Classification Module
Feb 2022

Predicting Customer Churn at PDAX

Prepared by: Li Sun

Introduction – Motivation & Problem Definition

- **Motivation**
  - PDAX is a Crypto Exchange, allowing Filipino customers to buy and sell Crypto Assets on their software platform
  - Despite PDAX's growth, they are not aware of the amount of customer churn in their customer base
  - Every month a significant amount of their customer who traded on PDAX's platform last quarter will churn and not trade at all.

- **Problem Definition -**
  - How can we predict which of PDAX's customers will churn next month?

- **Dataset**
  - Data is taken internally from PDAX's AWS & Metabase
  - Around 26k lines/events of data
  - Each event represent an active user who's traded in the last 10 weeks.
  - 8 columns of static data (age, gender, address, etc.)
  - 2 columns of dynamic data (ratio of trades / not trades + net monthly deposit)

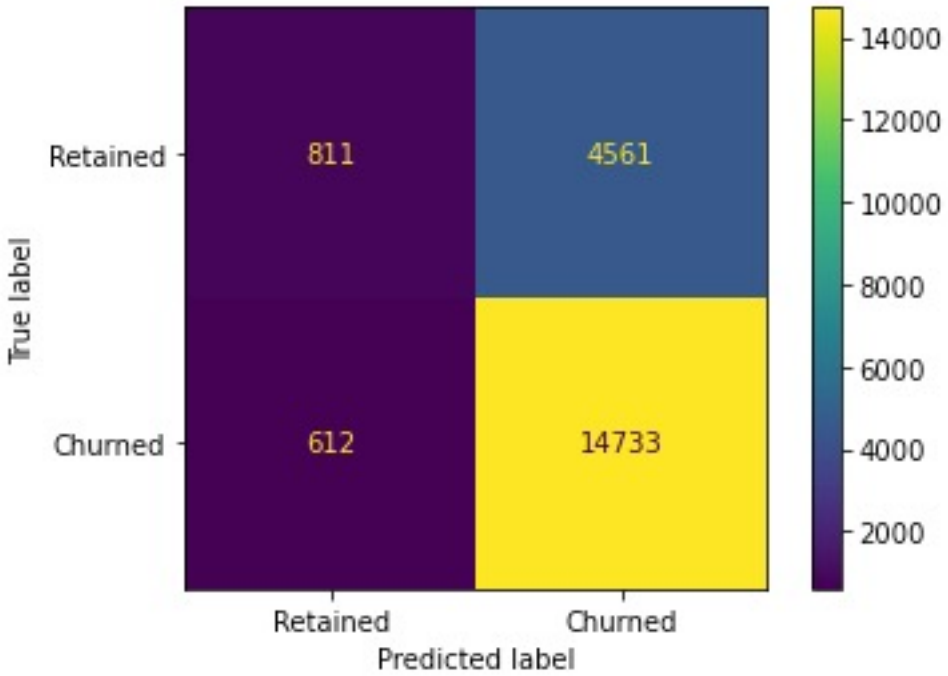
Methodology – Tools & Metrics

- **Tools –**
  - Pandas and numpy for data manipulation
  - Matplotlib and Seaborn for data visualization
  - Sklearn for classification models
  - SQL for data extraction

- **Metrics –**
  - Accuracy
  - Precision
  - Recall
  - F1 Score

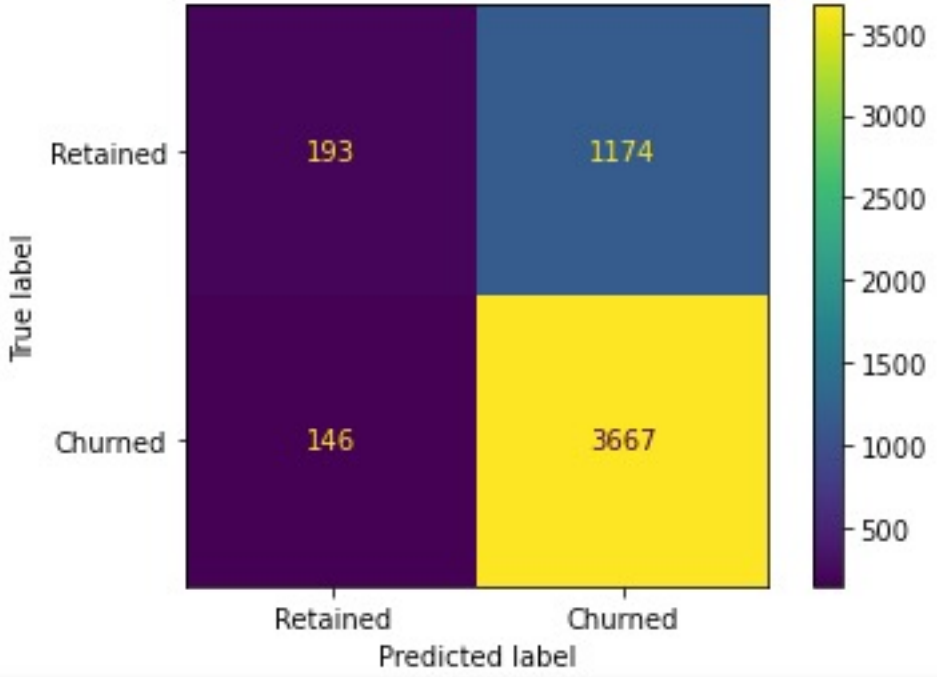
Results – Logistic Regression

- **Logistic Regression:**
  - Training Set:



	precision	recall	f1-score	support
False	0.57	0.15	0.24	5372
True	0.76	0.96	0.85	15345

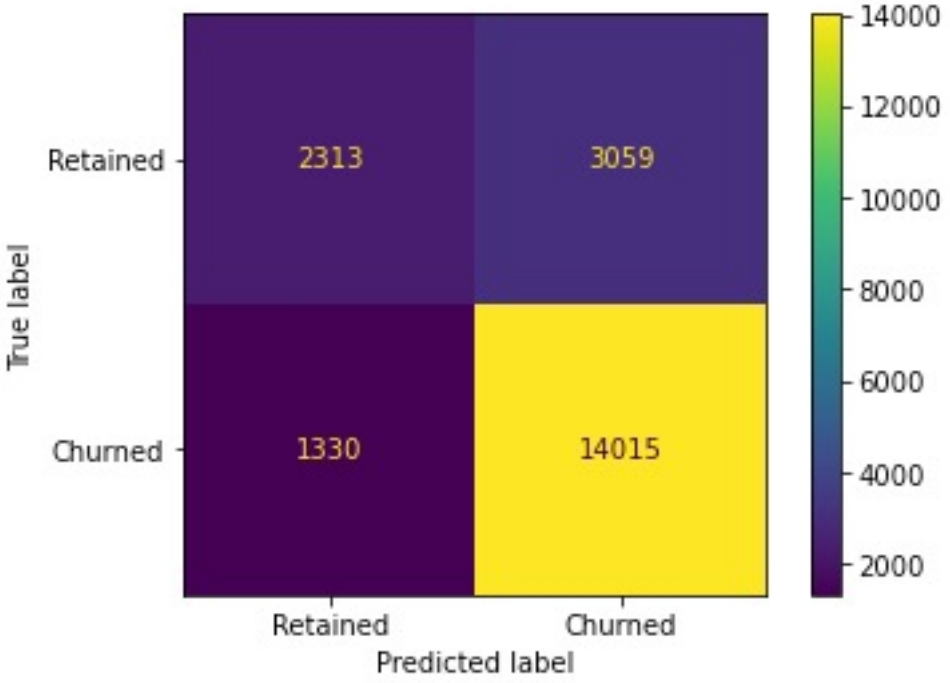
- **Logistic Regression:**
  - Test Set:



	precision	recall	f1-score	support
False	0.57	0.14	0.23	1367
True	0.76	0.96	0.85	3813

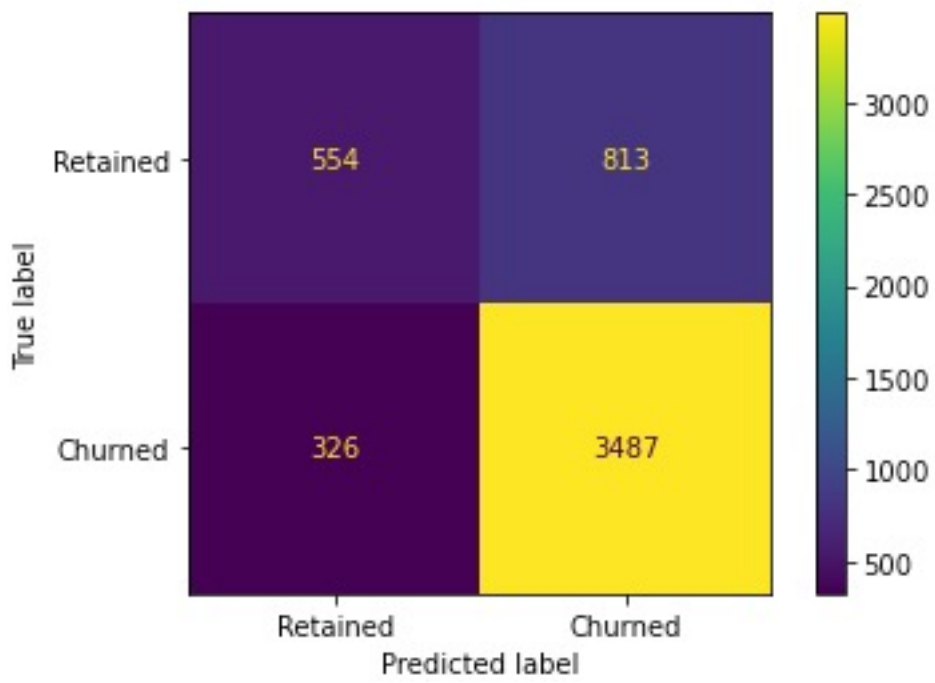
Results – Random Forest

- **Random Forest:**
  - Training Set:



	precision	recall	f1-score	support
False	0.63	0.43	0.51	5372
True	0.82	0.91	0.86	15345

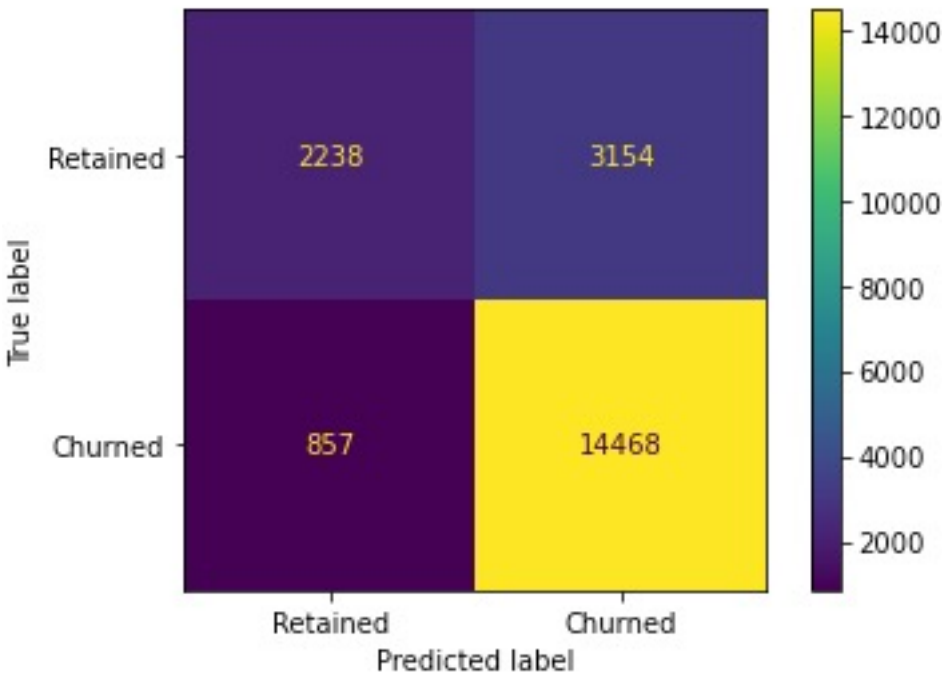
- **Random Forest:**
  - Test Set:



	precision	recall	f1-score	support
False	0.63	0.41	0.49	1367
True	0.81	0.91	0.86	3813

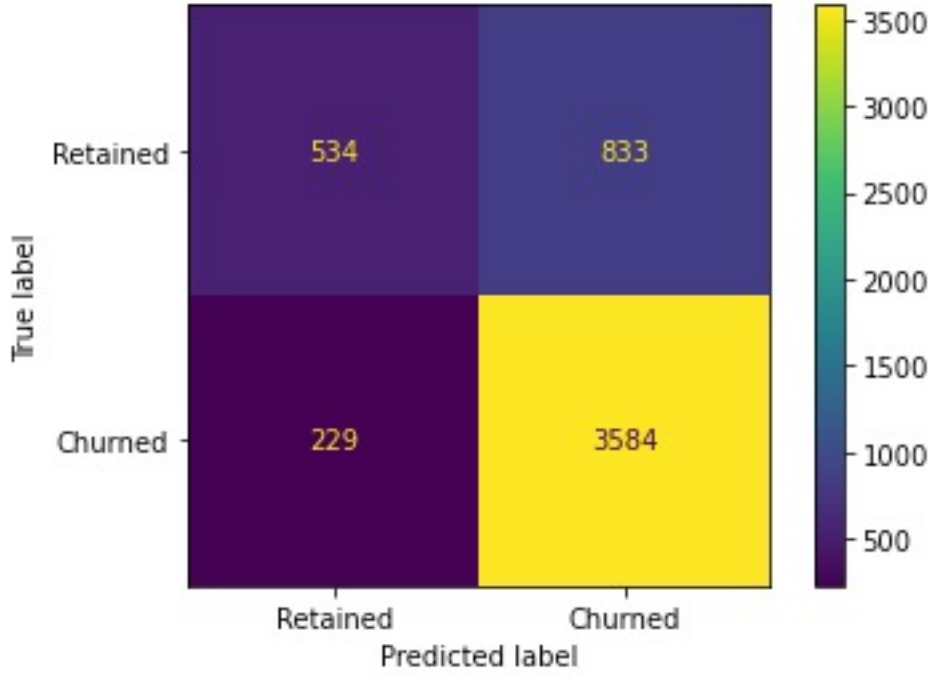
Results – XGBoost

- **XGBoost:**
  - Training Set:



	precision	recall	f1-score	support
False	0.72	0.42	0.53	5392
True	0.82	0.94	0.88	15325

- **XGBoost:**
  - Test Set:



	precision	recall	f1-score	support
False	0.70	0.39	0.50	1367
True	0.81	0.94	0.87	3813

Conclusion + Further Study

- **Conclusion:**
  - XGBoost is the best model in terms of performance with F1 score of 88%
  - However if interpretability is important then Logistic Regression may be best model to use.

- **Next Steps:**
  - Backtest the model with prior month's data. Backtest for at least 3 months
  - Consider redefining churn based on different time periods (e.g. 2 months)
  - Apply Shapely Values to demystify "blackbox" and explain XGBoost Models

Appendix