ROBERT H. SMITH
SCHOOL OF BUSINESS

Project Overrun Predictor

BUDT 751: Harnessing AI for Business

Spring 2025

Sanjay Dari Veerabasappa

Professor Tej & Professor McCoy

May 10th, 2025

1. Introduction

- Overview of the Role and Business Context:
  This project was approached from the perspective of a data science team within a consulting or engineering firm, tasked with tackling the significant business challenge of software project effort overruns. In an industry where delivering projects on time and within budget is critical for client satisfaction, profitability, and competitive standing, traditional estimation methods often prove insufficient against the complexities of modern software development. This context underscores the urgent need for more sophisticated, data-driven tools capable of offering predictive insights and facilitating proactive risk management.
- General Overview of Role / Business:
  Our team assumed the role of AI specialists and data scientists to conceptualize and develop an innovative tool, the "Overrun Advisor." This system is designed for project managers, solutions architects, and consultants who require robust methods to plan, execute, and deliver software projects successfully. Operating within the software development and IT consulting sectors—where overruns have severe financial and reputational consequences—the "Overrun Advisor" aims to provide substantial business value. It achieves this by enabling the early identification of at-risk projects, offering clear explanations for potential overruns through explainable AI, and suggesting actionable mitigation strategies via generative AI. This fosters a data-informed project management culture, moving beyond heuristic-based decision-making.

2. Methods and Methodologies

- Job Functions:
  The development of the "Overrun Advisor" integrated several key data science and AI job functions. The process began with Data Analysis and Preprocessing, involving exploration and cleaning of the china.csv dataset using SimpleImputer. A core activity was Data Engineering and Augmentation; to overcome the initial dataset's size limitation (499 records), a two-stage augmentation (Gaussian Mixture Models then an empirical method) expanded it to over 15,000 samples. This was followed by Feature Engineering, where domain-relevant features like overrun_pct (target), churn_rate, and team_intensity were created from 16 raw inputs.
  The **Machine Learning Modeling** phase involved selecting, training, and optimizing an XGBoost Regressor, which included stratified train/test splitting, sample weighting, and hyperparameter tuning with RandomizedSearchCV. **Model Evaluation and Critical Analysis** assessed performance against a baseline using RMSE, R^2, and MAE. **Explainable AI (XAI)** techniques, specifically SHAP (TreeExplainer), were implemented for transparency. A novel **Generative AI Integration** used the DeepSeek-R1 LLM to translate SHAP insights into actionable tips. Finally, **Application Development** involved prototyping an interactive Streamlit UI, deployed via pyngrok.
- AI Impact:
  Artificial Intelligence is central to the "Overrun Advisor's" effectiveness. The predictive analytics from the tuned XGBoost model provide quantitative forecasts of project effort overrun percentages, enabling proactive planning. The Explainable AI component, SHAP, demystifies these predictions by highlighting key risk drivers like N_effort or scope Deleted, building user trust and enabling targeted interventions. This aligns with the state-of-the-art demand for

transparent AI.
Furthermore, the **Generative AI** integration significantly enhances the tool's practical impact. The DeepSeek-R1 LLM translates complex SHAP outputs into simple, actionable natural language advice. This makes sophisticated insights accessible to project managers without deep AI expertise, directly supporting better decision-making and reflecting current trends in applied AI that prioritize user-centric, actionable intelligence.

- Future Technologies:
  The "Overrun Advisor" platform is designed for future evolution. Exploration of Advanced Deep Learning models (e.g., LSTMs for time-series project data) could offer more nuanced predictions with richer datasets. Automated Machine Learning (AutoML) could streamline model optimization. For dynamic advice, Reinforcement Learning presents an interesting avenue. More sophisticated NLP could extract features from unstructured project documents, while Federated Learning could enable collaborative model training on diverse, decentralized datasets while preserving privacy, reflecting cutting-edge AI paradigms.

3. Product Overview: The "Overrun Advisor"

The "Overrun Advisor" is an interactive web application serving as an intelligent assistant for project managers. It aims to provide early warnings for potential project effort overruns, offering clear, data-driven insights and actionable recommendations to improve project outcomes.

Key features include:

- **CSV Data Upload:** Users upload project data (expecting 16 raw features like AFP, scope changes, N_effort, based on china.csv).
- **Automated Feature Engineering:** The app calculates 4 critical engineered features (churn_rate, pdr_ratio, prod_nominal, team_intensity).
- **Effort Overrun Prediction:** A pre-trained, tuned XGBoost model (xgb_tuned_best_param.pkl) predicts the predicted_overrun_pct.
- **Interactive Prediction & Data Display:** Results are shown in a clear tabular format.
- **SHAP Driver Visualization:** Global SHAP feature importance (bar chart) and instance-level impacts (beeswarm plot) are displayed for the top 10 features.
- **AI-Generated Mitigation Tips:** A "Get AI Tips" button sends the top 5 SHAP drivers to the DeepSeek-R1 LLM, which returns concise, actionable, one-sentence tips for each driver.

4. Technical Implementation

The "Overrun Advisor" was developed through a systematic technical pipeline designed for robust data handling, effective modeling, and the generation of insightful outputs, with clear business relevance at each stage.

The project commenced with **Data Collection and Preprocessing** using the china.csv dataset (499 projects, 20 features covering size, scope, resources, and effort). After loading with pandas, SimpleImputer (most_frequent strategy) addressed potential missing values, though the dataset was largely complete. The Dev.Type feature, initially a constant 0, was noted for diversification during augmentation. This initial preparation ensured a reliable base for the subsequent, more complex stages.

To address the limited initial dataset size, a critical **Data Augmentation** strategy was executed in two stages, expanding the data to over 15,000 samples. Stage 1 involved **Gaussian Mixture Model (GMM) Augmentation**, where a GMM (5 components) generated 10,000 synthetic samples for numeric features from the cleaned china.csv data, with added noise (10% of feature standard deviation) and clipping to original bounds. Crucially, Dev.Type was sampled uniformly [0,1] to introduce variability. This formed china_synthetic.csv. Stage 2, **Empirical Augmentation**, used china_synthetic.csv as a base to generate an additional 5,000 samples by fitting another GMM on predictor features, sampling them with light jitter (2% standard deviation), empirically sampling overrun_pct, and then recomputing Effort. The 10,000 GMM-augmented samples were primarily used for model training, upweighted (sample weight of 5.0) to emphasize their foundational patterns. This augmentation was vital for developing a more stable and generalizable model.

The core **Model Implementation** began with **Feature Engineering**, creating the target overrun_pct and four key predictors (churn_rate, pdr_ratio, prod_nominal, team_intensity) from the 16 raw/GMM-derived inputs, resulting in 20 features for the model. An **XGBoost Regressor** was chosen for its performance and regularization capabilities. It was trained on an 80/20 stratified split (by overrun_pct bins) of the GMM-augmented data. **Hyperparameter Tuning** via RandomizedSearchCV (3-fold CV, 20 iterations) identified optimal parameters (e.g., max_depth=6, learning_rate=0.01, n_estimators=500, reg_alpha=0.1, reg_lambda=1), with the tuned model saved as xgb_tuned_best_param.pkl.

A **Critical Analysis** of this tuned model on the unseen test set yielded an RMSE of approximately **3.9876**, an R^2 of **0.6257** (explaining ~62.6% of overrun variance), and an MAE of **1.1141**. This significantly improved upon the baseline untuned XGBoost (Test R^2: 0.6030, Test RMSE: 4.1066). For **Explainability**, a shap.TreeExplainer identified N_effort (Mean |SHAP| = 2.104), Deleted scope (1.236), and prod_nominal (0.448) as top global drivers. The **Streamlit Application** (app.py) integrates these elements, allowing data upload, prediction, and SHAP visualization. Finally, **LLM Integration** with DeepSeek-R1 translates the top 5 SHAP drivers into actionable tips, demonstrating a state-of-the-art combination of predictive, explainable, and generative AI.

5. Bias Detection and Evaluation

Addressing potential biases is integral to responsible AI development. The primary source of potential bias for the "Overrun Advisor" is the **inherent nature of the foundational** china.csv **dataset**. Its likely specific geographical/organizational context might limit the model's global generalizability. The initial constancy of Dev.Type (value 0) suggested a possible skew, which augmentation aimed to diversify. However, the model's foundational learning remains influenced by this specific dataset.

Data augmentation, while crucial for dataset expansion, could reflect and potentially amplify biases from the original 499 projects if those patterns are unrepresentative of a broader population. The **choice of features and the XGBoost algorithm** could also inadvertently learn or perpetuate biases if features correlate with unobserved sensitive attributes or if the algorithm overemphasizes certain patterns from biased data.

To mitigate these risks, several **actions were implemented**:

- The two-stage **data augmentation** introduced controlled variability (e.g., for Dev.Type) to

enhance generalization.

- **Regularization techniques** (L1 via `reg_alpha=0.1` and L2 via `reg_lambda=1`) in XGBoost tuning helped prevent overfitting to potentially biased training data patterns.
- **SHAP for transparency** allows scrutiny of feature influences, helping identify disproportionate impacts that might signal bias.
- **Stratified sampling** for train/test splits (by `overrun_pct` bins) ensured representative model evaluation across different overrun levels.

Despite these efforts, **limitations** in bias mitigation exist given a single primary data source. Future work must include incorporating more diverse, global datasets and employing formal bias auditing toolkits (e.g., AIF360). Continuous refinement of LLM prompts for equitable advice is also essential. Our goal is transparency regarding data limitations and our efforts to address them.

6. Future Implementation

The "Overrun Advisor" prototype offers a strong platform for future enhancements. Key directions include **enhanced data integration and real-time capabilities**, potentially by connecting with project management systems like Jira for dynamic data input and continuous monitoring. This would transform the tool into a live risk management dashboard.

Advancements in predictive models could be pursued as richer datasets become available, such as exploring deep learning architectures (e.g., LSTMs for time-series project data). Incorporating **sophisticated NLP techniques** to extract features from unstructured project documents (proposals, reports) could significantly enrich model inputs by capturing qualitative risk factors.

The **LLM-generated advice could be further personalized**, tailoring recommendations based on richer project-specific contexts beyond just SHAP drivers. A highly valuable addition would be a **scenario planning ("what-if" analysis) interface** in the Streamlit app, allowing users to simulate the impact of changing project parameters on predicted overrun.

Finally, establishing a robust **user feedback loop and an MLOps pipeline** for continuous model monitoring, automated retraining, and LLM prompt refinement is crucial for long-term accuracy and relevance.

7. Challenges

Developing the "Overrun Advisor" involved several key challenges. **Initial data scarcity**, with only 499 projects in `china.csv`, necessitated the extensive two-stage data augmentation. Ensuring the **realism and generalizability of this augmented data** was a constant focus, as synthetic data may not perfectly capture all real-world nuances.

Interpreting SHAP values with potentially correlated input features required careful analysis to avoid misattributing influence. Significant effort was also invested in **LLM prompt engineering** for the DeepSeek-R1 model to ensure concise, relevant, and actionable tips.

Computational resources were a constraint, particularly for hyperparameter tuning

(RandomizedSearchCV with 60 model fits) and SHAP value generation. Lastly, a practical challenge for real-world adoption is ensuring **easy user data mapping** to the 16 specific raw input features required by the application.

8. Conclusion

The "Project Overrun Advisor" effectively demonstrates the power of an AI-driven system to predict software project effort overruns with a notable R^2 of 0.6257 and an RMSE of 3.9876 on test data, and more importantly, to provide explainable insights and actionable mitigation strategies. SHAP analysis transparently identified key drivers like N_effort (Mean |SHAP| = 2.104), Deleted scope (1.236), and prod_nominal (0.448), offering project managers clear focus areas. The innovative integration of this explainability with the DeepSeek-R1 LLM translates complex model outputs into practical, natural language advice, enhancing its business utility.

Addressing data scarcity through a two-stage augmentation process was fundamental to developing a robust model. While acknowledging potential biases from the foundational dataset, mitigation steps like regularization and transparent SHAP analysis were implemented. This project showcases a synergistic approach combining predictive, explainable, and generative AI, empowering a shift from reactive to proactive project risk management and holding significant potential for future real-world impact.