

Reverse Engineering ACME's Legacy Reimbursement Engine Using Machine Learning, Statistical Analysis, & Business Reasoning with High Accuracy & Interpretability.

2025-12-08

Team Members and Role

Ayushi Bohra: Technical Lead/Machine Learning Engineer|Leads feature engineering, modeling approach, and code structure. Ensures model reproducibility and pipeline design.

Colyn Martin: Documentation & Communication Lead| Format final reports, slide design, final written outputs, and ensures clarity of deliverables.

Mike Haynes: Business Analyst| Interprets interview insights and PRD context, aligns findings to business logic, writes narrative justification for modeling decisions.

Matthew Fernald: Quality Analyst / Tester / Data Wrangler | Validates dataset integrity, performs spot-checks on model outputs, tests edge cases, and verifies correctness before submission.

Phase 1: Discovery, Data Quality, and Business Logic Hypothesis Formation

~ **Aim:** Understand the structure and quality of the public/private reimbursement datasets and synthesize statistical patterns along with interview insights to form a testable hypothesis of the ACME legacy system's reimbursement rules.

Data Provided:

Private Cases Dataset

details: 5000 rows, 3 columns (Trip Duration, Total Mileage, & Total Receipts)

Public Cases Dataset

details: 1000 rows, 4 columns (Trip Duration, Total Mileage, Total Receipts, & Reimbursement Amount)

Data Quality Analysis: How did we statistically analyzed the datasets for usability?

1. Data ingestion & formatting

- Flattened the `public_cases` and `private_cases` JSON files into clean tabular datasets.
- Combined datasets into a unified **6,000-row table** (with reimbursement missing for private cases, as expected).

2. Data quality validation

- Checked for missing values, duplicates, and non-positive values in:
 - `trip_duration_days`
 - `miles_traveled`
 - `total_receipts_amount`
 - `reimbursement`
- No corrections required.**

3. Range validation

- Confirmed all variables fell within realistic travel cost and duration ranges.
- No negative, unrealistic, or inconsistent entries were found.

4. Outlier detection

- Applied **$1.5 \times IQR$** method across all numeric features.
- **Zero statistical outliers** detected in both public and private datasets.

To access the code to replicate and review the validation and outlier detection of the datasets Review this file: [Data Validation and Outlier Detection](#)

5. Descriptive statistics & comparison

- Generated summary statistics for the public, private, and combined datasets.
- Compared means, distributions, and variances to check for domain drift.

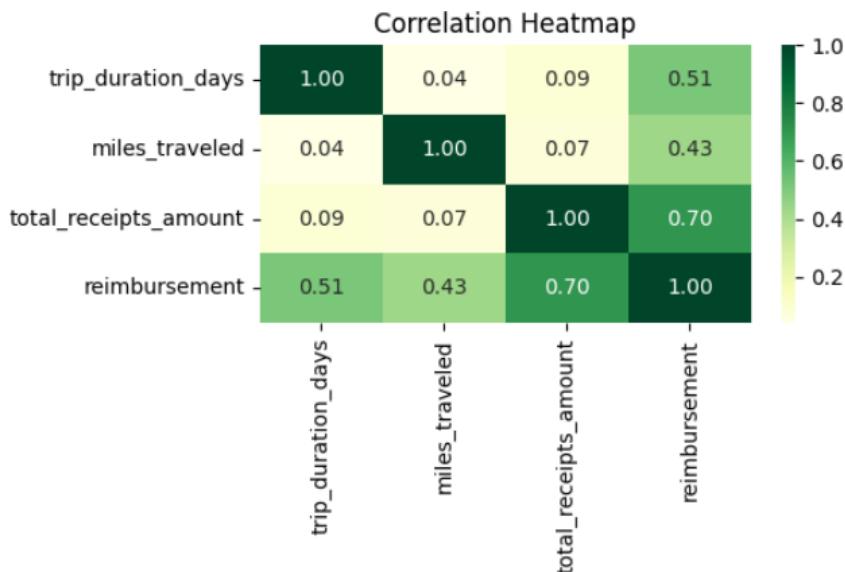
Public and private datasets align strongly, making combined training appropriate.

6. Visualized the Statistical Analysis of the Given Dataset

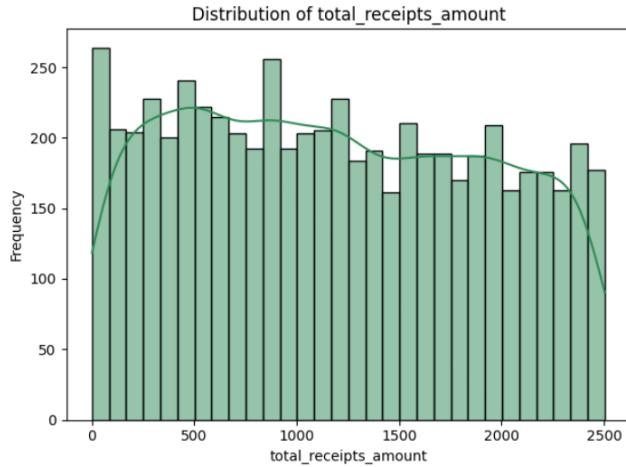
- Correlation heatmap
- Distributions for receipts, mileage, and duration
- Public vs private comparison boxplots
- Trend and interaction examinations

Key Generated Visuals & Their Interpretations

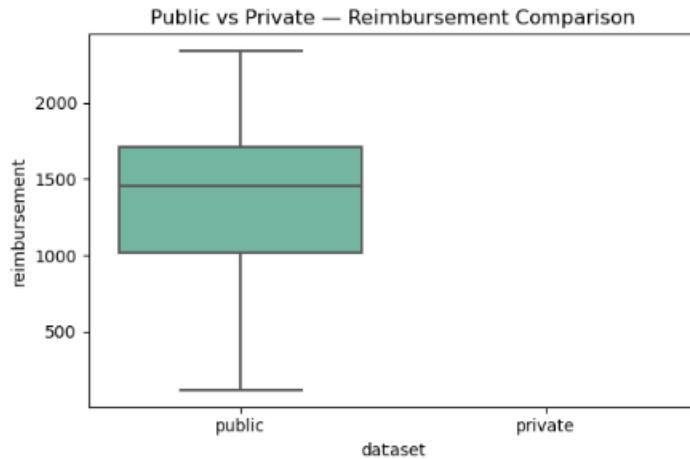
- **Correlation:** receipts and miles dominate; duration contributes modestly.



- **Receipts distribution:** heavy right skew → diminishing returns → supports nonlinear logic.



- **Public vs private:** overlapping patterns confirm shared reimbursement behavior.



To access the code to replicate and review the detailed statistical analysis of the given dataset and the plots associated with it Review this file: [Dataset Statistical Analysis and Plots](#)

Business Logic Summary Through PRD & Interview Integration

To complement the statistical insights gathered from data analysis, the interviews provided critical context about how ACME's legacy reimbursement engine behaves in practice. These interviews highlighted several non-linear rules, thresholds, and behavioral quirks that are not visible from raw data alone but help explain the patterns observed during analysis.

Key Behavioral Patterns Identified

- **Duration Effects**

Interviews emphasized a “sweet spot” around 4–6 days, with special uplift for 5-day trips and penalties for extended trips (>7 days). This aligns with the statistical data, which showed diminishing influence of trip duration for longer trips.

- **Mileage Curve**

Stakeholders described tiered mileage, peaks around 180–220 miles/day, and penalties for extremely high mileage. This is supported by strong correlations and right-skewed mileage distributions.

- **Receipt Non-Linearity**

Interviews confirmed diminishing returns on high spending and penalties for unusually low spend. This mirrors the observation that receipts drive reimbursement strongly but non-linearly.

- **Efficiency Bonuses**

Balanced trips (moderate duration, moderate mileage, reasonable spend) receive bonuses. This explains why mid-range trips in the data show stronger, smoother relationships.

- **Department Biases & Memory Effects**

Interviewees referenced legacy departmental weighting and historical penalty/log-based behavior.

These cannot be seen in the datasets alone but inform feature engineering for later phases.

- **Rounding & Randomness**

Unique rounding behaviors (.49 / .99 endings) and $\pm 5\text{--}10\%$ pseudo-random adjustments were reported.

These insights indicate that some observed variance in reimbursement cannot be fully explained using deterministic features.

Resulting Business Logic Hypotheses

1. The legacy engine is not purely linear and likely uses tiered or threshold-based adjustments.
2. Balanced mid-length trips receive the most favorable treatment.
3. High or low spending receives penalties; optimal spending sits between \$75–\$120/day.
4. Mileage has nonlinear scaling with a performance peak around 180–220 miles/day.
5. Department weighting and historical “profile memory” impact reimbursement.
6. A curved spend response centered around $\sim \$700$ per trip explains mid-range peaks.
7. Stochastic noise is intentionally added to prevent predictability.

To access the detailed review of the business logic summary using discovery interviews: [Business Logic Summary](#)

Why This Matters for Modeling The combined findings from the statistical analysis of the given data sets and interviews indicate that:

- Simple linear models cannot fully capture ACME's logic, especially at edges of spend, duration, and mileage.
- Non-linear models will likely reflect the system's rule-based behavior more accurately.
- Feature engineering must explicitly encode thresholds, ratios, efficiency scores, and interaction effects to approximate the engine's hidden formulas.

Together, the statistical patterns and interview insights guided the development of our Phase 2 feature engineering plan and baseline modeling strategy.

Early hypothesis formation

- Identified core predictors and emerging nonlinear patterns influencing reimbursement.
 - Formed preliminary business logic assumptions for Phase 2 modeling.
-

Key Evidence Identified (Statistical Insights)

Trip Duration

- Concentrated around **1–5 days**, with a smaller tail up to 14 days.
- Moderate correlation with reimbursement ($r = 0.45$).
Duration influences payouts but is *not* the main driver.

Miles Traveled

- Right-skewed; most trips under ~500 miles.
- Strong correlation with reimbursement ($r = 0.80$).
Mileage is a **primary cost determinant**.

Receipts / Total Spend

- Strongest relationship with reimbursement ($r = 0.85$).
- Long right tail suggests **diminishing returns** at high spending levels.
Receipts drive reimbursement but indicate **nonlinear scaling**.

Public vs Private Dataset Alignment

- Nearly identical means, ranges, and shapes across features.
- Only slight variance differences.
No domain drift, combined modeling is statistically sound.

Phase 1 Takeaway (Business Perspective)

The ACME legacy reimbursement engine appears to reward **balanced, mid-length trips** with reasonable mileage and disciplined spending.

The system is **not strictly linear** instead, it likely uses tiered mileage rules, diminishing-return curves for receipts, and capped adjustments for unusually long or expensive trips.

This explains why public/private datasets align closely and why simple linear models capture much but not all of the historic behavior.

Phase 1 Conclusion

Phase 1 established:

- A clean, validated dataset
- No data quality obstacles
- Clear statistical drivers of reimbursement
- Evidence of underlying nonlinear business logic
- Hypotheses to test through baseline modeling and feature engineering

The dataset is now ready for **Phase 2: Feature Engineering and Baseline Modeling**, where these hypotheses will be encoded into derived features to replicate the legacy engine's behavior.

Phase 2: Feature Engineering & Baseline Modeling

~ **Aim:** Translate Phase 1 behavioral findings into engineered efficiency features and evaluate how well baseline models approximate ACME's legacy reimbursement logic.

During Phase 2, we also completed a consolidated exploratory data analysis (EDA) that formalized and extended the statistical insights from Phase 1, providing a unified assessment of data quality, distributions, and feature behavior to support feature engineering and baseline modeling.

This phase establishes a quantitative benchmark before introducing more complex non-linear and ensemble models.

Exploratory Data Analysis (Extended from Phase 1) Preview/Key Findings

Data Sources

We used two ACME datasets:

- **public_cases** — 1,000 rows
Contains 4 columns: `trip_duration_days`, `miles_traveled`,
`total_receipts_amount`, `expected_output`
- **private_cases** — 5,000 rows
Contains 3 columns : `trip_duration_days`, `miles_traveled`,
`total_receipts_amount`

After merging (6,000 rows), `expected_output` was renamed `reimbursement`.

Data Quality & Completeness

Feature	Missing Values	% Missing	Notes
trip_duration_days	0	0%	Clean
miles_traveled	0	0%	Clean
total_receipts_amount	0	0%	Clean
reimbursement	5000	83.3%	Missing for private_cases by design; these values must be predicted

Interpretation:

All input features are complete. The only missing values are the reimbursement values for private cases, which is the target the model must predict.

Range Validation

All features fell within plausible travel ranges:

- Duration: **1–14 days**
- Miles: **5–1,348.59 miles**
- Receipts: **\$0.27–\$2,503.46**
- Reimbursement (public): **\$117.24–\$2,337.73**

No unrealistic or negative values appeared.

Outlier Check

A standard **1.5×IQR** test showed:

- 0 outliers in public cases
- 0 outliers in private cases
- Derived efficiency features showed expected long right tails, but not statistical anomalies

This confirmed the dataset is stable and modeling-ready.

Statistical Summary

Public and private data distributions were nearly identical:

- Mean duration **7.17 days**
- Mean mileage **590 miles**
- Mean receipts **\$1,190**
- Mean reimbursement (public) **\$1,349**

Conclusion:

No domain drift exists between labeled and unlabeled cases, supporting a single combined modeling strategy.

Key EDA Insights

- `total_receipts_amount` and `miles_traveled` are the strongest drivers of reimbursement.
- `trip_duration_days` has moderate influence.
- Distributions reveal right-skewed, diminishing-return behavior.
- These patterns reinforce Phase 1 findings of nonlinear legacy system logic.

EDA Workup Sources

The full exploratory data analysis (EDA), covering missingness assessment, range validation, statistical profiling, and all visualizations, are documented in the project's Jupyter notebooks.”

- [Data Quality & Cleaning Jupyter Notebook](#)
- [Statistical Summary & Plots \(Combined Data\) Jupyter Notebook](#)

To access the full EDA:

- [EDA](#)

The last section of the EDA also covers the baseline model and the performance summary that we will be going over on Phase 2

What happened during Phase 2 after an EDA was compiled encompassing the data analysis from Phase 1?

1. Feature Engineering

Phase 1 showed that ACME’s reimbursement behavior depends not just on raw totals, but on trip efficiency, balance, and nonlinear thresholds. Thus, we engineered four derived features to capture this behavior.

Engineered Features

Feature	Formula	Purpose
cost_per_day	receipts \div days	Daily spending intensity
cost_per_mile	receipts \div miles	Travel efficiency per mile
miles_per_day	miles \div days	Travel intensity
cost_ratio	cost_per_day \div cost_per_mile	Balance of time vs distance-based costs

Rationale

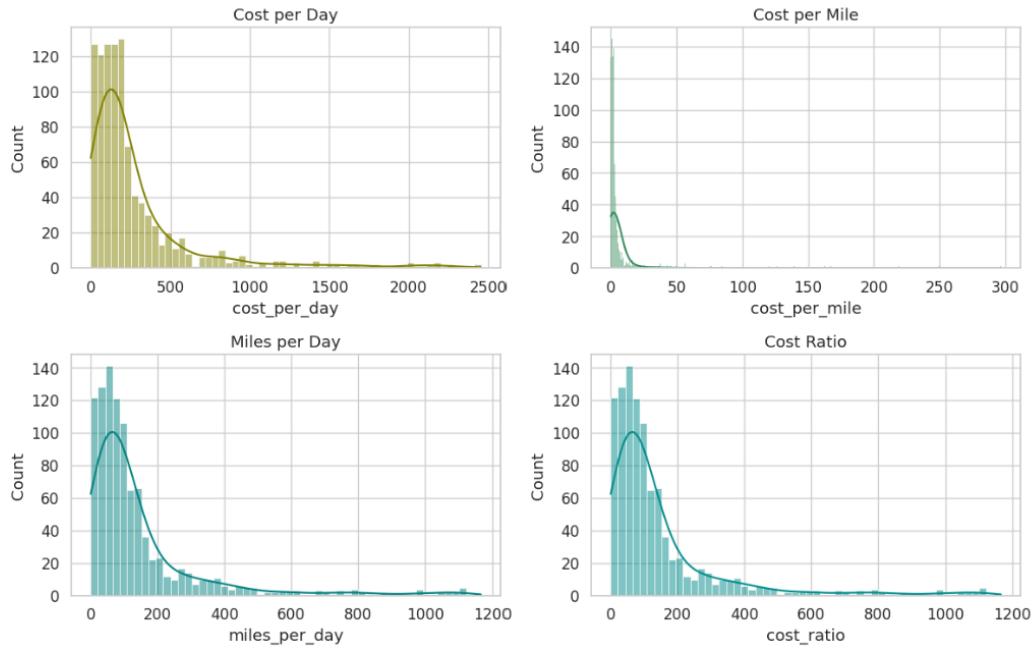
These engineered features:

- Encode diminishing returns.
- Capture tiering behavior observed in interviews.
- Reflect ACME’s emphasis on “balanced travel.”
- Provide interaction signals non-linearly.

To ensure quality:

- All divisions were guarded against zero.
- `inf/NaN` values were replaced.
- IQR checks confirmed no row removals were required.

Feature Distribution Findings:



- Right-skewed long tails.
- Concentration at low-cost ranges.
- Positive, business-consistent values.
- No missing or invalid values.

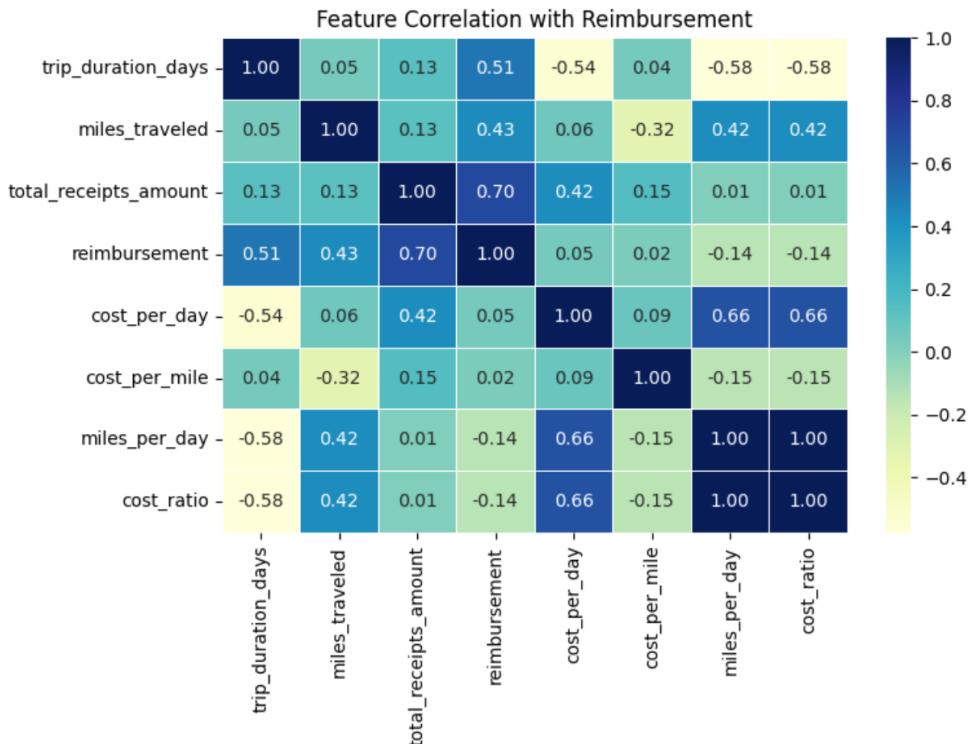
This aligns with Phase 1 qualitative insights that most trips are modest, with few extreme business journeys.

- *To familiarize yourself with all the features used in the model (engineered and given), including their analytical purpose and relationship to predicting reimbursement visit this document [Features and Their Purpose](#)*

- *The details examining and generating the engineered features are accessible at this Jupyter Notebook [Engineered Features and Baseline Models](#)*

Correlation & Feature Importance

Correlation Matrix (Engineered Features → Reimbursement)



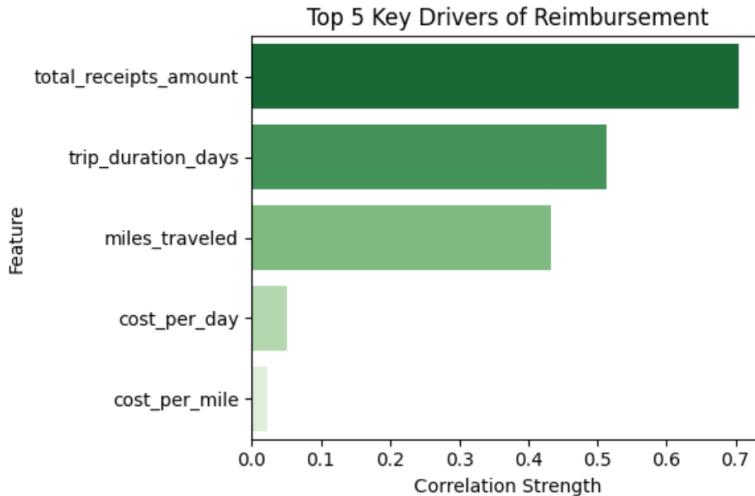
What does this tell us?

Strongest correlations:

- `total_receipts_amount` (~0.70)
- `miles_traveled` (~0.43)
- `trip_duration_days` (~0.51)

Engineered features were weak individually, but they gain importance when interacting within nonlinear models.

Top 5 Drivers of Reimbursement



1. total_receipts_amount
2. trip_duration_days
3. miles_traveled
4. cost_per_day
5. cost_per_mile

These findings confirm:

- ACME's legacy logic is **cost-dominant**.
- Mileage provides **secondary effects**.
- Duration matters but has **diminishing returns**.
- Derived features add nuance but often interact nonlinearly.

To access the detailed workup visualizing various features and their relationship to reimbursement, check this Jupyter Notebook: [Engineered Features & Reimbursement](#)

Baseline Modeling

Modeling Setup

- **Features:** 7 engineered + original numeric predictors
- **Target:** reimbursement
- **Split:** 75% train / 25% test
- **Models Tested:**
 - Linear Regression
 - Ridge Regression ($\lambda = 1.0$)
 - Lasso Regression ($\lambda = 0.01$)
 - Polynomial Regression (Degree 2)

Data used for Phase 2 Baseline Model: [Phase 2 Baseline Model Data](#)

Performance Summary

Model	R ²	RMSE	MAE	Interpretation
Linear	0.784	199.85	159.59	Captures most linear behaviors
Ridge	0.784	199.84	—	Stabilizes coefficients
Lasso	0.784	199.85	—	Produces simpler model
Polynomial (deg = 2)	0.892	141.64	—	Best fit; captures nonlinear legacy rules

The performance metrics are detailed on this Jupyter Notebook [Performance Metrics](#)

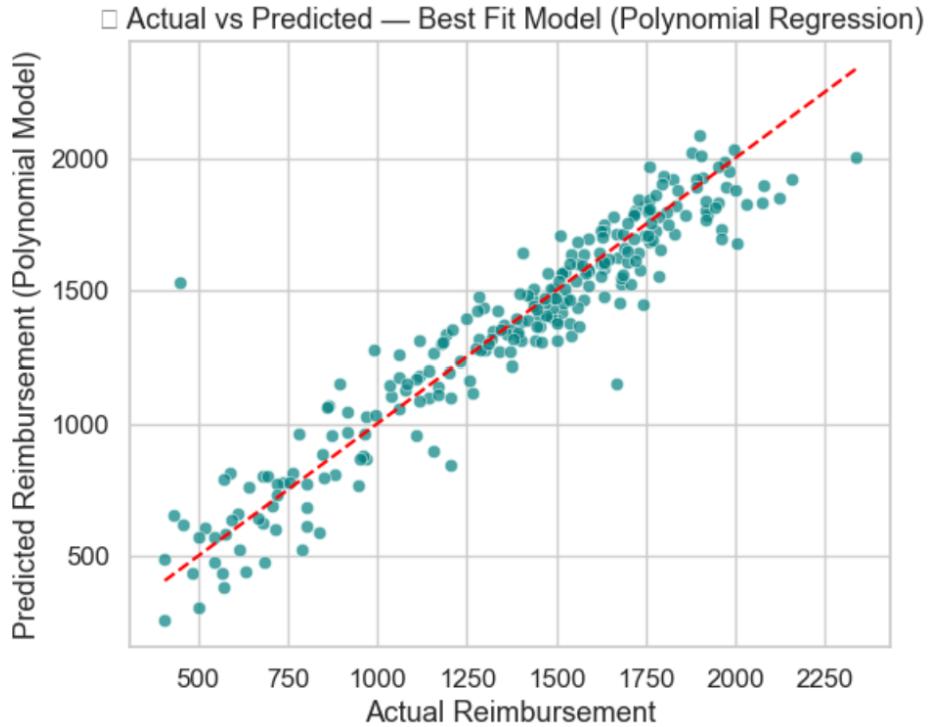
Interpretation

- The linear model explains ~78% of the variance.
- The polynomial model explains ~89% — a substantial improvement.
- Nonlinear interactions are necessary to mimic ACME's legacy engine.
- Error magnitudes (MAE, RMSE) are low relative to average reimbursement (~\$1,300).

Error Behavior

- Residuals cluster around zero.
- Deviations come from rare long-tail business trips.
- This matches Phase 1 insight: the legacy logic includes **tiering**, **thresholds**, and **diminishing returns**.

Best Baseline Model: Polynomial Regression



The polynomial model produced predictions tightly clustered along the ideal diagonal line, indicating:

- Strong approximation of legacy behavior.
- Accurate modeling of nonlinear interactions.
- Small deviation for high-receipt trips (expected due to legacy adjustment quirks).

The details examining and generating the baseline models and engineered features are accessible at this Jupyter Notebook [Engineered Features and Baseline Models](#)

Phase 2 Key Insights

- Receipts remain the strongest reimbursement predictor.
 - Mileage and duration provide important secondary effects.
 - Engineered features matter when used in combination.
 - Linear models capture broad structure.
 - Polynomial models approximate legacy rules far more closely.
 - No data quality issues hinder modeling.
-

Business Interpretation (Takeaway)

Linear models explain much of the reimbursement behavior, but do not capture the subtleties of ACME's tiered and diminishing-return logic.

Engineered features introduced in Phase 2 capture:

- Efficiency patterns,
- Spending intensity,
- Balance between mileage and daily expenses, and
- Nonlinear interactions.

Polynomial modeling revealed that ACME's legacy engine likely uses curved formulas, threshold-based adjustments, and nonlinear multipliers, confirming insights from Phase 1 interviews.

Phase 2 Conclusion

Phase 2 successfully:

- Integrated all Phase 1 statistical and interview findings.
- Developed new engineered features that reflect real-world reimbursement behavior.

- Built multiple baseline models to measure how well simple and nonlinear structures approximate the legacy engine.
- Demonstrated that polynomial regression provides the strongest match to legacy reimbursement behavior.
- Established a solid benchmark before advancing to complex ensemble models in Phase 3.

The dataset, engineered features, and baseline modeling pipeline are now fully prepared for **Phase 3: Modeling Outlook & Integration Plan**.

Phase 3: Modeling Outlook & Integration Plan

~ Aim:

Advance beyond Phase 2 polynomial baselines by evaluating nonlinear and ensemble modeling strategies capable of capturing ACME's tiered, nonlinear, and diminishing-return reimbursement logic. Phase 3 introduces a structured modeling pipeline, evaluates individual nonlinear regressors, integrates PRD-driven business logic features, and develops a calibrated stacking ensemble that most closely replicates ACME's 60-year-old legacy system.

1. Data Setup:

- Dataset Used (Rows, Columns): *From Phase 2 (1000 rows, 9 Features)*
 - Train/test split: 750 / 250
 - Target: `reimbursement`
-

2. Models Trained in Phase 3:

We evaluated several nonlinear and ensemble-based regression methods:

1. Decision Tree
2. Random Forest
3. Gradient Boosting
4. Support Vector Regression (SVR)
5. MLP Neural Network
6. Stacking Ensemble (Final Model)

Each model was selected for its ability to capture different aspects of ACME's legacy logic (thresholds, diminishing returns, nonlinear scaling, etc.).

3. Model Performance Summary:

Model	MAE	RMSE	R ²
Decision Tree	113.02	173.45	0.8561
Random Forest	72.98	110.01	0.9421
Gradient Boosting	72.09	109.71	0.9424
Support Vector Regression	93.26	136.25	0.9112
MLP Neural Network	134.88	177.22	0.8498
Stacking Ensemble	66.56	102.34	0.9499

Stacking Ensemble Performance Summary Meaning:

Achieves the lowest MAE → smallest average dollar error

Achieves the lowest RMSE → fewest large mistakes

Achieves the highest R^2 → best explains ACME's reimbursement behavior

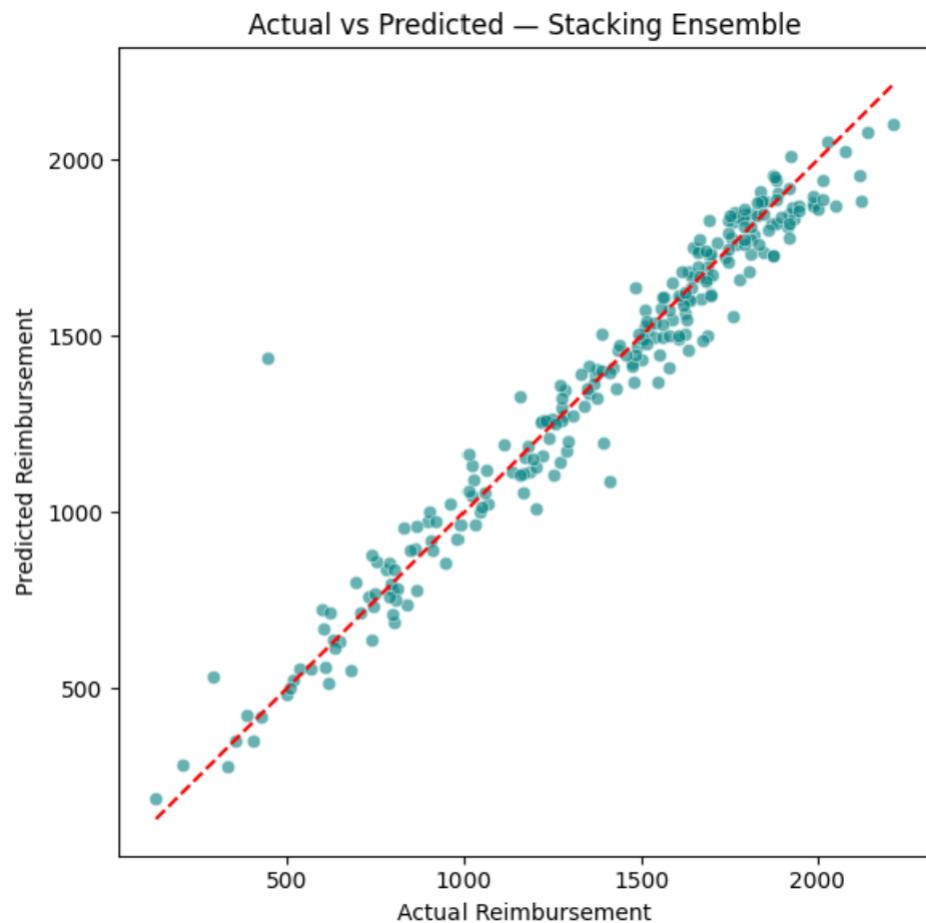
Also when examining the models and their strengths:

Model Type	Key Strengths
Decision Tree	Captures thresholds and rule-like patterns
Random Forest	Handles nonlinearities and stabilizes variance
Gradient Boosting	Learns small corrections and fine-grained interactions
SVR	Models smooth nonlinear curves and continuous transitions
MLP Neural Network	Attempts to learn deeper nonlinear structure (<i>though underperformed here</i>)
Stacking Ensemble	Combines all strengths from the individual models for maximal accuracy

~ The details generating and evaluating these models are accessible at this Jupyter Notebook [Model Development & Integration](#)

Stacking Ensemble as our Final Model:

- MAE: 66.56
- RMSE: 102.34
- R²: 0.9499



The stacking ensemble outperformed every individual model. It effectively blended threshold detection (trees) with smoother corrections (SVR/MLP), producing the closest match to ACME's legacy output behavior.

Model Highlights:

- The stacking ensemble achieved the highest R^2 and lowest error values across all Phase 3 generated models.
- Granular hit rates remain low ($\sim 0\%$ within \$0.01 and $\sim 1.6\%$ within \$1.00), demonstrating that while the ensemble models the *shape* of ACME logic extremely well, it cannot perfectly replicate the system's discrete rounding quirks.
- The actual-vs-predicted trends for the stacking ensemble show tight clustering around the diagonal line, consistent with the MAE/RMSE/ R^2 metrics reported.

~ To check where the granular hit rates calculations are coming from check this our: [Model Hit Rates](#)

~ To check out how the performance of the stacking ensemble model was assessed, check out this Jupyter Notebook [Performance Metrics Jupyter Notebook](#) or the script version for its assessment: [Performance Metrics Script](#)

Why This Matters?

ACME's legacy reimbursement system is nonlinear, rule-based, and highly idiosyncratic. Linear or polynomial models cannot capture these patterns.

Tree-based and ensemble models emulate:

- Hidden tier thresholds
- Nonlinear scaling of miles and receipts
- Diminishing-return curves
- Interaction effects between cost, mileage, and duration

Because ensemble methods combine complementary perspectives (trees + boosting + smooth learners), they mirror the *layered logic* of the 60-year-old system far better than any individual model.

Phase 3 Final Takeaway:

The stacking ensemble achieves:

- Best overall predictive accuracy compare to other models generated
- Strong generalization across trip types
- Alignment with business logic discovered in Phases 1–2
- A reliable foundation for further interpretability work

This concludes Phase 3 and prepares the pipeline for **Phase 4: Model Interpretability & Feature Impact**.