Kin Analytics Hackathon

Denis Trosman - 13/11/2023

Index

- Objective
- Data
- Analysis and cleansing
- Feature engineering
- Modelling and insights

Objective

Develop a two-year churn predictive model for Kin Security.

Kin Security, a successful security service company, faces challenges with clients canceling "Kin Safety" before two years, resulting in significant fixed costs. Seeking a solution, they aim to develop a classification model using credit bureau data to predict early cancellations. The model will inform targeted campaigns to reward long-term clients and penalize early cancellations based on credit scores and bank transaction data.

Data

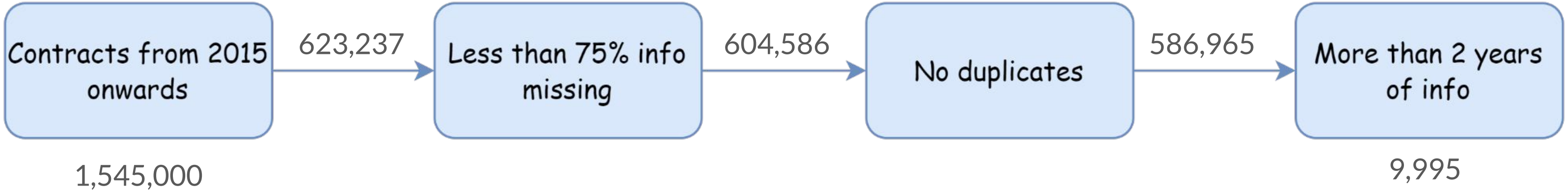
Four databases were delivered on 30/11/2019

- *Clients*: 1,545,000 x 10
- *Transactions*: 7,500,000 x 3
- *Products*: 3,739,192 x 3
- *Credit score*: 19,500,000 x 3

All of them were used to construct a solid dataset to develop a churn predictive model.

Analysis and cleansing

After applying the corresponding filters, the final dataset consisted of 9,995 observations.


Contracts from 2015 onwards
623,237
Less than 75% info missing
604,586
No duplicates
586,965
More than 2 years of info
1,545,000
9,995

*Date filters were built using *application_date*

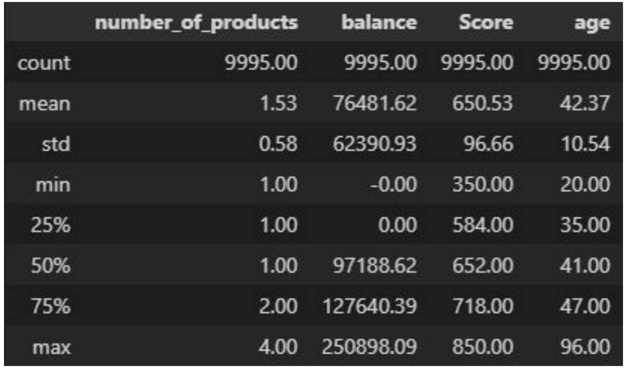
*For Italy, there are application dates after 2019, same than other countries. This should be talked with the client to better understand what they meant by that message, and will not be tackled in this exercise.

Feature engineering

Four new variables were created to build a better model

- **Number_of_products**: grouping by customerid, counting the number of products.
- **Balance**: pivoting by customerid, calculated as the sum of all transactions.
- **Score**: added credit score using the year-month as key to merge to clients data frame.
- **Age**: calculated using birth date

The construction of them can be seen at the attached notebook / shared repository.



	number_of_products	balance	Score	age
count	9995.00	9995.00	9995.00	9995.00
mean	1.53	76481.62	650.53	42.37
std	0.58	62390.93	96.66	10.54
min	1.00	-0.00	350.00	20.00
25%	1.00	0.00	584.00	35.00
50%	1.00	97188.62	652.00	41.00
75%	2.00	127640.39	718.00	47.00
max	4.00	250898.09	850.00	96.00

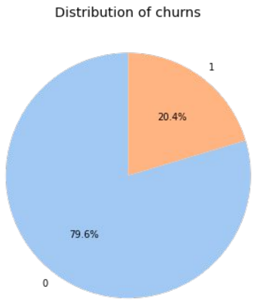
Modelling and insights

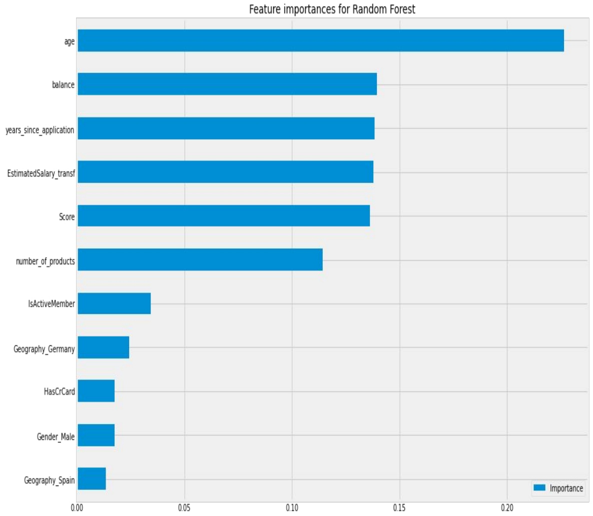
Churns were defined as those with an exit date after less than 2 years from application, and a final model showed an AUC of 0.85 in cross validation.

- After defining the target, 20% were churners (pie plot). Using AUC for this reason.
- After transformations and scaling, an 80-20 split was generated to train a model.
- Baseline model showed a poor performance of 50-50 accuracy.
- Optimized **Random Forest** showed a great AUC of **0.85** in cross validation and **0.72** in testing.
  - Random forest works greatly for classification problems, and helps avoiding overfitting.

Feature engineered variables helped the model gain a higher score.

Future steps: improve the model recall to have a better prediction of positive cases.


Distribution of churns


Feature importances for Random Forest
age
balance
years_since_application
EstimatedSalary_transf
Score
number_of_products
IsActiveMember
Geography_Germany
HasCrCard
Gender_Male
Geography_Spain
Importance