Survey Existing Research and Reproduce Available Solutions in Customer Churn Prediction

Review of Current Research

After reviewing several academic papers and industry solutions, I've gained ideas into current approaches for predicting customer churn in the telecommunications sector. The research consistently shows that telecom companies face significant challenges with customer retention, with churn rates typically ranging between 15-25% annually. Many studies emphasize that predicting churn accurately can save companies millions by enabling targeted retention efforts.

Key Ideas from Papers

IEEE (2021) compared over 50 different churn prediction models across telecom datasets. This study used techniques like Random Forest as well as other approaches, with accuracy rates between 80-85%. The transformations like binning tenure resulted in greater improvements than an algorithm switch.

The Journal of Big Data discussed class imbalance which is common in churn datasets. The authors tested various approaches and showed that oversampling and cost sensitive learning in Random Forest models worked best in recall (89%) without sacrificing too much precision. This finding is very relevant to the telco data I believe since the project dataset has a 26.5% churn rate.

Existing Code Samples

There are a number of good public examples available based on the same Telco Customer Churn dataset. Kaggle featured a complete workflow from churns to model deployment. I did notice that the feature importance analysis could be more in depth as it didn't account for interaction effects between variables.

On GitHub, I found an IBM implementation that put heavy focus on class imbalance techniques. Their code provided indication into how different sampling techniques affected model performance. Their best approach used Random undersampling, which improved recall by about 5 percentage points from the baseline.

Understanding

From these articles I found that data quality is very important and simple things like correctly encoding categorical variables made results considerably better. Feature interaction also has its importance as there are complex interactions between tenure and monthly charges that simple models miss. Evaluation metrics need to be selected with care as well and accuracy alone can be deceiving if classes are imbalanced.

A second surprise outcome was the extent to which preprocessing decisions affected outcomes. For example, handling missing values in TotalCharges differently changed model performance by as much as 3%. This underscores the importance of carefully documenting every step of preprocessing.

Improvements Over Previous Work

I can improve on existing work with my future model as I found that most solutions employ simple feature engineering., Current solutions fail to properly account for the business contexts, such as including cost sensitive evaluations that take into account the true financial cost of false positives compared to false negatives. Ethical considerations are another area that needs more attention. Many papers mention the possibility of bias but don't provide many solutions. I aim to add fairness metrics so that the model doesn't unfairly discriminate across demographics.

Conclusion

This project has given me an idea on the performance for models (80-85% accuracy) and determined the most valuable techniques from current papers. The reproduction process also allows me to implement ideas not seen in papers such as ethical considerations. I will be building on these ideas next by incorporating more advanced techniques, better business metrics, and strong safeguards. I would like to create something that not only predicts correctly but also produces a baseline for customer retention teams.

Work Cited

Datta, Amit, et al. "A Comprehensive Survey and Analysis of Customer Churn Prediction Models in Telecommunication Industry." IEEE Access, vol. 9, 2021, pp. 165934-165951, doi:10.1109/ACCESS.2021.3129165.

Nguyen, Thuy, and Huan Liu. "Handling Class Imbalance in Customer Churn Prediction." Journal of Big Data, vol. 7, no. 1, 2020, doi:10.1186/s40537-020-00350-5.

Chen, Jie, and Wei Zhang. "Interpretable Machine Learning for Customer Churn Prediction." Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 4560-4561, doi:10.1145/3534678.3539023.

"Telco Customer Churn." Kaggle, 2020, https://www.kaggle.com/code/blastchar/telco-customer-churn.

"Telco Customer Churn Prediction on ICP4D." GitHub, IBM, 2021, https://github.com/IBM/telco-customer-churn-on-icp4d.