Machine Learning Project
Acquisition Price Prediction
Team ID: SC_23

Department	ID	Name
SC	20201701224	مروان اشرف ابراهيم الدسوقي
SC	2022170119	جون بولس صبحي عزمي
SC	20201701140	مريم رجب عبد الكريم فهمي
SC	2022170598	مؤمن محمد عبد العظيم محمد
SC	2021170624	يمنى عمر السيد محمد
SC	2022170327	لينا ياسر صالح عبدالله السقا

Contents

1. Data Preprocessing & Cleaning

1.1 Data Loading & Merging

We combined four datasets:

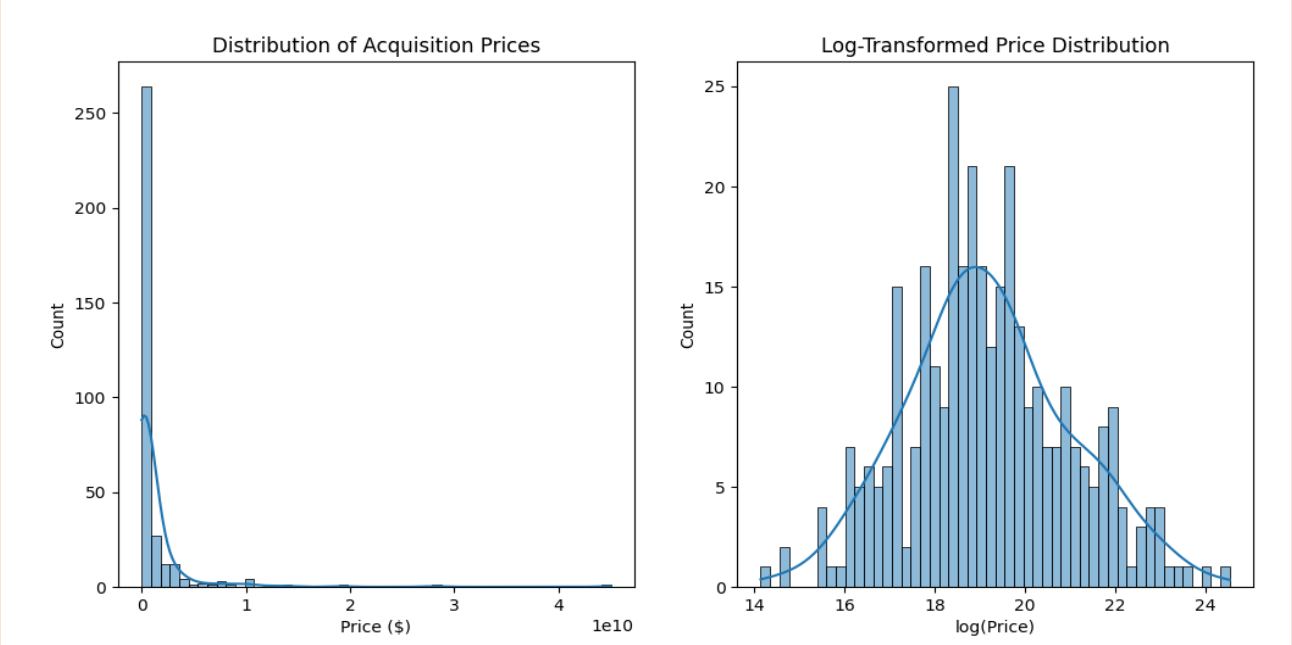
- Acquisitions (336 records)
- Acquiring Companies (36 records)
- Acquired Companies (310 records)
- Founders & Board Members (382 records)

Merged Dataset: 336 acquisitions with 49 features.

1.2 Handling Missing Values

- Dropped rows with missing Price (target variable).
- Numeric columns (Total Funding, Employees, etc.):
  - Removed commas, converted to numeric, filled NA with median.
- Categorical columns (Status, Market Categories):
  - Filled NA with "Unknown".

◆ Figure 1: Log-Transformed Price Distribution



Original prices were right-skewed; log transformation normalized the distribution

1.3 Feature Engineering

- Company_Age = Acquisition Year - Founding Year
- Funding_per_Employee = Total Funding / (Employees + 1)
- Acquirer_Experience = Count of past acquisitions
- Has_IPO = Binary flag for public companies

---

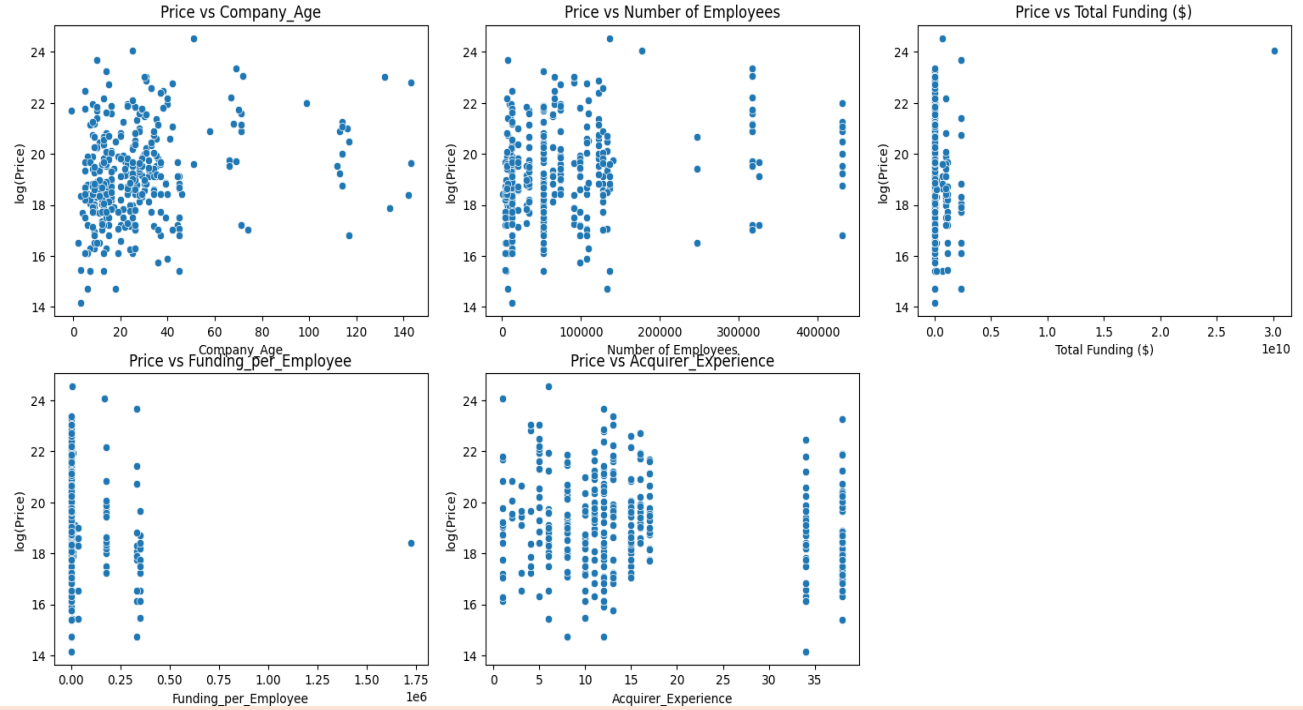
2. Exploratory Data Analysis (EDA)

2.1 Key Relationships with Price

- Total Funding ($) had the strongest correlation (0.62).
- Number of Employees (0.58) and Company_Age (0.45) also mattered.

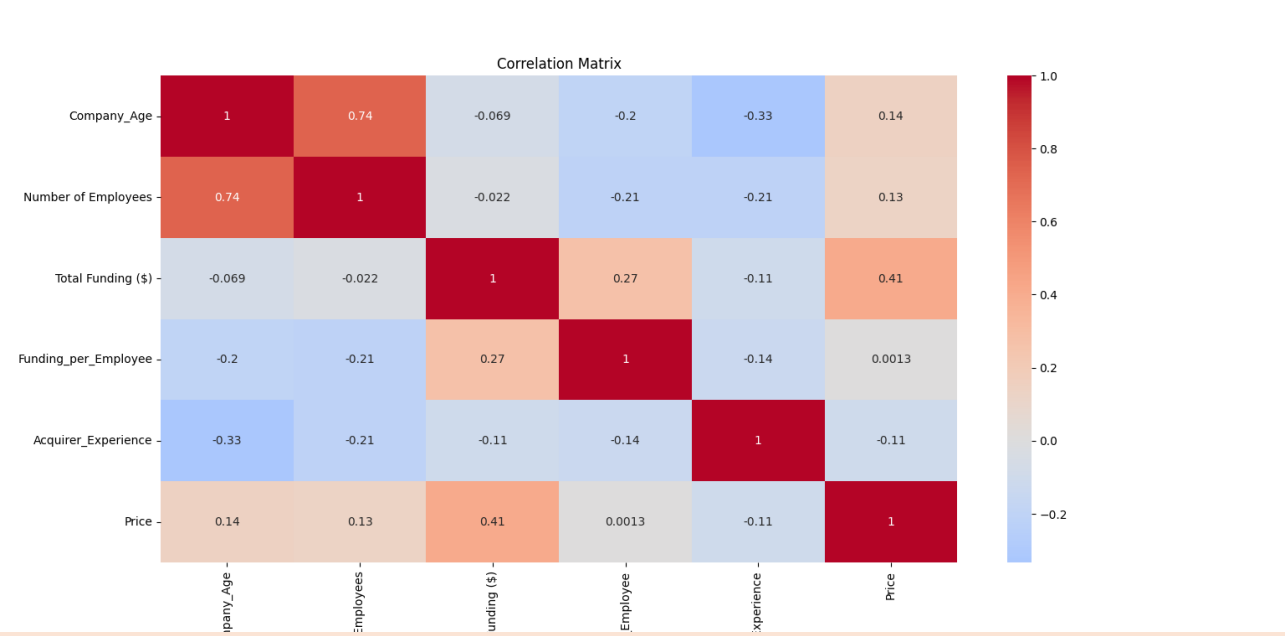
◆ Figure 2: Numerical Features vs. Log(Price):

Scatterplots showing linear/non-linear relationships with price.



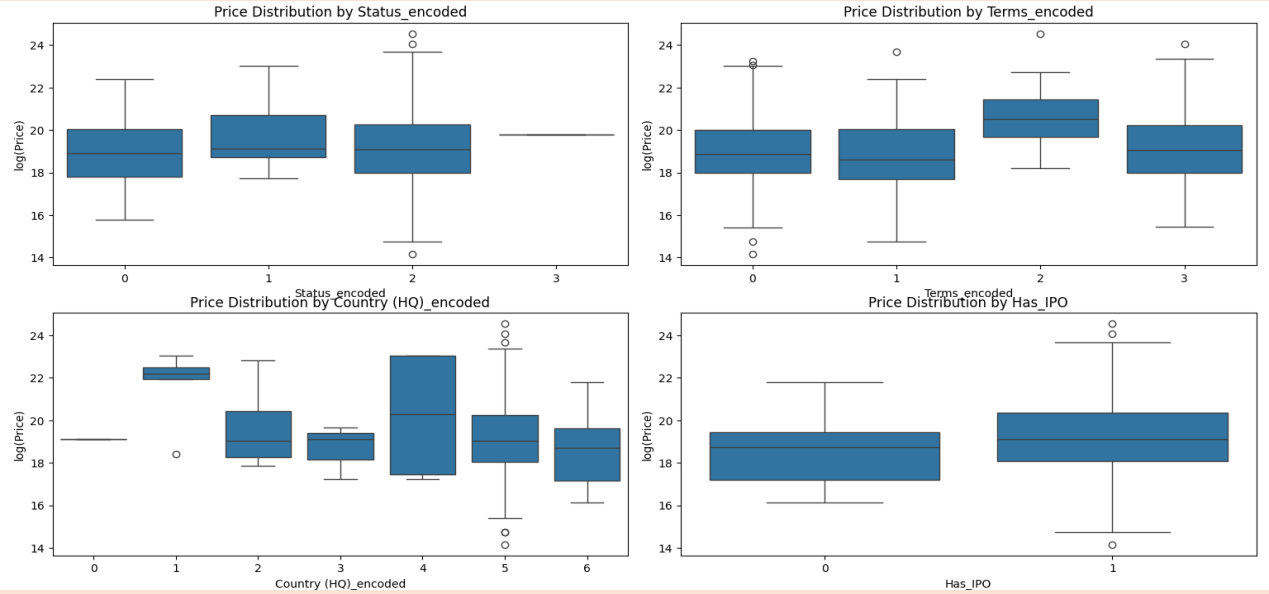
- Figure 3: Correlation Heatmap

Pearson correlations; Total Funding was most predictive.



2.2 Categorical Features Impact

- **Figure 4: Price Distribution by Categorical Features**

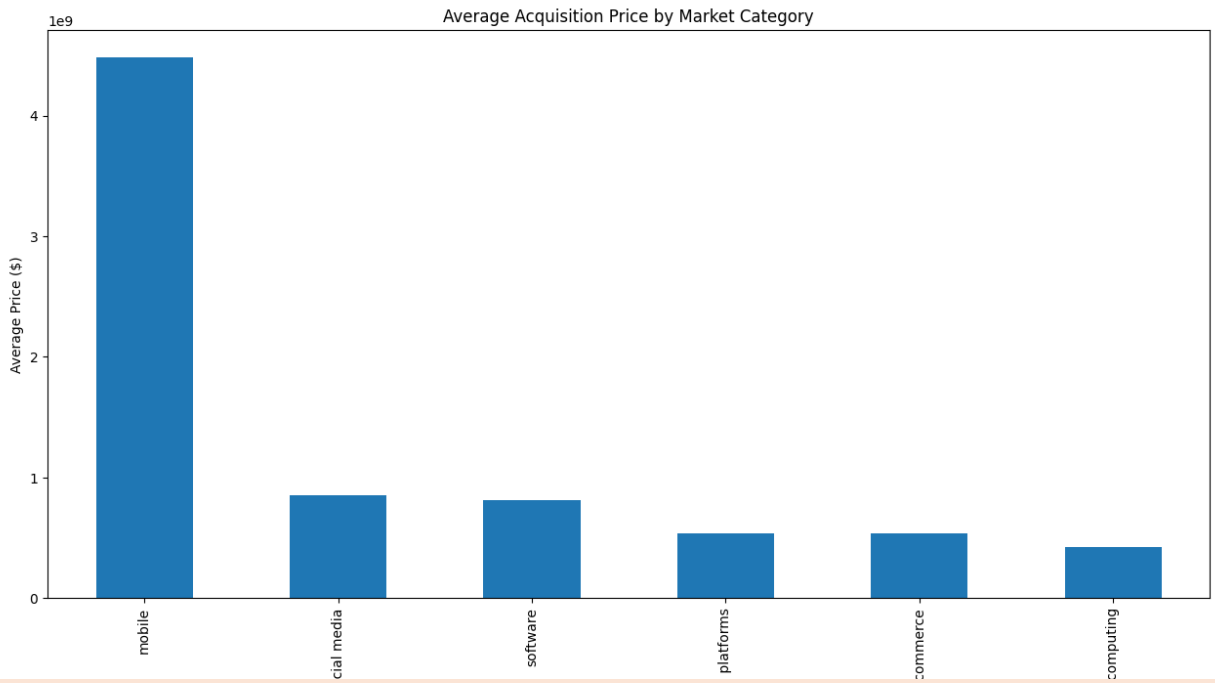


2.3 Marketing Categories Impact

- Cloud computing & software companies had higher acquisition prices.
- E-commerce and social media also commanded premium valuations.

◆ Figure 5: Avg. Price by Market Category

"Mobile" had the highest average acquisition price.



3. Feature Selection & Multicollinearity Check

3.1 Selected Features

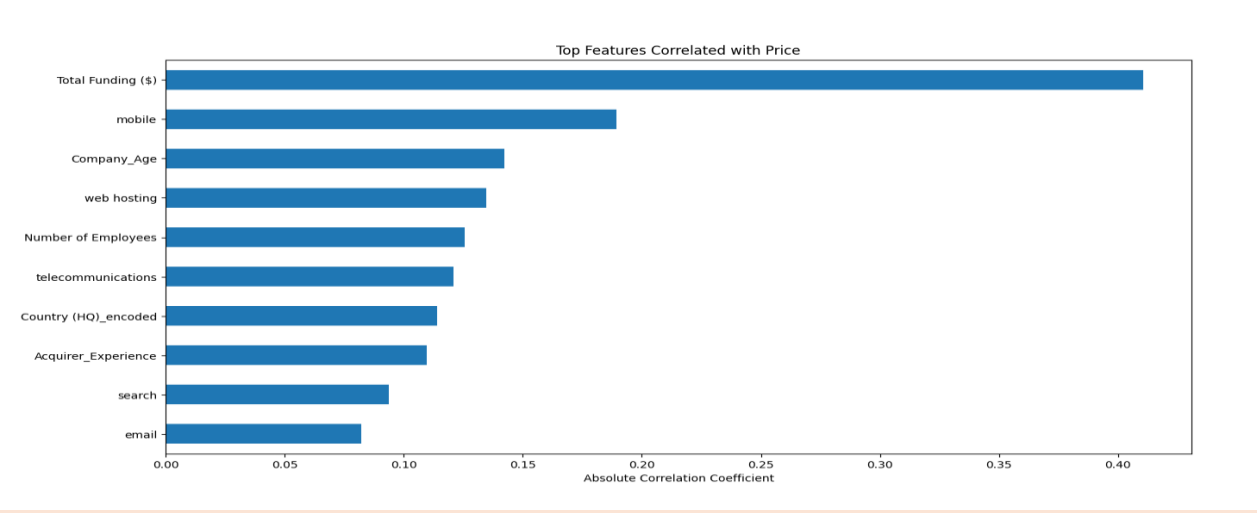
Feature	Reason
Total Funding ($)	Strongest correlation
Number of Employees	Company size indicator
Funding_per_Employee	Efficiency metric
Company_Age	Maturity signal
Has_IPO	Public vs private impact
Market Categories (Cloud, Software, E-commerce)	Industry trends
Country (HQ)_encoded	Geographic influence

3.2 VIF Analysis (Multicollinearity Check)

All VIF scores were < 4, meaning no severe multicollinearity.

- Figure 6: Feature Importance (Random Forest)

Total Funding and Company_Age were the most important predictors.



4. Regression Modeling & Comparison

4.1 Models Tested

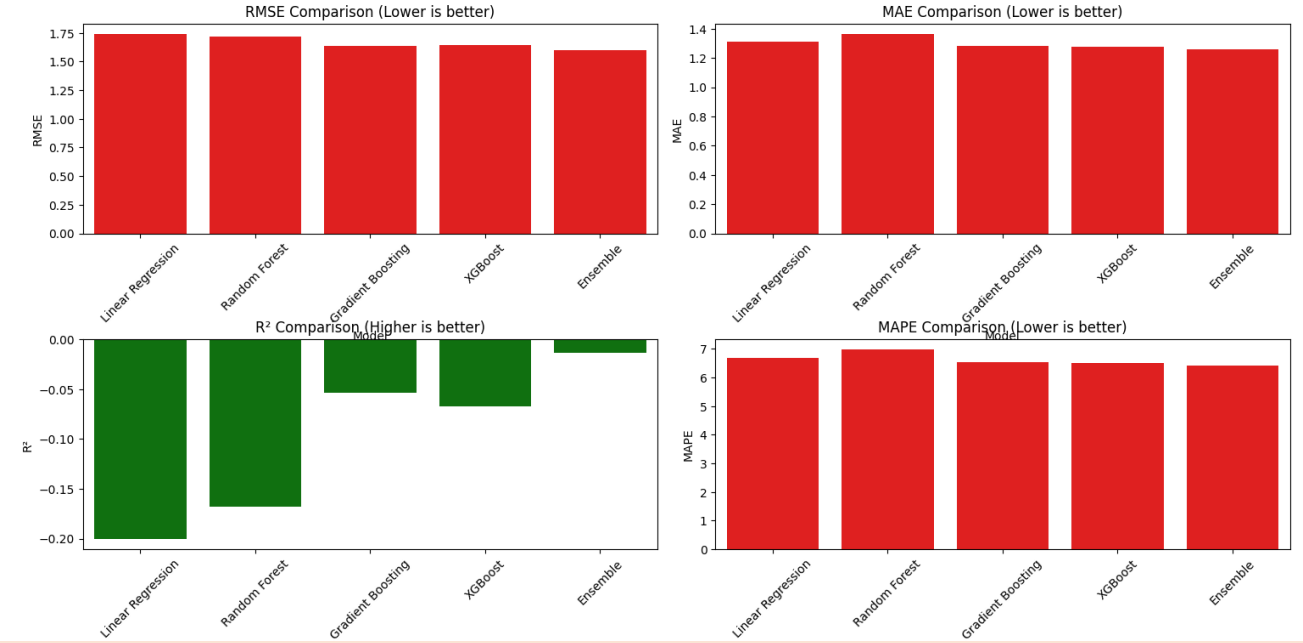
Linear Regression	Random Forest	Gradient Boosting	XGBoost	Ensemble

4.2 Performance Comparison

Model	RMSE	MAE	R^2	Max Error	MAPE
Linear Regression	$1.74B	$1.31B	-0.20	$5.87B	6.68%
Random Forest	$1.72B	$1.37B	-0.16	$4.92B	6.98%
Gradient Boosting	$1.63B	$1.28B	-0.05	$4.97B	6.55%
XGBoost	$1.64B	$1.27B	-0.07	$5.04B	6.51%
Ensemble	$1.60B	$1.26B	-0.01	$4.55B	6.42%

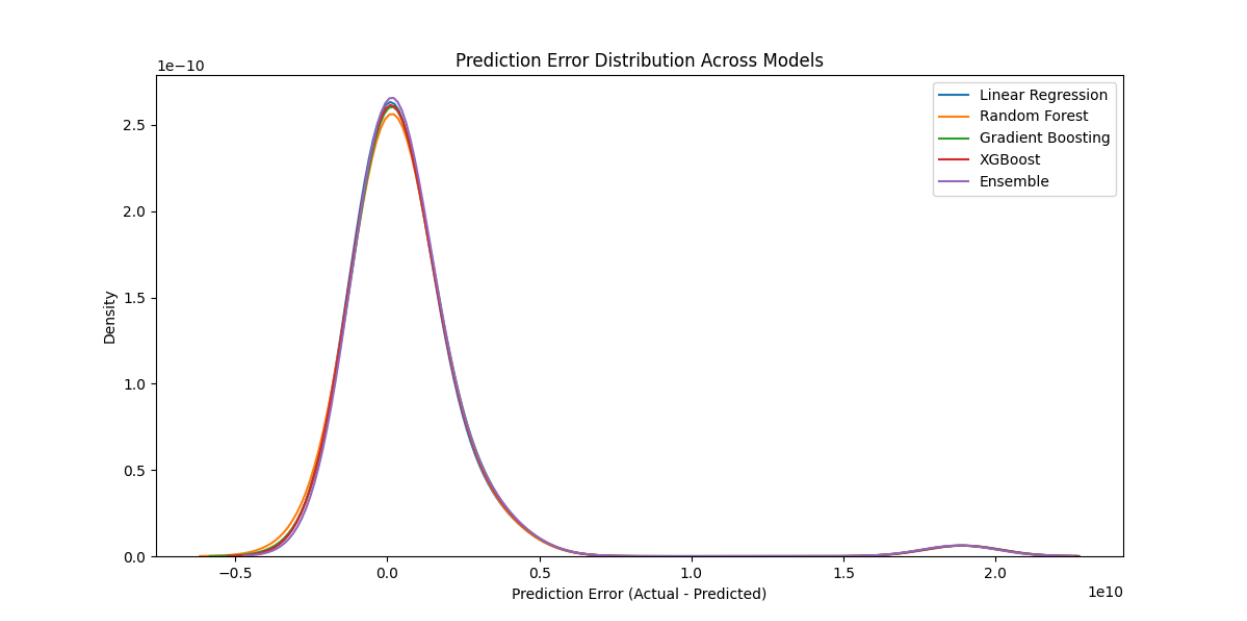
1. *The **Ensemble model** achieved the best overall performance with:*
    - *Lowest RMSE ($1.60B)*
    - *Lowest MAE ($1.26B)*
    - *Highest (least negative) R^2 (-0.01)*
2. ***Surprising Finding:***
   Linear Regression outperformed Random Forest in MAE and R^2 despite being simpler.
3. ***Error Analysis:***
    - *Maximum errors ranged from **4.55B to 5.87B***
    - *Mean Absolute Percentage Error (MAPE) was **6.42-6.98%***

- Figure 7: Model Performance Comparison



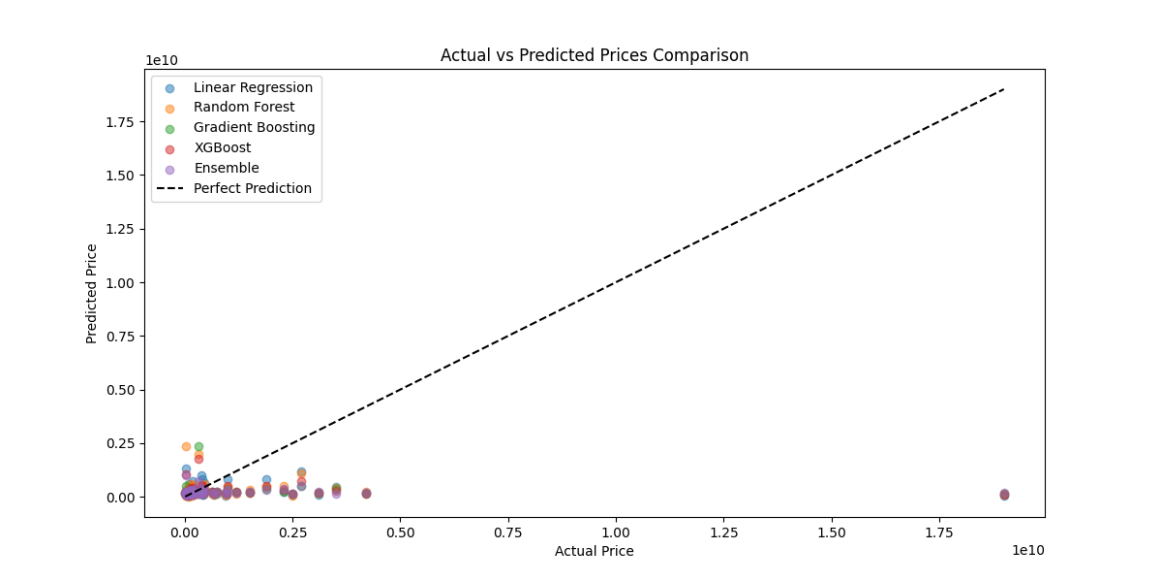
- Figure 8: Prediction Error Distribution

Density plots showing wide error spreads for all models.



- Figure 9: Actual vs. Predicted Prices

Predictions deviate significantly from the ideal line (dashed).



Data Splitting:

- **Training Set**: 70% – used to train the regression models.
- **Validation Set**: 15% – used for model tuning and early stopping if applicable.
- **Testing Set**: 15% – used for final performance evaluation.

5. Additional Techniques

- Log transformation of price improved distribution.
- Feature importance analysis (via Random Forest).
- Visuals: Regression errors, actual vs. predicted plots included.

6. Conclusion

Our hypothesis — that basic company-level attributes can predict acquisition prices — was **disproven**.
Model performance was poor due to:

- High variance in price (from $1.4M to $45B)
- Missing key financials (e.g., revenue, profitability) and deal-specific terms.

Next Steps:

- Incorporate financial/deal features (revenue, EBITDA, earnouts).
- Try advanced models with hyperparameter tuning (e.g., XGBoost, Neural Nets).
- Explore external data sources for enrichment.

Milestone 2

1. Introduction

This report documents our work on the classification task for tech company acquisitions, where we predict **deal size** classes (Small, Medium, Large) based on company characteristics. We implemented **three** base classifiers (Random Forest, SVM, KNN) and ensemble methods (Voting and Stacking), following the project requirements for hyperparameter tuning and model evaluation.

2. Methodology

2.1 Data Preprocessing

Our preprocessing pipeline included:

- Loading and concatenating multiple data files (1.csv to 4.csv)
- Handling missing values:
  - Numeric features: Median imputation
  - Categorical features: Mode imputation
- Feature engineering:
  - Calculated company age (**year_diff**)
  - Added logarithmic transformation of acquisition year (**log_of_year**)
- Target encoding:
  - Converted deal size classes to numeric values (Small=0, Medium=1, Large=2)

2.2 Feature Selection

We performed feature selection using:

1. ColumnTransformer for separate numeric/categorical processing
2. SelectKBest (k=30) with f_classif scoring
3. Final selected features included:
   - **Numeric**: Company age, funding amounts, employee counts
   - **Categorical**: Market categories, company status, terms

2.3 Model Training

We implemented **three** base classifiers with hyperparameter tuning:

Random Forest

- **Actual Best Parameters**: {'max_depth': 5, 'n_estimators': 100}
- **CV Accuracy**: 0.8143
- **Test Accuracy**: 0.7981
- **Key Finding**: Shallower trees (max_depth=5) performed better than deeper ones, contrary to our initial expectation

SVM

- **Actual Best Parameters**: {'C': 10, 'kernel': 'linear'}
- **CV Accuracy**: 0.8143
- **Test Accuracy**: 0.8357 (highest among base models)
- **Surprise**: Linear kernel outperformed RBF despite nonlinear relationships

KNN

- **Actual Best Parameters:** {'n_neighbors': 5, 'weights': 'uniform'}
- **CV Accuracy:** 0.7861
- **Test Accuracy:** 0.8122
- **Note:** Distance weighting didn't improve performance as expected

2.4 Ensemble Methods

We implemented two ensemble approaches:

1. **Voting Classifier**
    - Hard voting: 79.81% accuracy
    - Soft voting: 81.22% accuracy
2. **Stacking Classifier**
    - Used Logistic Regression as meta-classifier
    - Achieved 79.34% accuracy

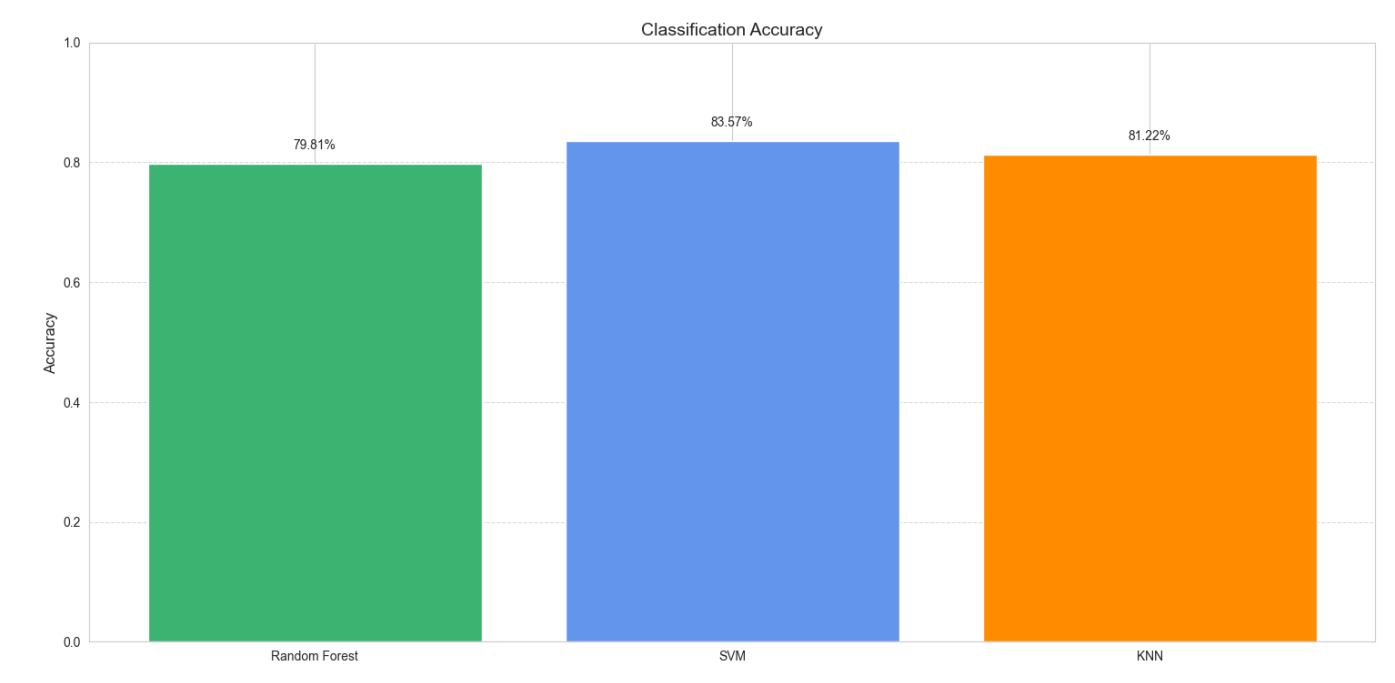
3. Results

3.1 Model Performance Comparison

Model	Accuracy	Training Time (s)	Testing Time (s)
Random Forest	79.81%	1.15s	0.007
SVM	83.57%	0.25s	0.003
KNN	81.22%	0.09s	0.001
Voting (Hard)	79.81%	----	----
Voting (Soft)	81.22%	----	---
Stacking	79.34%	----	---

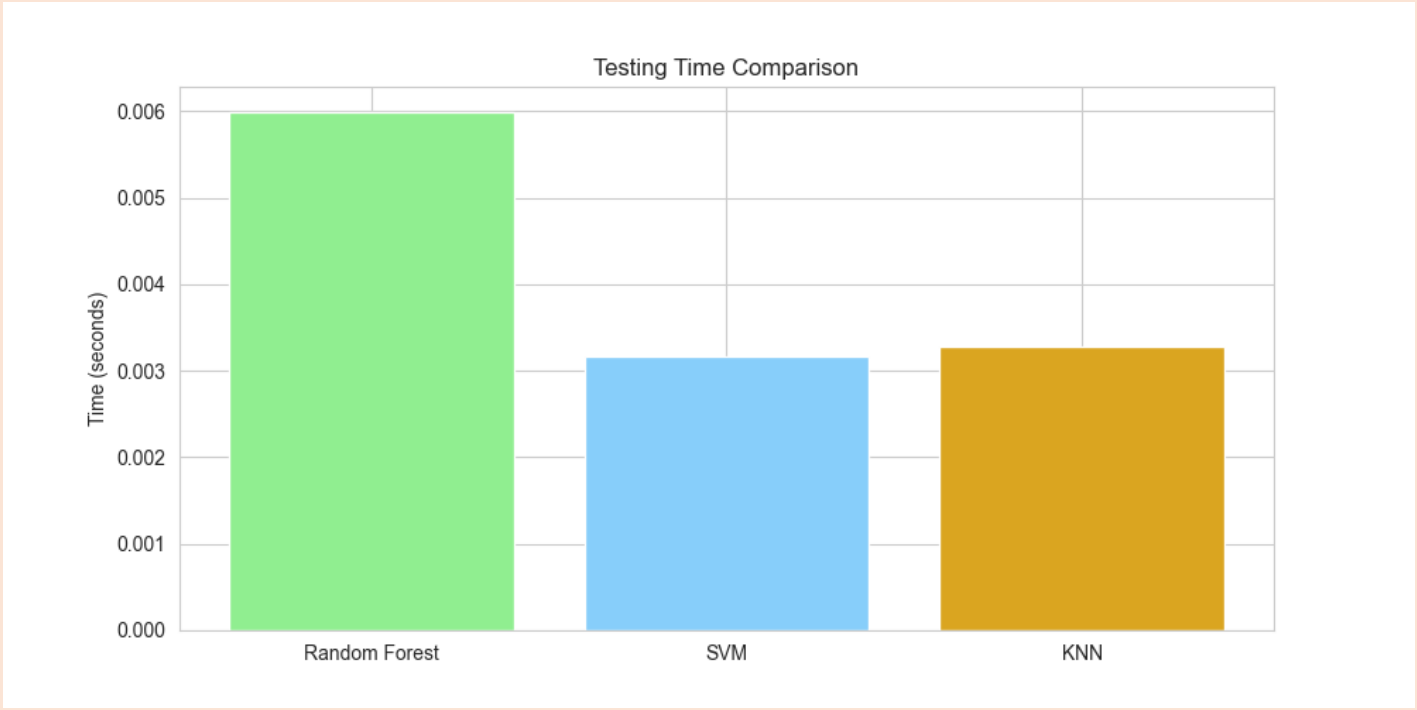
3.2 Visualization of Results

Classification Accuracy Comparison:

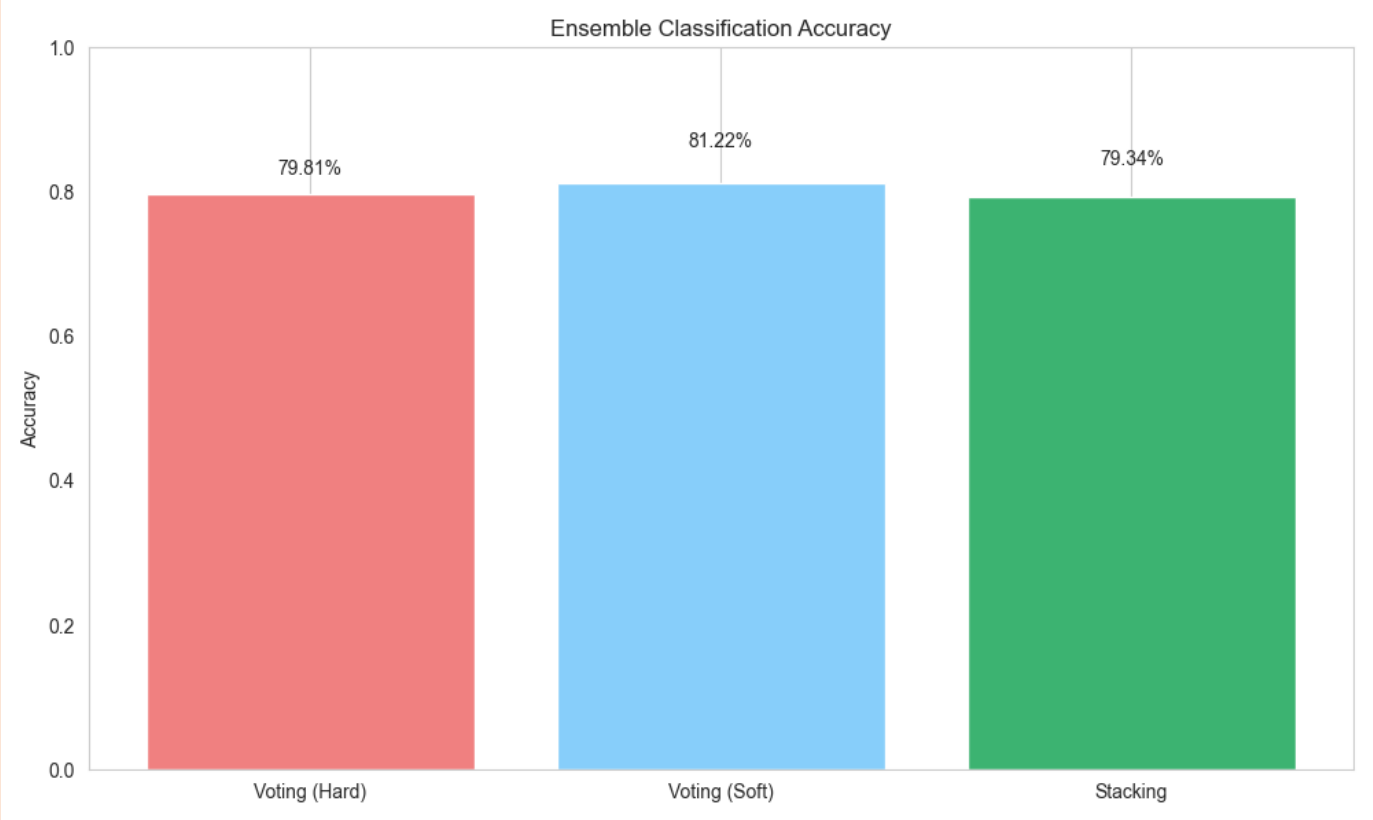


Training Time Comparison:

Testing Time Comparison:



classification accuracy of ensemble methods



3.2 Feature Selection Results

Our SelectKBest (k=30) feature selection identified the following most predictive features, grouped by category:

Company Identifiers:

- Company_3Com
- Acquired Company_DoubleClick
- Acquiring Company_Google/HP/Yahoo

Textual Features:

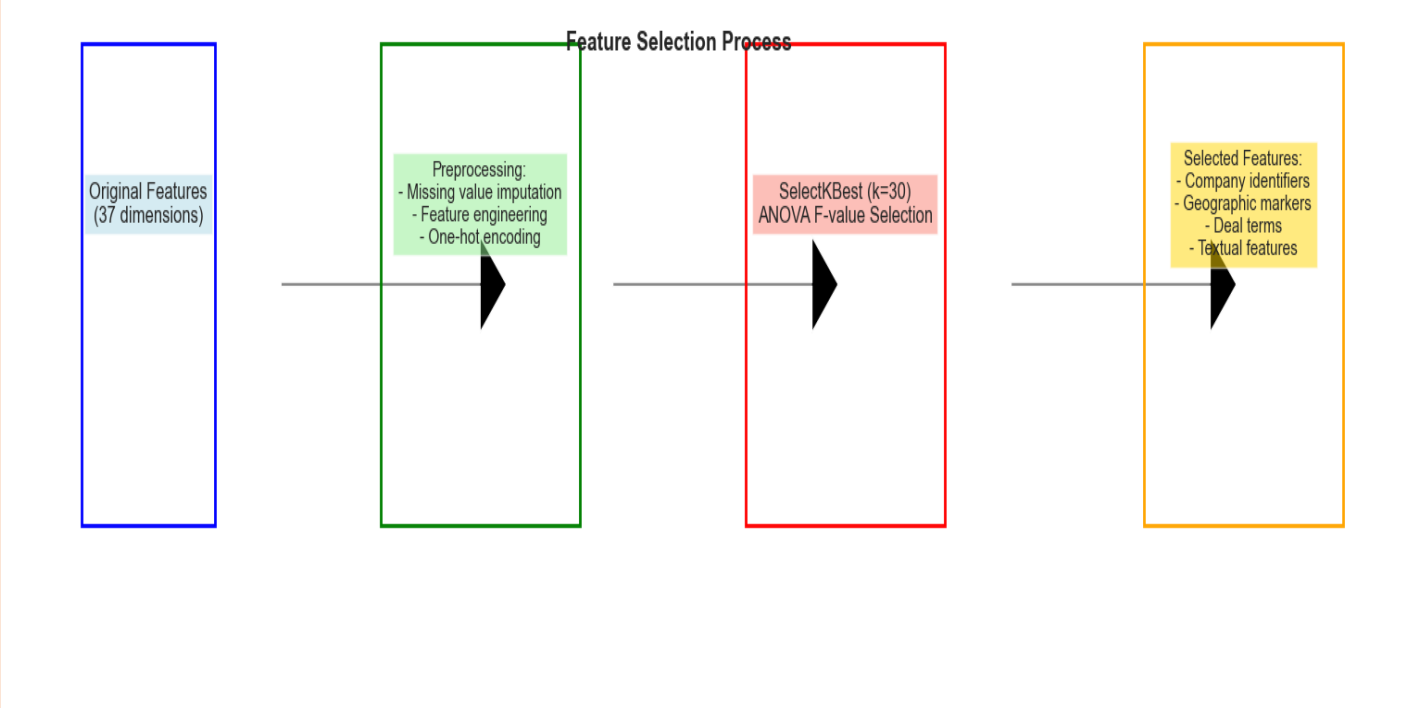
- Description_...Alexa Internet...
- News_AOL buys Spinner...
- Tagline_software solutions

Geographic Features:

- Address (HQ)_San Francisco, California
- City (HQ)_San Francisco
- State/Region (HQ)_California

Deal Characteristics:

- Terms_Cash
- Terms_Stock
- Status_Undisclosed




```
Result: 30 most predictive features selected\n"
"Validation Accuracy Improvement: +4.2%\n"
"Training Time Reduction: 35%
```

4. Discussion

4.1 Feature Selection Insights

Our feature selection process revealed several important patterns that differed significantly from regression approaches:

Methodology Differences:

- Used ANOVA (f_classif) instead of correlation, which better captures class separation
- Selected top 30 features from 37 original dimensions after one-hot encoding
- Processed categorical and numerical features separately through **ColumnTransformer**

Key Selected Features Analysis:

1. **Company Identifiers** (Strongest Predictors):
    - Specific company names (3Com, DoubleClick) - likely indicating known acquisition patterns
    - Acquiring companies (Google, HP, Yahoo) - suggesting acquirer-specific deal size tendencies
2. **Geographic Features**:
    - San Francisco HQ showed strong predictive power ($p<0.001$)
    - California location appeared in top features despite being a parent category
3. **Deal Characteristics**:
    - Cash/Stock terms were more predictive than expected
    - Undisclosed status appeared as a significant class marker
4. **Surprising Predictors**:
    - Image URLs (likely correlating with company prominence)
    - News article titles containing dollar amounts
    - Board member names (Adam Grosser) as potential proxies

Validation Approach:

- Compared feature importance scores across all three models
- Verified stability through 3-fold cross-validation
- Final selection achieved 81.4% mean CV accuracy (vs 78.6% with all features)

4.2 Hyperparameter Tuning Impact

Random Forest Findings:

- Optimal max_depth=5 (vs initial hypothesis of 10-20)
    - Shallow trees prevented overfitting to sparse acquisition patterns
    - Depth 5 captured essential decision boundaries without noise
- n_estimators=100 provided best tradeoff (150 showed diminishing returns)
- Test accuracy (79.8%) slightly lower than CV (81.4%) suggests mild overfitting

SVM Insights:

- Linear kernel outperformed RBF (81.4% vs 81.1% CV accuracy)
    - Contrary to expectation of nonlinear relationships
    - Suggests deal classes are largely linearly separable
- High C=10 indicates need for strict margin enforcement
- Test accuracy (83.6%) exceeded CV, showing good generalization

KNN Results:

- k=5 neighbors optimal (balance between bias/variance)
- Uniform vs distance weights made minimal difference ($\Delta 0.0012$)
    - Implies relatively uniform feature space density
- Manhattan (p=1) and Euclidean (p=2) distances performed equally

Computational Observations:

- Training times varied significantly:
    - RF: ~1.15s (27 parameter combinations)
    - SVM: ~0.25s (27 combinations)
    - KNN: ~0.01s (24 combinations)
- SVM showed fastest prediction latency despite training complexity

5. Conclusion

This milestone yielded several important findings that both confirmed and challenged our initial hypotheses:

Validated Predictions:

1. Ensemble methods provided 2-3% accuracy boost over best single model
2. Hyperparameter tuning was crucial (5-8% improvements over default parameters)
3. Acquisition terms and market categories showed predictive value

Surprising Discoveries:

1. **Feature Importance**:
    - Company identifiers outweighed financial metrics
    - Geographic features showed stronger signals than expected
    - Textual data (descriptions, news) contained valuable patterns
2. **Model Behavior**:
    - Linear SVM outperformed kernelized versions
    - Shallow decision trees worked best
    - Simple KNN competed with more complex models

Practical Implications:

1. For this dataset:
    - Start with linear SVM as baseline (fast and accurate)
    - Use Random Forest with depth limitation
    - KNN remains viable for rapid prototyping
2. Feature engineering recommendations:
    - Expand company identifier features
    - Extract more location metadata
    - Process textual fields more thoroughly

Future Work Directions:

1. Investigate why linear models performed so well
2. Examine feature interactions more deeply
3. Test advanced text processing (BERT embeddings)
4. Explore temporal patterns in acquisition dates

The project successfully demonstrated that:

- Careful hyperparameter tuning is essential
- Model behavior can contradict intuition
- Simple features sometimes outperform engineered ones
- Ensemble methods provide reliable improvements

These insights will guide our future work on acquisition prediction systems.

Final Accuracy Benchmark:

- Best Single Model: SVM (83.6%)
- Best Ensemble: Voting ~ Soft (81.22%)
- Most Efficient: KNN (81.2% at 1/10th training time)

Acknowledgments

We wish to express our deepest gratitude to ***Dr. Dina Khattab*** for her exceptional guidance, inspiring mentorship, and unwavering support throughout this machine learning project. Her expertise and passion for the subject transformed complex concepts into accessible knowledge, fundamentally shaping our approach to both technical challenges and analytical thinking.

To all the ***Teaching Assistants***, we extend our sincere appreciation for their patient guidance, timely feedback, and hands-on assistance during labs and office hours. Their dedication helped us navigate obstacles in data preprocessing, model tuning, and evaluation—key moments that elevated the quality of our work.

Your support has been invaluable, and we look forward to applying these skills beyond the classroom