Predictive Modeling Case Study Report: Customer Churn

By Trey Poore

Introduction

This report reviews a machine learning case study that developed predictive models for customer churn in the telecommunications industry (Khanzhina et al., 2023). The study demonstrates how large-scale customer data can be transformed into actionable insights through systematic preprocessing, model experimentation, and post-deployment monitoring. It also highlights ethical practices and lessons aligned with the machine learning development lifecycle and iterative learning emphasized in this course.

Data Preprocessing

The project began with 84,000 historical customer accounts for training and 250,000 live accounts for testing. The team defined churn as a binary Yes/No outcome, a critical decision that prevented data leakage. Results showed 73% churned while 27% did not, providing a clear baseline.

Initially, the dataset contained 1,400 features, many of which were irrelevant or incomplete. Through exploratory data analysis (EDA) and principal component analysis (PCA), features were narrowed to 340 key variables. PCA highlighted the attributes with the strongest weights, while features with high null-value rates were eliminated to maintain data quality (Jolliffe & Cadima, 2016).

EDA also uncovered distinct behavioral patterns between churners and non-churners, strengthening the dataset's interpretability. These steps ensured that the final feature set was both statistically sound and practically meaningful.

Model Choice

The researchers adopted a two-phase modeling strategy:

- **Random Forest (RF):** Selected as a starting model due to its robustness with tabular data, ability to capture non-linear relationships, and provision of feature importance rankings (Breiman, 2001). RF acted as a benchmark and helped identify the most relevant features.
- **Light Gradient Boosting Machine (LGBM):** After refining the feature set, the team transitioned to LGBM to boost predictive power. LGBM excels at handling large datasets with reduced feature sets, efficiently uncovering subtle patterns and offering superior scalability (Ke et al., 2017).

This progression reflects best practices: beginning with an interpretable baseline, then advancing to a more complex, high-performance model.

Performance Metrics

The model was evaluated across multiple performance metrics:

- **Accuracy:** 90%, showing strong overall correctness.
- **F1 Score:** 93%, reflecting balance between precision and recall, especially important because missing churners is costlier than false positives.
- **Area Under the Curve (AUC):** 88%, demonstrating effective discrimination between churners and non-churners (Bradley, 1997).

When deployed, the model achieved an **80% hit rate** in real-world testing. Using multiple metrics ensured that evaluation reflected business priorities rather than relying on accuracy alone.

Post-Deployment Plan

The team recognized that **customer behavior evolves**, requiring ongoing monitoring and retraining to combat **data drift**. Their strategy included:

- **Monthly dashboards** to monitor accuracy, F1, and AUC.
- **Quarterly retraining** with updated data.
- **Annual audits** to review assumptions, feature relevance, and potential biases.

This iterative maintenance ensured the model remained both accurate and ethical over time.

Ethical and Privacy Considerations

To comply with privacy standards, all direct personal identifiers were removed. Instead, the model relied on anonymized behavioral and contractual features. This protected customer data while maintaining predictive strength.

Ethically, the reliance on behavioral rather than demographic variables helped mitigate risks of discriminatory outcomes. The integration of ongoing monitoring further ensured that bias or drift could be detected and corrected over time.

Critical Reflection

This case study strongly aligns with the **machine learning development lifecycle**:

1. **Problem Definition:** Clear target (churn: Yes/No).
2. **Data Preparation:** Rigorous preprocessing, feature selection, and PCA.
3. **Modeling:** Use of RF as a baseline, followed by LGBM for scalability and subtlety.
4. **Evaluation:** Multi-metric validation aligned with business needs.
5. **Deployment and Monitoring:** Continuous oversight and iterative updates.

The study also reflects this class's emphasis on **iterative learning**—from initial data exploration through post-deployment retraining.

For future projects, several insights stand out:

- **Define the target clearly.** Precise definitions prevent errors and leakage.
- **Combine domain knowledge with statistics.** Each strengthens the other.
- **Prioritize feature quality.** Clean, meaningful variables outperform complex models built on noisy data.
- **Select the right metrics.** F1 and AUC often capture business priorities better than accuracy.
- **Plan for change.** Models must evolve with customer behavior, making monitoring and retraining essential.

Conversely, risks to avoid include retaining weak features with excessive null values, over-relying on accuracy, and assuming deployment marks the end of the process.

Conclusion

This case study demonstrates how predictive modeling can guide churn management by pairing rigorous preprocessing with strategic model selection. The combined use of RF and LGBM reflected methodological progression, while multi-metric evaluation ensured robust validation. Ethical safeguards and monitoring reinforced trustworthiness and long-term viability.

The study illustrates how predictive modeling aligns with lifecycle values: defining the problem, preprocessing, modeling, evaluation, and continuous monitoring. For my own work, the clearest lessons are the value of **clear target definitions, quality feature selection, and iterative maintenance**. These principles will guide the development of predictive models that remain accurate, ethical, and business-relevant over time.

References

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Khanzhina, N., Fedorova, A., Shestakov, A., Karanina, E., & Rogov, S. (2023). Development of predictive models of customer churn in the telecommunications industry. *PeerJ Computer Science, 9*, e1179. https://doi.org/10.7717/peerj-cs.1179