Project Update 2: Customer Churn Prediction in the Telecom Industry

Team Members: Seokhoon Shin, Joshua Nahm, Jiwon Choi, Shinyeong Park, Jaejoong Kim

1. Work Completed Since Update 1

Since the first update, our team has made considerable progress in enhancing model performance and refining the approach to maximize prediction accuracy and interpretability. The main tasks completed include:

1) **Hyperparameter Tuning**: We applied both grid search and random search to optimize the parameters for **Random Forest, XGBoost, CatBoost, and LightGBM** models. This tuning helped improve each model's accuracy, particularly for Random Forest and LightGBM, which demonstrated the best performance.
2) **One-Hot Encoding for Categorical Variables**: We converted Label Encodings to One-Hot Encoding for variables like ‘Contract Type’ and ‘Payment Method’, as this provided a more comprehensive representation of categorical features, especially important for models like CatBoost.
3) **Model Comparison and Ensemble Testing**: We tested several ensemble models (Random Forest, XGBoost, CatBoost, and LightGBM), each displaying unique characteristics in handling the data. Comparing their performance metrics allowed us to determine which model best aligns with our objectives of accuracy and efficiency.

2. Challenges Faced

During this phase, we encountered several issues:

1) **Class Imbalance**: While SMOTE initially improved class balance, we observed that it occasionally introduced noise, impacting certain models' test set performance. We are exploring alternatives, such as NearMiss and Tomek Links, to better handle class imbalance without compromising data integrity.
2) **Choosing the Best Model**: With multiple high-performing models, determining the final choice has been challenging. Each model provides different advantages, with Random Forest being consistent, CatBoost handling categorical features well, and LightGBM offering efficiency. Balancing accuracy, interpretability, and efficiency will guide our final decision.
3) **Interpreting Confusion Matrices**: The confusion matrices highlight each model's strengths and weaknesses in classifying churn and non-churn accurately. For example, while Random Forest and CatBoost showed relatively high accuracy, the misclassifications among other models underscored the need to choose the model that best addresses both classes.

3. Goals Before Project Submission

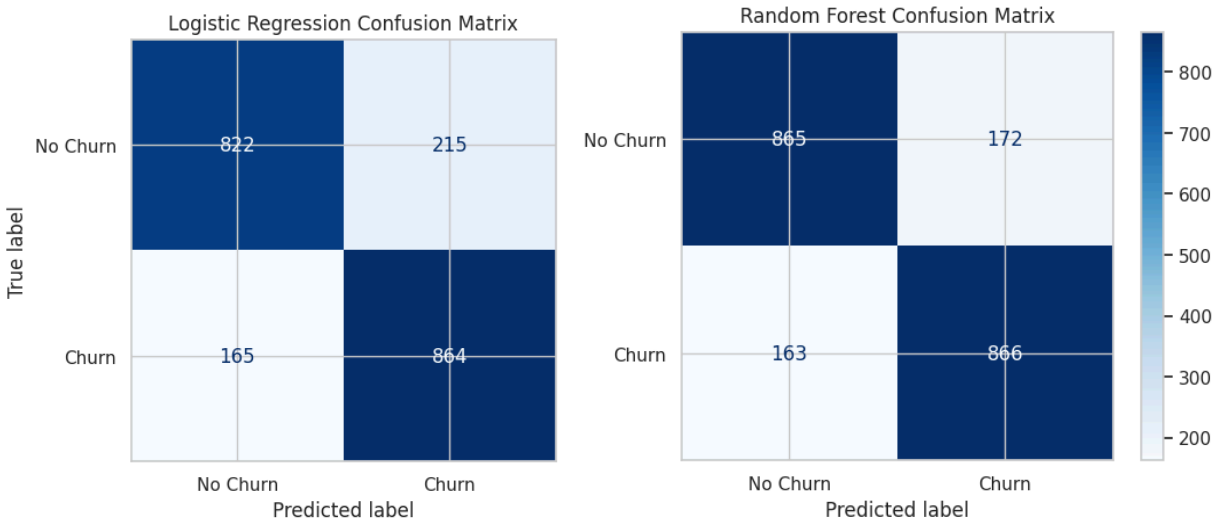
Moving forward, we aim to:

1) **Select the Best Model**: Based on accuracy, precision, recall, and F1 score, we will choose the final model for submission. Currently, Random Forest and LightGBM lead in performance.
2) **Generate Actionable Insights**: Using feature importance and interpretability analyses, we plan to identify top predictors of churn, which will be shared with stakeholders to guide retention strategies.
3) **Optimize Final Model**: We will perform final tuning and adjustments on the selected model to ensure it's robust and reliable for practical application.

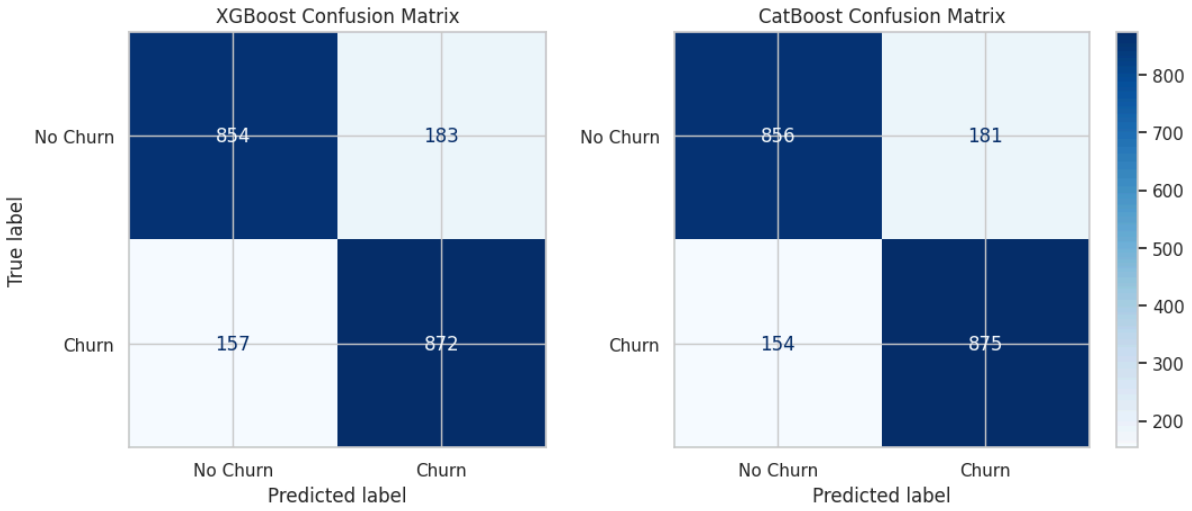
4. Insights on Model Choice and Differences

The provided confusion matrices for each model reveal critical differences and insights:

1) **Logistic Regression**: Serving as our baseline, Logistic Regression correctly classified most cases, though it misclassified a substantial portion of churn cases as non-churn. Its simplicity aids interpretability, but it lacks the nuanced predictive capability of more advanced models.
2) **Random Forest**: With high accuracy and balanced classification, Random Forest achieved robust performance, as seen in the confusion matrix. This model offers stability and interpretability through feature importance, making it a strong candidate for the final model.



3) **XGBoost**: XGBoost achieved a comparable accuracy level to Random Forest but with a slightly higher misclassification rate. Its gradient-boosting approach effectively captures complex patterns, though it's computationally heavier.
4) **CatBoost**: CatBoost handled categorical variables efficiently and delivered high accuracy. Its confusion matrix shows a balanced performance, though processing time was higher than LightGBM, leading us to consider LightGBM for efficiency-focused applications.



5) **LightGBM**: LightGBM demonstrated competitive accuracy with a faster processing time than XGBoost and CatBoost. This model is particularly useful for larger datasets or real-time applications, making it appealing for scaling purposes.