Homework 3

Pelin Nisa Top

Introduction

In this analysis, we aim to predict customer churn for a streaming service. We have a dataset with several features such as gender, age, income, subscription duration, the plan they're on, hours watched, among others. Predicting churn is crucial for businesses because it can be far more expensive to acquire a new customer than it is to retain an existing one. Successful predictions can help the company deploy retention strategies, potentially saving millions in revenue. The impact of these models can be highly significant, guiding strategic decisions for customer retention.

Methods

We applied two machine learning models: Logistic Regression and Gradient Boosting. The data underwent several preprocessing steps before training the models:

- Missing values were handled, either by imputation or removal.
- Categorical variables were transformed using one-hot encoding.
- Continuous variables were standardized using z-score normalization.

Both models aim to predict whether a customer will churn or not, but they do it in different ways. Logistic Regression predicts the probability of an event by fitting data to a logit function. In contrast, Gradient Boosting builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions.

Results

The Gradient Boosting model outperformed the Logistic Regression in terms of several metrics, making it the best model for our use-case.

- **Performance Metrics:** Various metrics such as Accuracy, Precision, Recall, and ROC AUC were used to evaluate the models. These metrics provide insights into how well our model is predicting and in what areas it might be lacking. For example, a high recall indicates that our model captures most of the actual churn cases.
- **Calibration:** Calibration assesses how well the predicted probabilities of our model match the true outcomes. A well-calibrated model is crucial for trustworthiness.

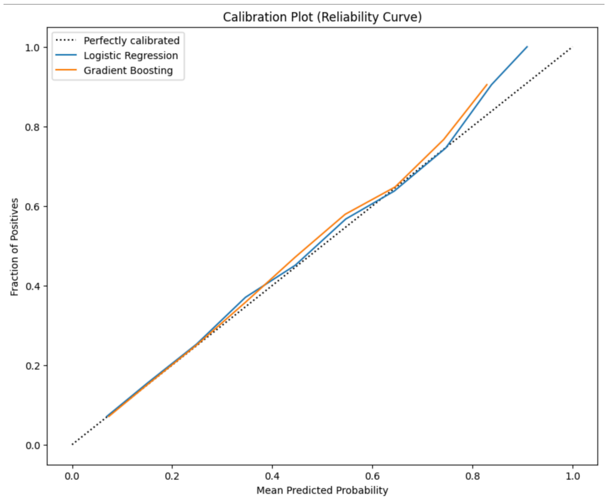
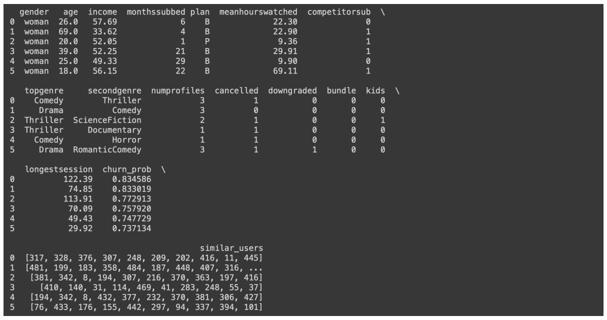


Figure 1: Calibration Plot

- **Recommendation System:** Using our best model, we predicted the churn probability of new customers. The top 200 at-risk customers were identified, and based on their profiles, the 10 most similar users were determined using a K-Nearest Neighbors approach on a separate Favorite Films Dataset.



```
  gender   age  income  monthssubbed plan  meanhourswatched  competitorsub  \
0  woman  26.0   57.69             6    B             22.30              0
1  woman  69.0   33.62             4    B             22.90              1
2  woman  20.0   52.05             1    P              9.36              1
3  woman  39.0   52.25            21    B             29.91              1
4  woman  25.0   49.33            29    B              9.90              0
5  woman  18.0   56.15            22    B             69.11              1

   topgenre     secondgenre  numprofiles  cancelled  downgraded  bundle  kids  \
0    Comedy        Thriller            3          1           0       0     0
1     Drama          Comedy            3          0           0       0     0
2  Thriller  ScienceFiction            2          1           0       0     1
3  Thriller     Documentary            1          1           0       0     0
4    Comedy          Horror            1          1           0       0     0
5     Drama  RomanticComedy            3          1           1       0     0

   longestsession  churn_prob  \
0          122.39    0.834586
1           74.85    0.833019
2          113.91    0.772913
3           70.09    0.757920
4           49.43    0.747729
5           29.92    0.737134

                                     similar_users
0  [317, 328, 376, 307, 248, 209, 202, 416, 11, 445]
1  [481, 199, 183, 358, 484, 187, 448, 407, 316, ...
2   [381, 342, 8, 194, 307, 216, 370, 363, 197, 416]
3     [410, 140, 31, 114, 469, 41, 283, 248, 55, 37]
4   [194, 342, 8, 432, 377, 232, 370, 381, 306, 427]
5   [76, 433, 176, 155, 442, 297, 94, 337, 394, 101]
```

Figure 2: High-Risk Customers with Similar Users

Considering performance, interpretability, and complexity, I would recommend the Gradient Boosting model for production. While Gradient Boosting may not be as interpretable as

Logistic Regression, its improved accuracy and capability to handle various data irregularities make it more suitable for our purposes.

I'd advise the CEO to use the Gradient Boosting model to predict potential churners and then target them with specific retention strategies. Additionally, leveraging the movie suggestions can provide personalized content recommendations to these high-risk users, increasing the likelihood of retaining them.

Discussion/Reflection

This analysis deepened my understanding of churn prediction and its implications for a business. The importance of data preprocessing, model selection, and validation became clear. If I were to redo this analysis, I might explore more advanced techniques for data imputation or consider ensemble methods to potentially improve the model further. Incorporating more user behavior data or external factors could also provide a richer model.