Capstone Project Final Report

Post Graduate Program in Data Science
(Specialization: Generative AI)

”The Impact of Generative AI on Enterprise Workforces”

Domain of Project: Technology
Group Number: 2
Mentor: Vibha Santhanam
Date: 06th October, 2025

Team Members:

- B. Sai Tejaswini (Team Leader)
- Shivam Dhar
- Yash N
- Darshan M. Goswami
- Keerthi Shetty

1. Summary of Problem Statement, Data, and Findings

This project investigates how generative AI (GenAI) adoption influences enterprise workforce outcomes, focusing on employee productivity and sentiment. The analysis uses a Kaggle dataset of 100,000 records related to enterprise AI adoption, containing categorical features (Company, Industry, Country, GenAI Tool, and raw sentiment text) and numeric features (Adoption Year, Employees Impacted, New Roles Created, Training Hours, and Productivity Change). The primary objective is to predict workforce productivity change (as a percentage, binned into Low/Medium/High categories) and employee sentiment polarity based on the context of AI adoption. These predictive insights aim to provide actionable guidance for HR and management in shaping AI implementation strategies.

The data science workflow involved extensive preprocessing (scaling, encoding, text vectorization) and feature engineering to capture contextual effects. Our final model, a LightGBM classifier, achieved about 81% accuracy in classifying productivity change categories and 88% accuracy in sentiment polarity classification (F1-score = 0.86). Key findings include consistent productivity gains on the order of 15–20% across industries after GenAI adoption and the importance of contextual factors in predicting these gains. Top predictive factors were found to be training hours per employee, number of new roles created, and overall employee sentiment. These results suggest that positive employee perception correlates strongly with higher productivity outcomes, emphasizing the role of targeted training and sentiment management.

Key insights include:

- Productivity Gains: Adopters experienced productivity improvements averaging 15–20% after GenAI implementation.
- Model Performance: The tuned LightGBM model significantly outperformed baseline approaches, indicating the value of advanced modeling and rich features.
- Feature Importance: Employee enablement factors (e.g. training hours, new roles) and sentiment polarity emerged as critical predictors of success.
- Organizational Implications: Positive sentiment is linked to better outcomes, highlighting the need for focused HR interventions around training and communication.

1.1 Overview of the Final Process

A structured data science pipeline was applied to address the problem. The key dataset was a Kaggle "Enterprise GenAI Adoption & Workforce Impact" dataset (100,000 records × 10 features) covering global companies and AI tools. Initial preprocessing confirmed data quality (no missing or duplicate entries). Numeric variables (such as Training Hours and Productivity Change) were standardized, while categorical variables (Industry, Country, etc.) were label-encoded. Employee sentiment text was processed by tokenization, stop-word removal, and TF-IDF vectorization, combined with a polarity scoring approach. The target variable (productivity change) was discretized into Low, Medium, and High categories for classification.

Data and Preprocessing: The dataset comprised both structured (categorical and numeric) and unstructured (text sentiment) data. After verifying data cleanliness, numeric features were scaled and categorical features were encoded. Sentiment text was converted into TF-IDF vectors with an accompanying sentiment polarity score.

Feature Engineering: To capture contextual effects, new interaction features were created (e.g. Industry×Country, Tool×Sentiment). Ratio features were derived, such as training hours per impacted employee and new roles per employee. These engineered features enriched the dataset to capture nonlinear relationships that simple models might miss.

Modeling: Initial models (logistic regression, random forest, and XGBoost) were trained as baselines but underperformed. The final solution employed a

LightGBM classifier, chosen for its efficiency with tabular data. Hyperparameters (such as learning rate, number of leaves) were optimized via cross-validation, and balanced class weights were applied to mitigate target class imbalance.

The overall methodology combined these steps iteratively: insights from exploratory analysis guided feature engineering, and modeling results informed further refinements. This integrated approach led to a robust final model for predicting productivity changes and sentiment in response to GenAI adoption.

Step-by-Step Walkthrough of the Solution:

Data Loading and Preprocessing: The Kaggle dataset of 100,000 enterprise adoption records was examined to ensure data quality. No missing or duplicate values were present. Numeric features (e.g. Training Hours, Productivity Change) were scaled using a Standard Scaler, and categorical features (Company, Industry, Country, AI Tool) were label-encoded. The productivity change percentage was binned into three discrete classes (Low, Medium, High) to frame a classification problem.

Exploratory Data Analysis: Trends in the data were analyzed. Adoption of GenAI peaked around 2023, affecting between 100 and 20,000 employees across companies. The employee sentiment distribution in the dataset was roughly 23% positive, 54% neutral, and 23% negative. Correlation analysis revealed weak linear relationships between numeric features and productivity, suggesting that contextual factors are more influential. These findings motivated the creation of contextual and interaction features to better capture the underlying patterns.

Feature Engineering: Contextual interaction features were constructed, such as Industry×Country and Tool×Sentiment. Ratio features were also added (e.g. Training Hours per impacted employee, New Roles per employee). The sentiment text was further processed: after tokenization and stopword removal, TF-IDF vectors were computed and a sentiment polarity score was assigned. These enriched features aimed to incorporate non-linear effects and capture the nuanced impact of workforce factors on productivity.

Initial Modeling: Logistic regression, random forest, and XGBoost classifiers were trained on the prepared data to establish baselines. These initial models achieved only about 33% accuracy in predicting the three productivity classes, effectively no better than random guessing. Predictions were heavily biased toward the Medium productivity class, reflecting the underlying class imbalance and indicating that linear models were insufficient for this task.

Model Refinement and Final Solution: Given the poor initial results, a LightGBM classifier was selected for its performance with heterogeneous data. Hyperparameters were tuned using grid search and cross-validation. Class imbalance was addressed by assigning balanced class weights. The final LightGBM model significantly improved performance, achieving roughly 81% accuracy (F1-score = 0.75) on productivity classification. A SHAP (SHapley Additive exPlanations) analysis of the model confirmed that Training Hours per employee, New Roles Created per employee, and Sentiment Polarity were the most influential features driving predictions.

Each step built on the previous one insights from EDA informed feature engineering, which in turn enabled the final model to capture complex patterns. The iterative process culminated in a robust predictive solution for enterprise workforce outcomes.

2. Model Evaluation

The final model is a LightGBM classifier optimized for multi-class prediction of productivity change.
Hyperparameters such as learning rate, number of leaves, and tree depth were tuned via cross-validation on the training set. The model used a multi-class objective with balanced class weights to handle the uneven distribution of productivity categories.

Evaluation metrics demonstrated strong performance and robustness. On a holdout test set, the model achieved an accuracy of ~81% and an F1-score of 0.75 for productivity classification.
Class-wise evaluation showed a recall of 0.94 for the High-productivity class (indicating most high-impact cases were correctly identified) and a precision of 0.92 for the Low-productivity class (indicating few false alarms in low-impact predictions). These metrics represent a substantial improvement over the baseline (33% accuracy). The sentiment classifier also performed well, with 88% accuracy and F1 = 0.86 in predicting positive, neutral, or negative sentiment polarity.

To validate the solution's reliability, we examined confusion matrices and found that all classes were predicted with reasonable balance, unlike the initial bias toward Medium-class predictions. The high recall for critical classes (High productivity) and high precision for low-impact cases suggest that the model can be used confidently for decision support. SHAP analysis provided interpretability: it confirmed that features related to workforce enablement (training hours, new roles) and employee sentiment were the top drivers of the model's predictions, aligning with domain expectations and lending credibility to the findings.

Overall, the evaluation indicates that the final model is both accurate and robust, capable of capturing the complex interplay of factors that determine workforce productivity changes due to GenAI adoption.

3. Comparison to Benchmark

The final solution was compared to both naive and initial benchmarks. A simple baseline (predicting the majority class or random guessing among three classes) yielded only ~33% accuracy, as expected by chance. The initial models (logistic regression, random forest, XGBoost) similarly hovered around this baseline, confirming that more sophisticated techniques were needed.

In contrast, our tuned LightGBM model delivered an accuracy of 81%, dramatically outperforming the benchmark. This improvement reflects the combined effect of advanced ensemble methods and the enriched feature set. In practical terms, the final model made over twice as many correct predictions as the baseline. The higher F1-scores and balanced class-wise performance further demonstrate that the model improved upon benchmarks for each category of productivity change.

In summary, the project substantially exceeded the initial benchmarks. The use of LightGBM and comprehensive feature engineering made the solution much more effective than simpler approaches, validating our final methodology over the baseline assumptions.

4. Implications

The findings of this analysis have several important implications for enterprise strategy and HR management:

Data-Driven Adoption Planning: Enterprises can use predictive models like the one developed here to forecast productivity outcomes of GenAI projects. This enables more informed decision-making about where and how to allocate resources for AI deployment.

Focus on Employee Enablement: Training hours and new role creation were identified as top predictors of productivity gains. This suggests that companies should emphasize comprehensive training programs and thoughtful role redesign when implementing GenAI tools. Investing in employee skills development appears likely to maximize the benefits of AI adoption.

Sentiment Monitoring: Positive employee sentiment correlated with higher productivity improvements. Therefore, organizations should regularly gauge employee sentiment through surveys or advanced NLP analysis of feedback to identify concerns early. Proactive interventions (e.g. additional support or communication) can then be made to maintain positive perceptions and smooth transitions.

Advanced Analytics: The use of ensemble models and text analytics proved valuable. Companies are encouraged to invest in advanced analytics capabilities (including transformer-based NLP models like BERT or GPT) to better interpret employee feedback and workforce metrics, enhancing the accuracy of predictions and the depth of insights.

Continuous Evaluation: Given that productivity gains varied by context, enterprises should implement ongoing tracking of GenAI initiatives. Longitudinal studies or time-series monitoring can validate model predictions over time and adjust strategies as needed. This will help ensure sustainable improvement and organizational learning.

These implications underscore that successful GenAI adoption is not just a technical challenge but a socio-technical one. Data-driven HR strategies focusing on training, role management, and employee sentiment are key to realizing the productivity benefits indicated by our analysis.

5. Limitations

Several limitations of the current solution should be noted:

Data Scope: The analysis relies on a single Kaggle dataset of enterprise GenAI adoption. This dataset may not capture all industries or geographies uniformly, and it may omit important contextual variables (such as workforce skill levels or company size). Therefore, the model's applicability might be limited outside the data's domain.

Target Discretization: We transformed productivity change into categorical bins (Low/Medium/High). While this enabled classification modeling, it loses information about the magnitude of change. A regression approach predicting exact percentage change might capture more nuance but would be more sensitive to noise.

Feature Limitations: Some potentially important factors were not included. For example, economic conditions, organizational culture, or detailed employee demographics could influence outcomes but were unavailable. The sentiment analysis used a basic polarity score, which may not reflect nuanced emotions or contextual subtleties. More sophisticated NLP could improve accuracy.

Temporal Dynamics: The model is static and does not account for time-dependent effects. GenAI adoption impacts may evolve over months or years (e.g., learning curves, technology maturation), which our snapshot model cannot capture. A time-series or longitudinal model would be needed for dynamic predictions.

Generalizability: Although the model performed well on the given data, real-world deployment could introduce unforeseen variations (such as changes in AI technology or workforce structure). The model may require re-training or adaptation when applied to new datasets or different company contexts.

Explainability: Like many ensemble models, LightGBM is relatively complex. While SHAP analysis provided some interpretability, the model is not fully transparent. Stakeholders may require further explanation of individual predictions or simpler surrogate models for certain decisions.

Acknowledging these limitations is crucial. Future work should aim to gather more diverse data, incorporate additional relevant features, and explore alternative modeling approaches to address these gaps.

6. Closing Reflections

This project provided valuable learning experiences in applying data science to a complex, real-world problem. Key takeaways include the importance of combining domain knowledge (in human resources and organizational behavior) with technical modeling. The iterative process reinforced that thorough feature engineering and model tuning are often more impactful than any single algorithm choice. We also gained insight into the strengths of ensemble methods like LightGBM for handling heterogeneous data and capturing nonlinear relationships.

In future work, we would expand the scope and depth of the analysis. This could include incorporating additional data sources (such as country-level digital readiness or workforce skill indices) to enrich the dataset. We would explore more advanced NLP techniques (e.g. transformer-based models) for sentiment analysis to capture subtle employee concerns. Experimenting with regression or time-series models could provide finer-grained insights into productivity trajectories. Finally, validating the approach with real-world case studies or pilot implementations would strengthen confidence in the model's practical utility.

Overall, the capstone process deepened our understanding of how generative AI can reshape enterprise workforces and how data-driven methods can inform strategic decisions. The lessons learned and methodologies developed here will guide our future projects involving AI and workforce analytics.