Predicting Churn for KKbOX – Spark MLlib, H2O and Amazon ML

Maroon Team 4 – Bekzat Alish (alish004@umn.edu), Bryce Quesnel (quesn012@umn.edu), Ishwarya Ravikumar (ravik020@umn.edu), Karthik Narapareddy (narap004@umn.edu), Shuyu Sui (sui00002@umn.edu)

KKbox

KKbox is one of the leading music streaming service providers in Southeast Asia market. In hopes of understanding customer churn, KKbox is looking to predict users who are likely to leave in the next 30 days

The KKbox churn dataset includes data about historic churn, purchase and transaction data, user profile data and most importantly about 31 gigabytes of customer user logs

Data transformation

We utilized Amazon S3 and EMR to manipulate and transform our data for churn prediction analysis

Modeling

For PySpark ML and H2O, we tried both logistic regression and random forest classification models. For AML, only logistic regression is available

Evaluation

We did a 70%, 30% train-test split and ran the model on the test dataset for evaluation

Parameters	APACHE Spark		H_2O.ai
Performance	★ ★ ★	★	★ ★ ★
Flexibility	★ ★ ★	★	★ ★
Execution	★ ★	★	★ ★ ★
Learning Curve / Simplicity	★	★ ★ ★	★ ★
Final Score	★ ★ ½	★	★ ★ ½

Key Takeaway

PySpark ML

PySpark and Spark ML provides a powerful machine learning solution by combining the computational power of distributed data and flexibility of libraries such as pandas that are run locally. Although the learning curve is high, SparkML provided the largest range of algorithms which are more customizable than in other platforms.

H2O

Using H2O, multiple advanced models such as Deep Learning, Random Forests could be run on big data in a relatively easy manner. It is highly automated - auto-detecting variable encoding, multi-platform support and multi-language such as Python and Java. On the other hand, simpler models could be more suitable for beginners.

Amazon ML

Inspite of support of AWS, AML is highly limited by the number of models available, making it a less than ideal choice for machine learning on big datasets