Project Proposal: Implementing a Churn Model on simulated employee data

Claudine Chen

In our vibrant entrepreneurial world, companies and organizations need to constantly keep their competitive edge, whether it's to retain customers or the best employees. A way to monitor and react to what is happening in these contexts is to collect and analyze data on human behavior and demographics, looking for which factors have the greatest impact on the desired result. With a model of customer behavior, a company can make informed decisions on how to provide a better experience and to optimize income. Machine learning and statistics offer tools to do exactly that, especially on a large scale.

The specific study of customers cancelling subscriptions or not returning using data is call churn modeling; churn is the overturning of something, and in this case means when a customer is lost. Alternatively this is also called customer lifetime value modeling.

Companies with subscription models, like Netflix, will want to keep members subscribed for as long as possible, and will look for key factors for why subscribers cancel their memberships, so that they can actively prevent this. Online retailers like Amazon don't have monthly subscription fees, but they similarly will be interested in keeping customers returning to Amazon to buy things. Companies that don't have contractual agreements with customers will define churn in other ways, such as by time between visits. Regardless of the definition, when looking for measures that can optimize income, companies will want to implement changes that have the highest probability of making an impact.

Making changes comes with an inherent cost, and proposals that are not based on evidence could potentially be detrimental. Relying on data for insights and inspiration for improvements is likely to be the most cost effective and efficient way to grow a company.

The data I will use to explore churn modeling is employee information, offered by Medium.com on Kaggle (https://www.kaggle.com/ludobenistant/hr-analytics), with an eye on understanding employee churn - why are they leaving when they do. It contains 15000 simulated employees, and features include employee satisfaction level, their last evaluation, the number of projects, their average monthly hours, time spent at the company, whether they have had a work accident, whether they have had a promotion in the last 5 years, their department, and salary level.

To gain insights, we want to understand what information, out of all the data we have access to, has a significant impact, positive or negative, on if the employee stays or leaves. The specific steps that need to be taken to do this are as follows:

1. Look at correlations between individual columns of data and the churn variable using logistic regression and decision trees.
2. Look for interactions between columns and if they impact the churn variable.
3. Apply other model frameworks (random forest)
4. Apply model validation, performance metrics, and classification metrics

The deliverables will be:

1. all code (in Python) used to conduct the analysis, in the form of Jupyter notebooks.
2. A final report detailing my process and the outcomes of the analysis
3. A slide deck summarizing the project for the general public

All deliverables will be publicly available in my Github account.