ATB customer delinquency analysis

Brett Soprovich, Ramon Zamilpa, Jiaqiang Yi

ATB Financial

Content

1. Problem exploration
2. Methods
3. Insights
4. Next Step

1. Problem exploration

Background:

ATB faces the challenge of delinquency from customers.

Objective:

We want to forecast the potential delinquency issues so that ATB can prepare for this

Data set:

Label: Status

Feature: Count and value of transactions in different channels

2. Methods

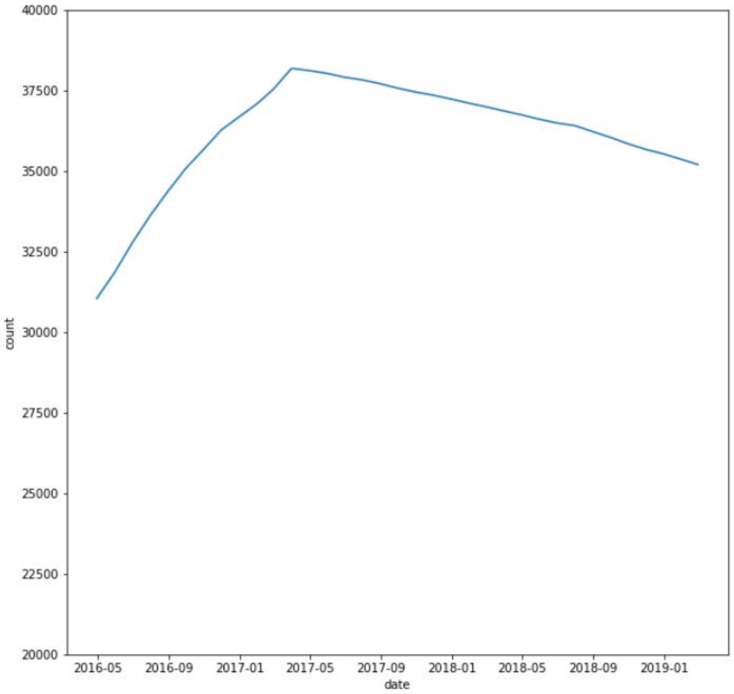
1. Exploratory data analysis (EDA)
    a. Time series analysis
2. Feature selection
3. Machine learning prediction

3. Insights - EDA

Label analysis

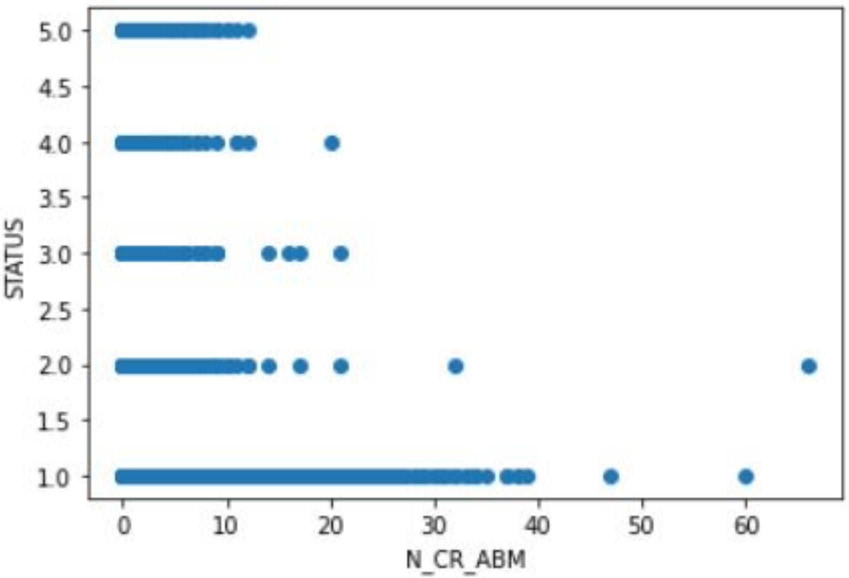
```
1    1227777
2      20011
5       8303
3       7221
4       3480
Name: STATUS, dtype: int64
```

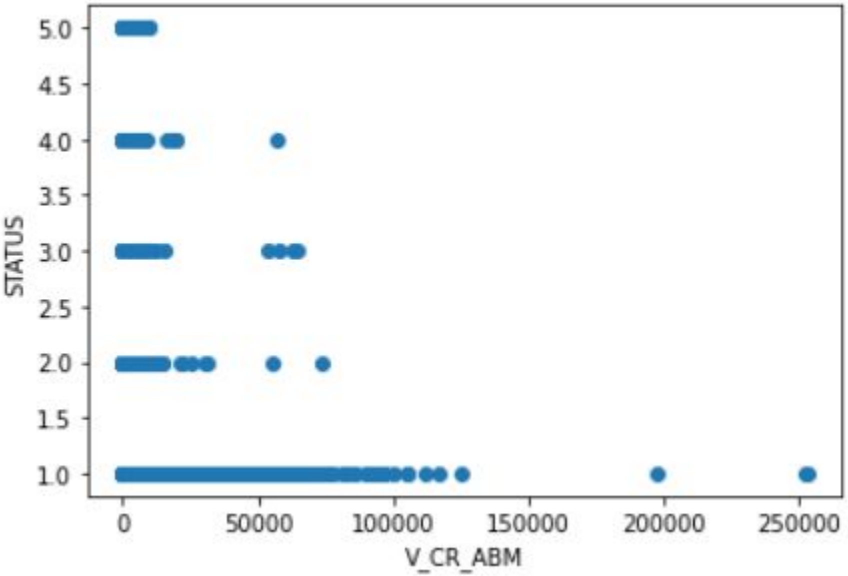
Customer number: **41326**



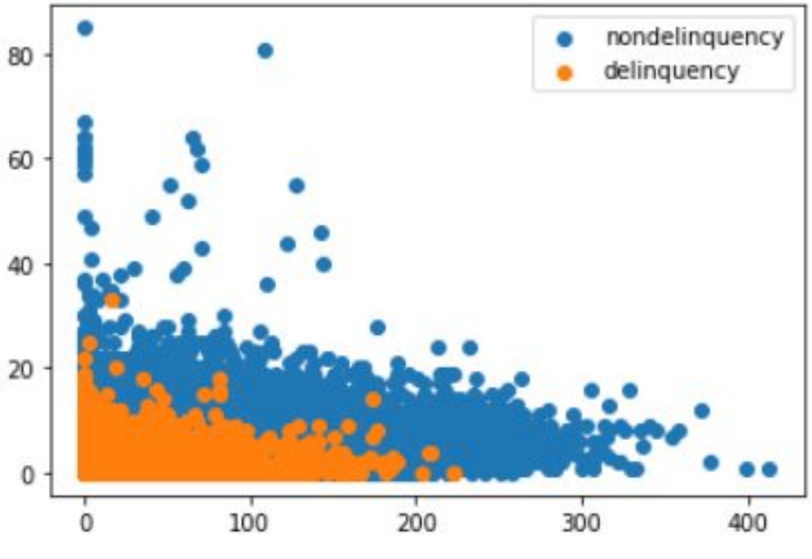
3. Insight - EDA

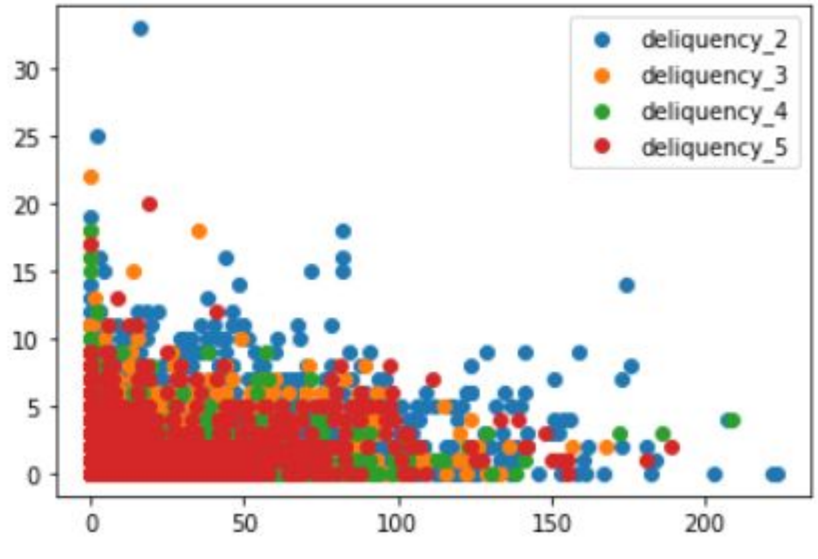
Feature analysis





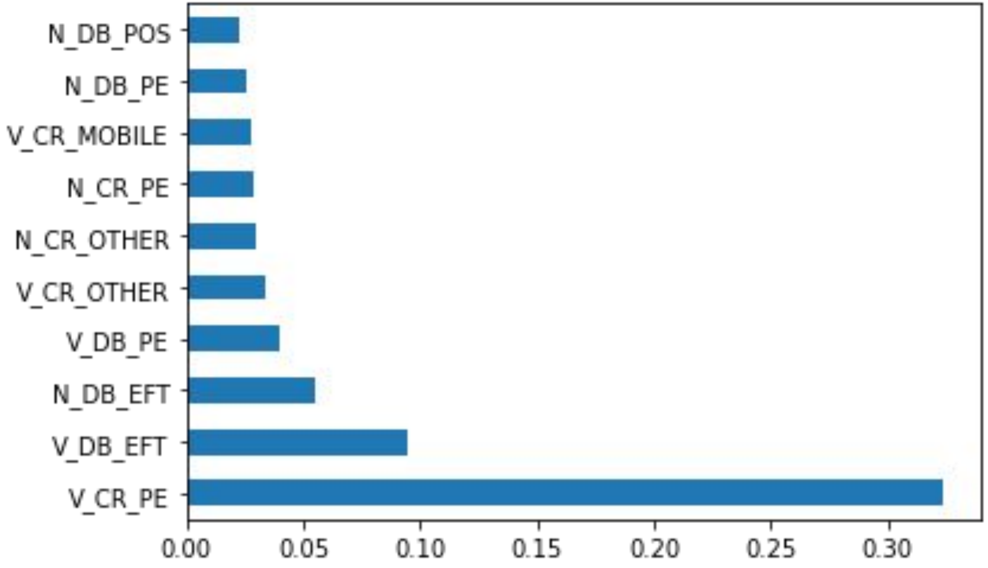
3. Insight - Feature selection

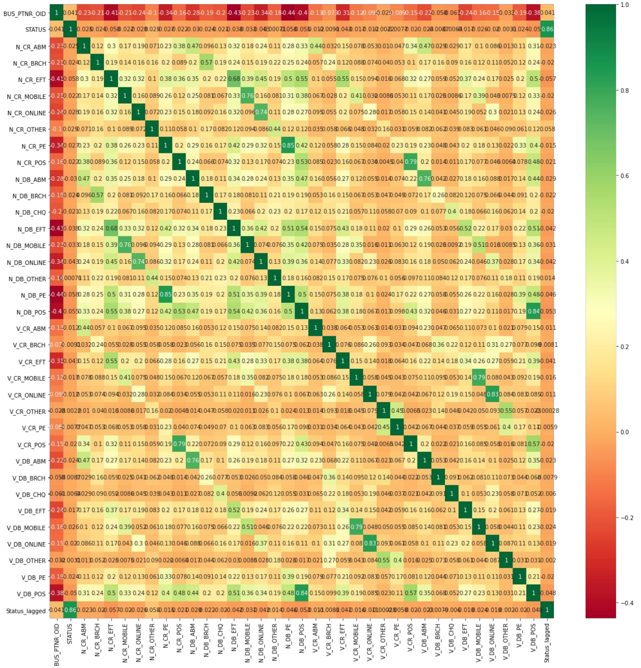




3. Insight - Feature selection

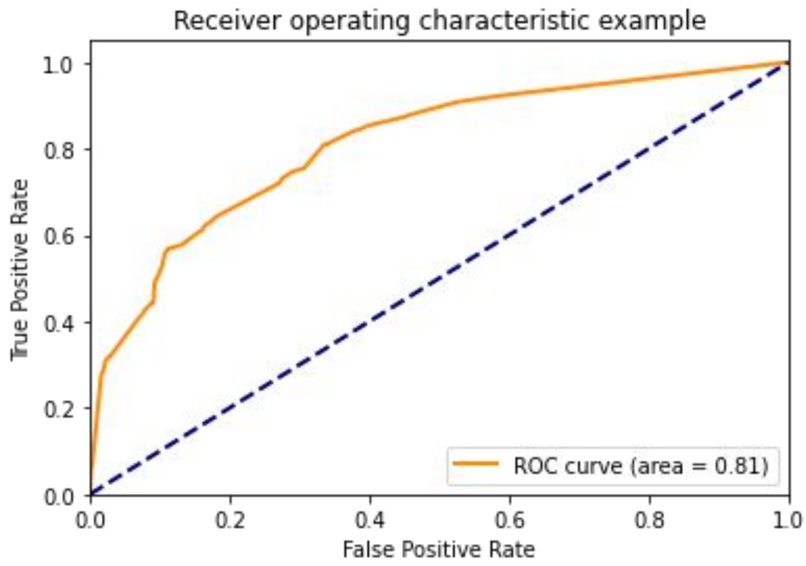
Extra Trees Classifier and Heatmap considerations





3. Insight - Random Forest

- First Model tried was Random Forest
- Used selected features on the dataset
- Initial Model didn't perform overly well
  - Dataset has very few example of defaulters
  - Skewed data over predicts "non-defaulters"
    - Predicted non-defaulter 100% of time
  - Utilized the predictions on a probability
    - Helps encourage model to predict default
- Overall Recall of model: 0.028
- Overall Precision of model: 0.548
- Room for improvement and tweaking
  - Add more features
  - Rolling window for customer ID for longer term trends
  - More in-depth analysis historical data of defaulters for trends on those specific customers



3. Insight - Random Forest

Initial Random Forest:

- Equal probability to default/non-defaulter
  - Recall Score: 0.00
  - Precision Score: 0.00
- Probabilities > 0.14 assigned to non-defaulters
  - Recall Score: 0.03
  - Precision Score: 0.55
- Probabilities > 0.1 assigned to non-defaulters
  - Recall Score: 0.06
  - Precision Score: 0.31
- Probabilities > 0.05 assigned to non-defaulters
  - Recall Score: 0.49
  - Precision Score: 0.09
- Probabilities > 0.01 assigned to non-defaulters
  - Recall Score: 0.49
  - Precision Score: 0.09

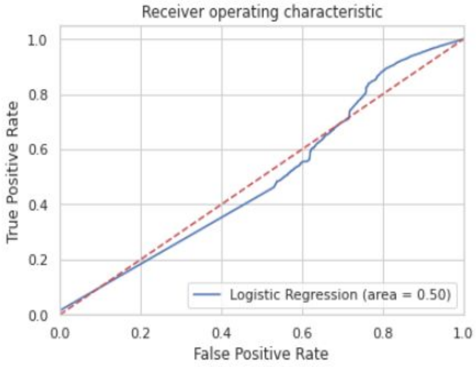
3. Insight - Logistic Regression

- Second Model tried was Logistic Regression
- Room for improvement and tweaking

Spliting training data by Dates

Overall Recall of model: `0.0015`
Overall Precision of model: `0.0353`

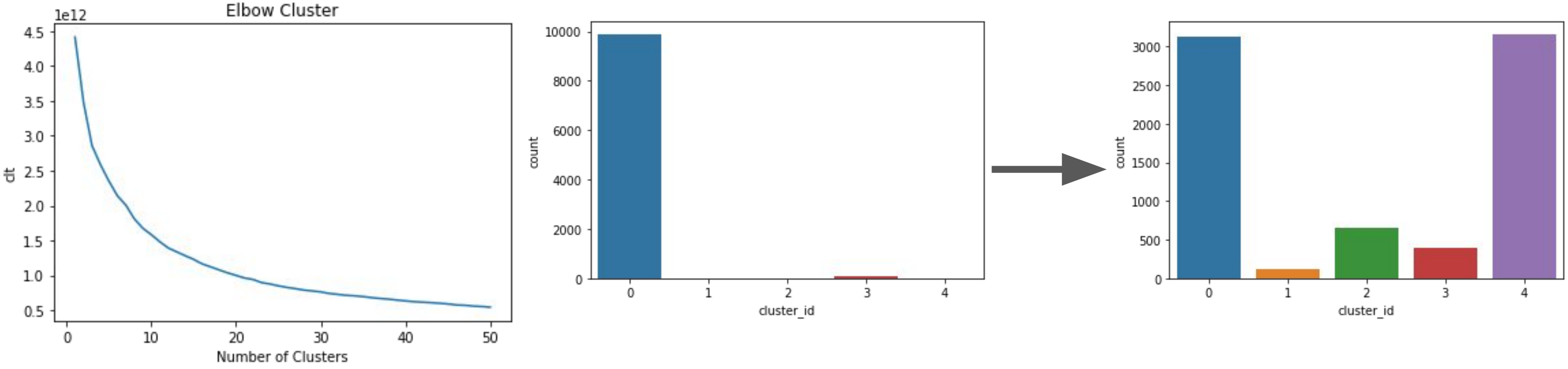
Spliting training data by Customers



Overall Recall of model: `0.7045`
Overall Precision of model: `0.0344`

3. - Insight - Initial Clustering

Initial clustering K Means attempt on features indicated further need for outlier removal as the clusters were initially ineffective as clearly splitting the customers apart.



4. Next step

1. Cluster of different customer groups
2. Apply other machine learning models to see the performance
3. Implement further feature engineering on the data to explore the effect
4. Investigate more in depth of the features of business partner ID’s that have defaulted for further clues
5. Try a rolling exploratory data of customers to have more robust considerations