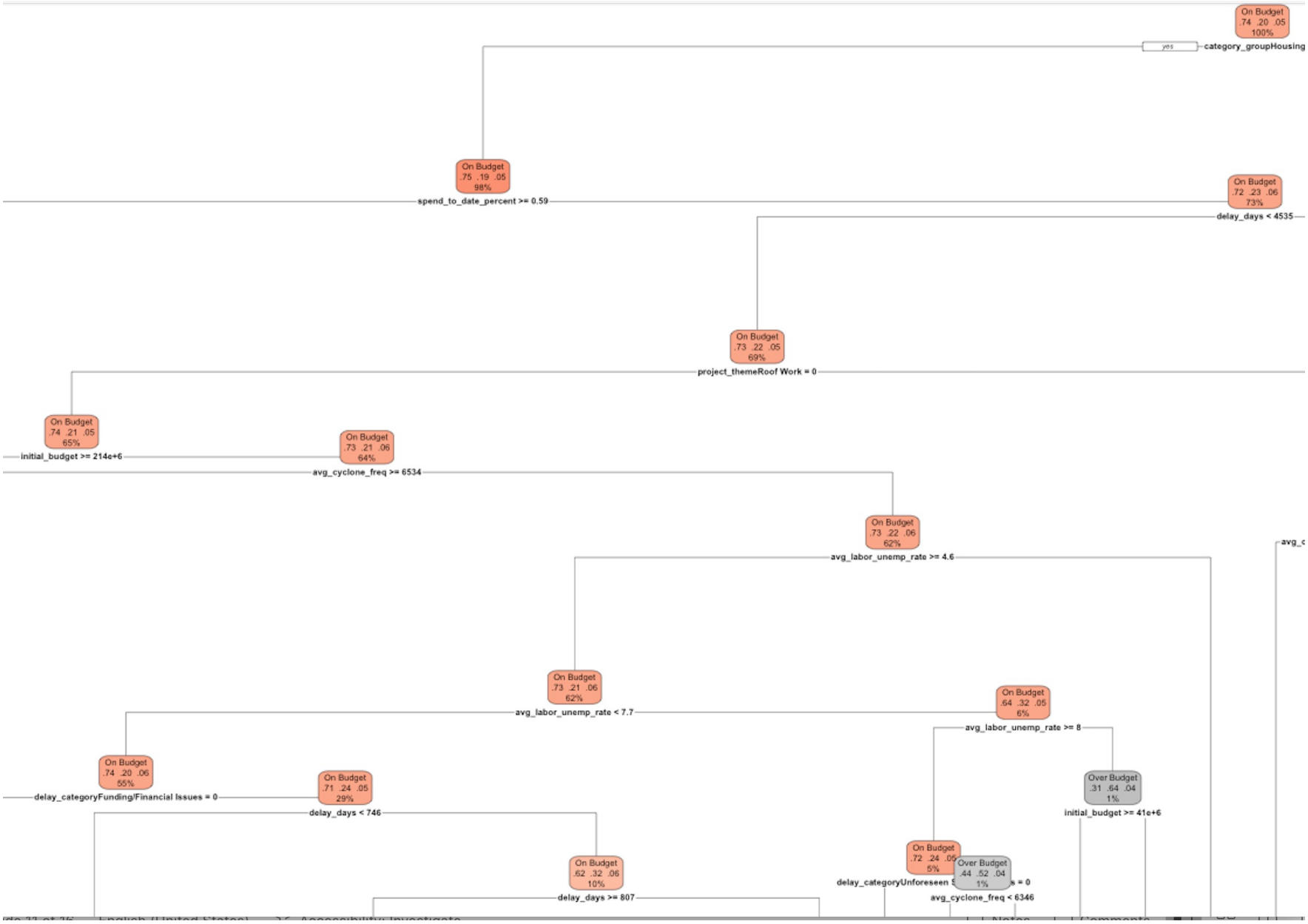
- Variables like budget size, construction jobs, unemployment, temperature anomaly, and wind speed had strongest influence

20 Most important Variables	Cost Risk	Exceptional Performance	Fast but Costly	High Risk	Lean Delivery	On Track	Schedule Risk	Strong Delivery	Time Risk, Cost-Saving
spend_to_date_percent	61.10	23.96	30.88	84.00	35.95	66.49	100.00	41.98	61.51
initial_budget	82.61	33.48	38.12	93.01	56.68	78.06	94.80	60.05	77.21
avg_temp_anomaly	61.97	24.55	36.59	77.49	49.76	65.95	68.23	44.98	71.30
avg_max_wind_speed	68.82	29.55	41.60	76.75	54.77	70.36	69.29	51.11	66.71
avg_labor_unemp_rate	58.47	23.39	38.04	74.72	45.30	59.77	69.00	41.99	68.95
avg_construction_jobs	53.98	22.89	28.96	68.41	44.03	52.14	65.43	36.95	69.69
avg_cyclone_freq	57.36	26.73	34.93	66.10	50.47	60.30	65.57	45.27	62.76
project_themeRoof Work	29.43	20.75	24.83	63.19	16.18	32.38	43.58	13.83	45.19
project_themeInterior Renovation	26.34	22.46	28.47	47.19	33.71	41.74	55.56	30.97	40.33
category_groupHousing	32.65	8.48	9.10	55.12	12.45	23.69	34.96	9.45	19.66
project_themeSafety/Street Improvements	38.60	13.30	17.74	50.28	25.81	40.96	48.94	25.31	34.29
current_phaseConstruction	35.83	15.69	23.59	50.02	17.47	42.46	47.74	25.32	37.00
boroughManhattan	35.65	12.37	25.23	49.85	26.62	40.21	41.69	19.80	31.74
category_groupParks & Recreation	38.07	20.42	25.85	44.63	31.67	44.29	48.71	30.20	41.15
boroughBrooklyn	36.38	17.84	16.22	48.07	21.29	42.79	43.07	25.50	36.26
delay_categoryFunding/Financial Issues	43.75	22.44	30.19	48.00	33.20	45.11	46.22	37.46	42.38
boroughQueens	29.18	18.52	19.76	43.47	28.17	38.73	47.87	23.61	39.35
category_groupPublic Buildings	36.00	26.84	17.10	47.77	30.20	39.93	45.72	19.10	46.24
category_groupTransportation	31.80	10.52	21.24	39.73	22.63	39.65	47.66	30.36	27.51
category_groupMiscellaneous / Other	32.47	10.87	23.26	44.05	23.96	45.34	47.54	16.92	36.60

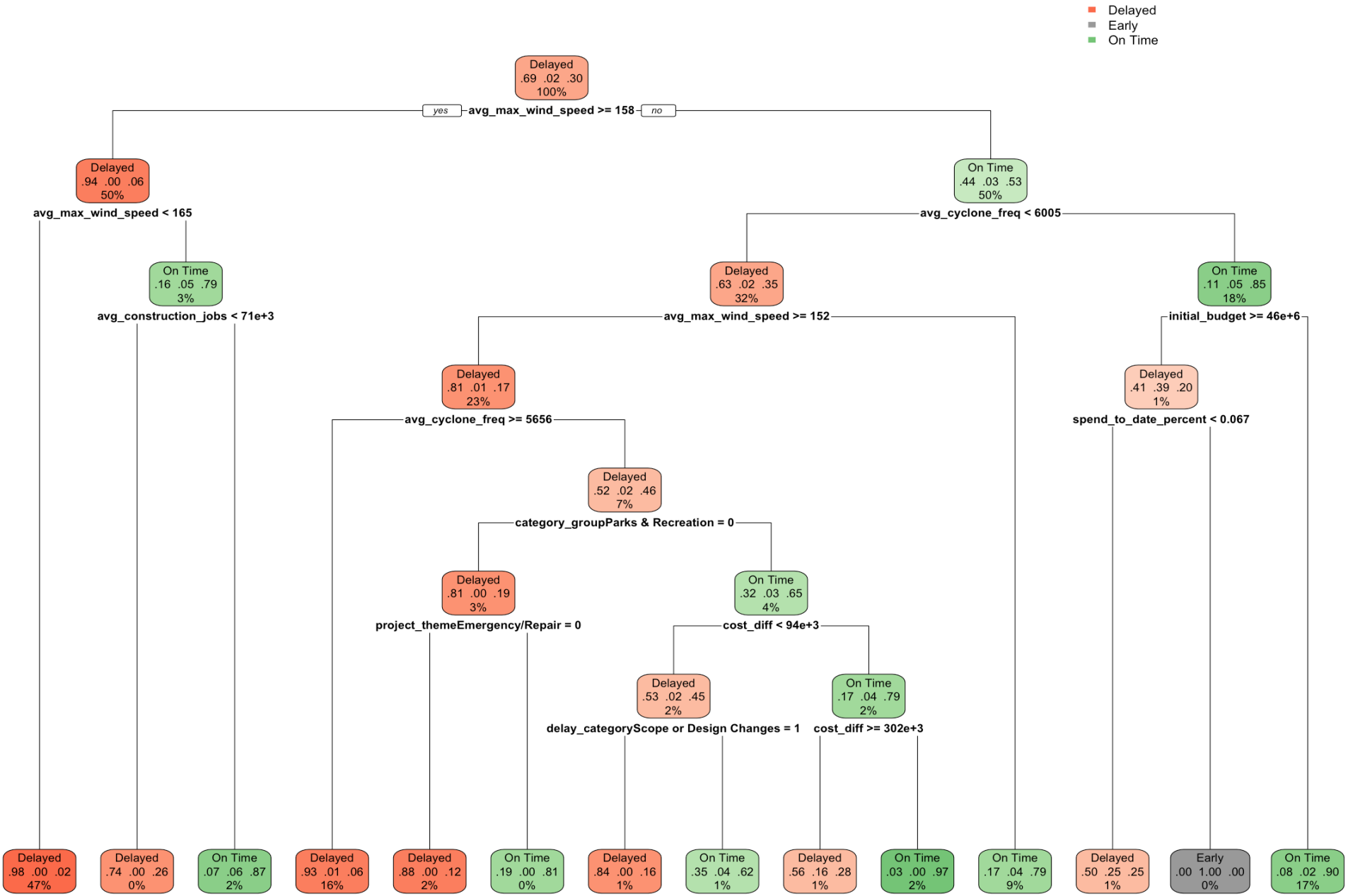
Modeling Results Summary



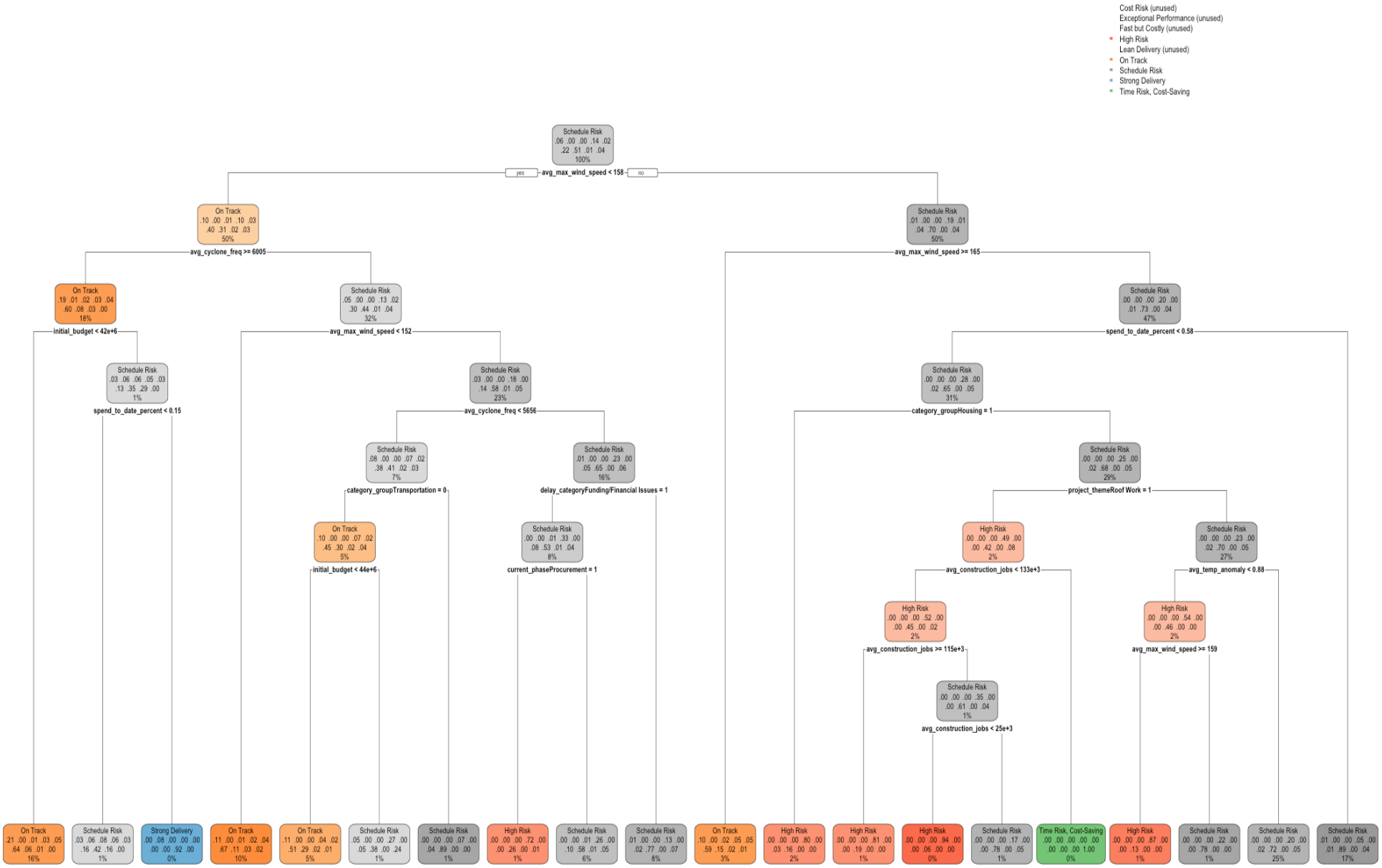
Modeling Results Summary



Modeling Results Summary



Modeling Results Summary



Summary

- XGBoost consistently achieved the highest accuracy and kappa scores
- Project profile classification offers a powerful way to summarize and flag risks
- High wind, labor shortages, and high initial budget are common indicators of late or overbudget delivery

Why Were Our Models So Accurate?

✅ **Structured Data**: Key predictors like budget, spend %, and delay days were well-defined and cleanly separated classes.

🔍 **Strong Signal**: Features such as spend_to_date_percent and initial_budget showed high predictive power.

🌲 **Model Choice**: Tree-based models (Random Forest, XGBoost) effectively captured nonlinear patterns.

📊 **Robust Evaluation**: Used 5-fold cross-validation to ensure reliable performance.

⚠️ **Limitation**: Lacked access to complex project factors — e.g., mscope changes, site complexity, inter-agency dynamics.

📉 **Implication**: High accuracy reflects data structure, not necessarily full real-world complexity.

Conclusion and What's Next

🚩 **Proactively flag high-risk projects** based on learned patterns from budget, timeline, and weather-labor dynamics

🧮 **Support auditors and capital planners** with data-driven insights, not just retrospective reviews

📊 **Inform performance dashboards** to increase transparency for agencies and the public

🛠️ **What's Next:**

- Add unsupervised learning to discover hidden project clusters
- Incorporate **contractor/vendor history** into the risk models
- Expand to other cities or federal datasets
- Integrate into capital project dashboards