Class imbalance: Through the Eyes of our Working Example

Our Story

The data science team at AAVAiL has a number of datasets to analyze to support its core product, the audio/video service. These data grow and change over time. For example, all the data that are used to study customer retention will only grow larger and more complex over time. There will be additional services, new markets and other features that enrich AAVAiL's data for customer retention, but one thing that is unlikely to change is that the data will generally be *unbalanced*.

If one or more classes in your data are underrepresented you could be introducing bias into your models. Working with imbalanced data may also lead your analysis to incorrect or misleading conclusions. Understanding this will help you keep class balance and bias in perspective as you proceed through the data analysis pipeline.

THE DESIGN THINKING PROCESS

During the *Define* and *Ideate* phases of the design thinking process, there will be a lot of work accomplished that is related to the identification of data sources that can be used to create machine learning models. You'll be building and testing numerous pipelines, comparing the performance of each one with the others, and utilizing the best ones.

Utilizing them for what? Supporting the Define and Ideate processes. During these phases of the design thinking process the most important rule is to use your creativity to help develop a view of the challenges facing the company, and also to develop possible solutions for those challenges. It is impossible to know ahead of time what those challenges are. As a result, you will need the ability to effectively iterate using pipelines to do things like compare strategies that deal with imbalanced classes. Iteration and the ability to concisely summarize the finding is critical to your ability to support your team during the Define and Ideate stages of design thinking.