📊 Project Performance Report: Customer Satisfaction Prediction

🎯 Executive Summary

This report summarizes the efforts and outcomes of the machine learning project to predict customer dissatisfaction. Utilizing a dataset with 370 anonymized features, a highly accurate ensemble model was developed. The final result achieved was an AUC score of 0.8407, a competitive and excellent outcome that validates the applied methodology.

📈 Insights from Exploratory Data Analysis (EDA)

Severe Class Imbalance: The most critical discovery was that only 4% of customers in the training data were classified as "dissatisfied." This challenge guided the subsequent model-building and evaluation strategies. The AUC Score metric was chosen over Accuracy to ensure a fair and robust evaluation.

Data Quality: The dataset was remarkably clean with no missing values, which streamlined the preprocessing phase. A significant outlier (-999999) was identified in the var3 feature, which was treated as a coded placeholder rather than a true numerical value.

🛠️ Data Preprocessing & Feature Engineering

To enhance data quality and focus the model on the most impactful information, the following steps were successfully implemented:

Noise Reduction: 34 constant and 29 duplicate features were removed, as they offered no predictive value.

Intelligent Feature Selection: Using a RandomForestClassifier, the feature set was strategically reduced from 306 to 153 features. This step was crucial for improving model training speed and mitigating the risk of overfitting.

💡 Model Development & Evaluation

Two state-of-the-art Gradient Boosting models were implemented and evaluated, yielding the following results:

XGBoost: Achieved a strong baseline performance with an AUC score of ≈ 0.83 after hyperparameter tuning.

LightGBM: Delivered a highly competitive performance with an AUC score of ≈ 0.836, along with a significant advantage in training speed.

🚀 Advanced Technique & Final Result: Ensembling

To maximize predictive performance, an advanced ensembling technique was deployed:

Strategy: The final prediction was calculated by taking the arithmetic mean of the probabilities generated by the optimized XGBoost and LightGBM models.

Result: This technique successfully boosted performance, achieving a final estimated AUC Score of 0.8407. This confirms that combining the "perspectives" of multiple strong models often leads to a more accurate and robust final decision.

✅ Conclusion

The project successfully delivered an effective and reliable predictive system for identifying dissatisfied customers. The excellent final score places this solution at a highly competitive level and underscores the value of a methodical data science approach—from deep data cleaning to applying advanced techniques like ensembling. The resulting model is a valuable asset for proactively identifying at-risk customers and implementing targeted retention strategies.