Main Notebook: Final_notebook.IPYNB

Machine Learning Modeling

The machine learning algorithms alone are not enough to make an efficient model that a Bank can rely on. So, some other methods should be taken into considerations.

Working with data, in real-life data are imbalanced. For example, predicting whether a patient has cancer or not, fraud detections, detecting defects, etc. These data are going to be imbalanced.

Learning what best method to use, **Cost-sensitive learning using class weights** stands out because it does not change our data. It instead pays much attention to one class we are interested in during training, the churn.

SMOTE is popular but it has negative side because it changes the data. It makes over sampling by producing synthetic examples. Just blindly copying current sample and create new samples. And from the what modeling done before, it has out weighted by **Cost-sensitive learning using class weights.**

Performance Metrics and Business Interest

From the visual like confusion matrix, the thing is clear: the 0.4 threshold isn't just a technical pick but rather a golden spot.

The reason is because that's where we make the most money: $201,600 in net value, which is the highest we can get. It's better spot where we're not missing too many churners, catching about 76% of them, but we're also not wasting budget on people who aren't actually leaving, we have only 537 false positives.

This means, we only have to reach out to 846 customers, which is way less work than blasting 1,900 or more at lower thresholds. Actually, our recall drops a bit from 0.1 threshold, but that drop is totally worth it, we save way more in costs than we lose in missed churners.

Go lower to 0.3, and we start paying for a ton of extra false positives for very little gain. Go higher to 0.5, and we let too many real churners walk away untouched. So, **0.4** keeps the company efficient, focused, and profitable. It's the number that turns our churn model from a warning system into a real money-maker.

Business Impact and Key Churn Insights

Losing customers is expensive, and once a customer leaves, it is often too late to recover them. These points show why banks need to spot at-risk customers early and why using ***recall-focused***, **cost-sensitive** models makes practical sense for churn prediction.

- *Customer retention is cheaper:* Retaining an existing customer typically costs 3–5 times less than acquiring a new one, making churn prevention financially critical.
- *Small churn reductions matter*: A 5% reduction in churn can increase profits by 25–40%, especially in banking where customer lifetime value is high.
- *Churn is imbalanced:* In most banks, only 15–25% of customers churn, which is why accuracy alone is misleading and recall-focused models are preferred.
- *Missed churners are costly:* Failing to identify a churner can result in the loss of years of future revenue, while a false positive usually costs only a promotion or incentive.
- *Early action works:* Customers identified and contacted before disengagement are significantly more likely to stay than those contacted after service usage drops.

Category of high Churners

- High balance customers leave because they don't feel special.
  - Bank should give them their own personal banker and better interest rates.
- Customers in their 40s to 70s leave because bank is not helping with what comes next in life, like retirement.
  - Bank need to call them and offer real retirement and estate planning help.
- People leave when bank only talk to them when there's a problem.
  - Check in with them regularly, before they think of leaving, just to see how they're doing.
- They leave because another bank offers them a better deal.
  - We should proactively give our best customers better rates before they have to ask, so they never feel the need to look elsewhere.

Appendix:

Live model Link: https://ml-customer-churn-pred.streamlit.app/