Machine Learning

CHURN PREDICTION – CAR INSURANCE

ELENI TAPTA

Churn Prediction

Churn rate is the number of individuals or items moving out of a collective group over a specific period of time

Statement of the Problem

The problem appeared as a ***frequent and increasing trend of customers switching companies*** for their car insurance, in order to take advantage of a competitors' offer.

Churn models identify the ***customers with a high likelihood of leaving the company***. These customers cancel their contract, the policy, in order to benefit from better conditions (a lower premium) with another company.

For the company, churn prediction is one of the fundamental issues in the ***prevention of revenue loss*** and it is therefore an important way to improve competitiveness.

Feature Selection – Overview*

Policy Holder

- Age
- Customer satisfaction
- Gender
- Highest premium paid Life events
- Lifetime
- Location identifier (ZIP code)
- Maximum duration of all policies owned by the same customer
- Network attributes
- Number of policies in force (in all lines of insurance)
- Returned Customer
- Segment selected by the company
- Sum of premiums (both canceled and in force)
- Occupation

Policy

- Brand credibility
- Change in premium
- Contracted care
- Discounts applied or Bonus-Malus level
- Guarantees
- Payment method
- Premium price
- Product Usage
- Type of Insurance
- Loss Ratio
- **Policy status (ex. Canceled)** → Our target variable

Customer/Company Variables

- Customers mention that they are going to switch
- Duration of current insurance contract
- Elapsed time since the last complaint
- Handling time of authorizations and declarations
- Number of complaints
- Number of contact moments
- Number of declarations
- Number of times subscribed
- Outstanding charges
- Reaction on marketing actions
- Type of contact (email, call, etc.)

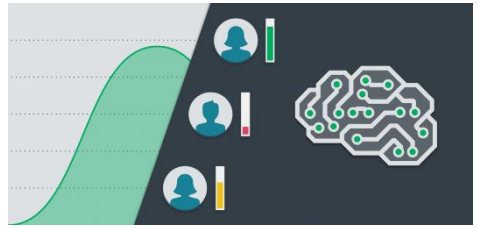
* Details provided in readme.md (feature selection table)

Feature Engineering

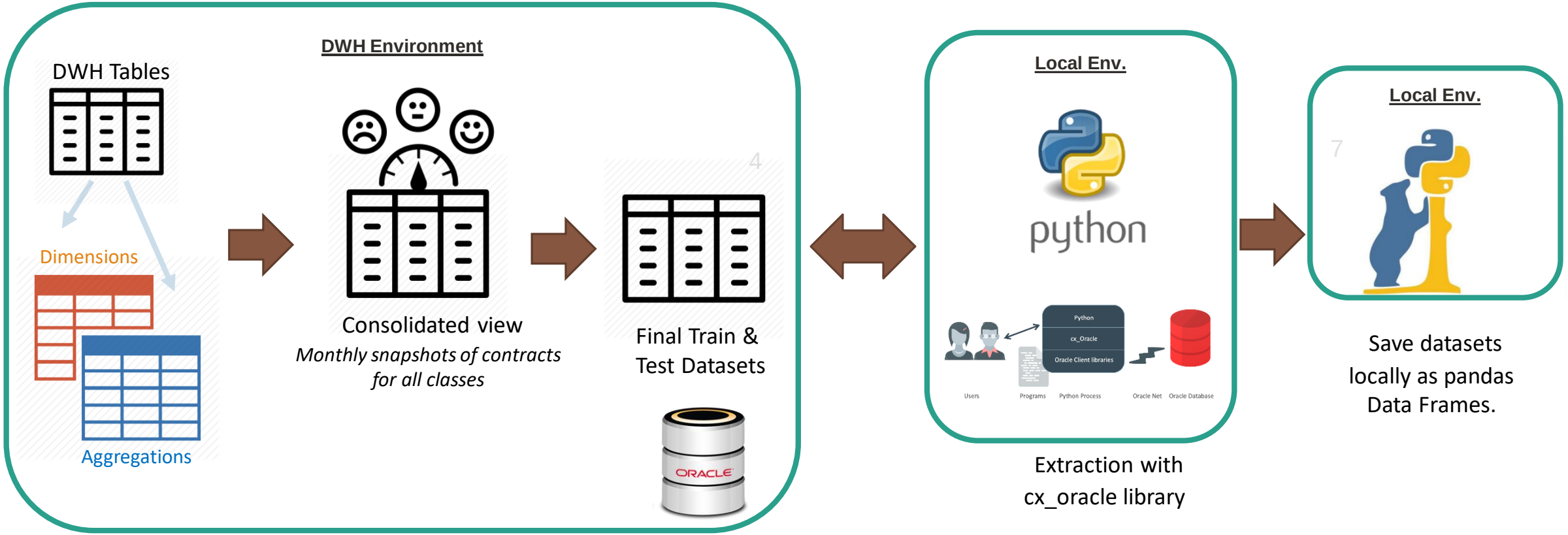
Various techniques have been applied on a database level, such as:
Imputation, Handling Outliers, One-hot Encoding, Aggregation Operations

	INS_COV_KEY	INS_COV_CD	INS_COV_NAME
1	13888	028	ΠΡΟΣΩΠΙΚΟ ΑΤΥΧΗΜΑ ΟΔΗΓΟΥ 3000 ΕΥΡΩ-ΑΕΡΟΜΕΤΑΦΟΡΑ
2	13920	065	ΠΡΟΣΤΑΣΙΑ BONUS-MALUS ΚΑΙ ΚΑΛΥΨΗ Υ.Ζ. ΑΝΑΣΦΑΛ.
3	14076	108	ΘΡΑΥΣΗ ΚΡΥΣΤΑΛΛΩΝ
4	14000	077	AUTOCARE
5	14008	068	ΘΡΑΥΣΗ ΚΡΥΣΤΑΛΛΩΝ ΧΩΡΙΣ ΑΠΑΛΛΑΓΗ
6	14046	011	ΥΛΙΚΕΣ ΖΗΜΙΕΣ ΤΡΙΤΩΝ (Ν.489/76) (Ανά Ατύχημα)
7	14206	002	Α. Ε. ΕΝΤΟΣ ΦΥΛΑΣ. ΧΩΡΩΝ, FERRY ΚΑΙ ΕΞΩΤ.
8	14407	109	ΘΡΑΥΣΗ ΚΡΥΣΤΑΛΛΩΝ ΑΠΟ ΤΡ. ΑΤΥΧΗΜΑ / ΑΠΟΠ. ΚΛΟΠΗΣ
9	14687	001	ΣΩΜΑΤΙΚΕΣ ΒΛΑΒΕΣ ΤΡΙΤΩΝ ΚΑΙ ΕΠΙΒΑΙΝΟΝΤΩΝ(Ν.489/76)
10	14611	070	ΡΥΜΟΥΛΚΗΣΗ ΣΥΝΕΠΕΙΑ ΑΤΥΧΗΜΑΤΟΣ (ΤΗΛ.2111075709)
11	14663	078	ΘΡΑΥΣΗ ΚΡΥΣΤΑΛΛΩΝ ΜΕ ΑΠΑΛ. 10% ΚΑΙ ΕΛΑΧ. 30 ΕΥΡΩ
12	14802	026	ΠΡΟΓΡΑΜΜΑ EASY COVER 2 (ΕΠΑΡΧΙΑΣ)
13	15010	027	ΠΡΟΓΡΑΜΜΑ EASY COVER 3 (ΛΟΙΠΗ ΕΛΛΑΔΑ)
14	15104	048	ΘΡΑΥΣΗ ΚΡΥΣΤΑΛΛΩΝ (ΑΠΑΛ. 200 ΕΥΡΩ)

	CONTACT_NO	CURRENT_POLICY_STATUS	COV10_PREMIUM	COV11_PREMIUM	COV4_PREMIUM	COV12_PREMIUM	COV13_PREMIUM	COV14_PREMIUM	COV15_PREMIUM	COV16_PREMIUM
1	*02****7	1	1,14	17,9	1,04	1,78	0,95	1,64	0	3,04
2	*0**7264*	0	0	0	0	0	0	0	0	0
3	*0370*607	0	0	0	0	0	0	0	0	0
4	*0443607	0	0	0	0	0	0	0	0	0
5	*0***252*	0	0	0	0	0	0	0	0	0
6	*025073*7	0	0	0	0	0	0	0	0	0
7	*0376**7*	1	1,71	0	5,88	0	5,13	0	0	2,86
8	*0*5520**	0	0	0	0	0	0	0	0	0
9	*0***657*	1	0,9	17	3,12	2,03	0,9	1,87	0	2,88
10	*032635*3	1	3,09	25,77	4,65	0	0	0	0	0
11	*023**2**	0	0	0	0	0	0	0	0	0
12	*026**335	0	0	0	0	0	0	0	0	0



Data Acquisition



Data Acquisition

- ✓ **Data Extraction logic:** Exclude repetitive or misguiding records from the dataset.
  - ✓ For example, the company is only interested in contracts that are about to be renewed.
  - ✓ Exclude contracts with 0 duration or ‘NA’ (-1) values

```
SELECT *
FROM ML_CHURN_FINAL_DATA
WHERE
MO_KEY>20200301 AND  MO_KEY<20210401 AND --Only take into account last years' data.
TOTAL_YEAR_INSURED_SYMB>=0 AND TOTAL_YEAR_INSURED_SYMB<=10 AND --Only a specific insurance period is valid according to the business logic specified by the company.
POLICY_STATUS=1  AND --Take into account only active snapshots of contracts.
CURRENT_POLICY_STATUS!=-1 AND --Do not take into account null values for our target variable.
IS_RENEWAL=1 AND  -- Only take into account monthly snapshots corresponding to the month of renewal for each contract.
SYMB_DURATION!=0  -- Do not take into account contract duration outliers.
AND INS_PKG_KEY!=-1 -- Do not take into account null attributes.
AND AGENT_CTGR_KEY!=-1-- Do not take into account null attributes.
AND TAXK_INC_ZONE_KEY!=-1-- Do not take into account null attributes.
```

Data Acquisition

- ✓ **Resampling:** The majority of our records represent the Positive class (Not Churn), ~80% of the dataset), while the Negative class (Churn) is only present in ~20% of the dataset.

  - ✓ **Avoid repetitive records:** Monthly snapshots → multiple views of the same attributes, if the customer never churns or makes any significant changes in the contract rules. Only distinct features made it to the final dataset (extracted via Python script)

  - ✓ **Over-sampling under-represented class:** Enrich the dataset with views of records on customers that churned in past months (not included in the original training dataset).

Data Acquisition

CATEGORY	TRAIN	TEST
DATASET NAME	ET_ML_CHURN_FI_TRAIN	ET_ML_CHURN_FI_TEST
NO OF RECORDS	884.258	85.522
CHURN RECORDS	567.184	74.089
NOT CHURN RECORDS	356.453	11.433
NO OF FEATURES AVAILABLE	62	62

Feature Selection

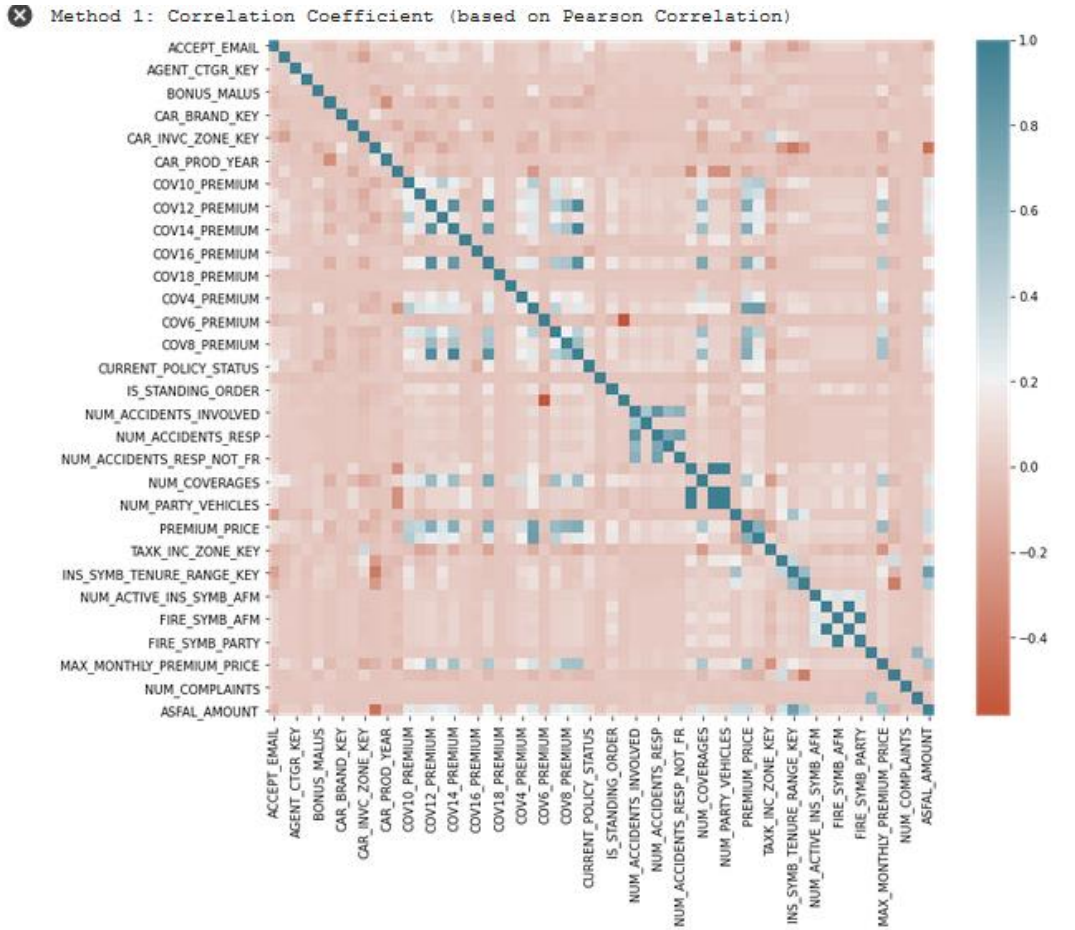
Our **final datasets contain over 60 attributes** in order to provide an overview of the most useful features for churn prediction.

Benefits of performing feature selection:

- **Reduces Overfitting**: Less redundant data means less opportunity to make decisions based on noise.
- **Improves Accuracy**: Less misleading data means modeling accuracy improves.
- **Reduces Training Time**: fewer data points reduce algorithm complexity and algorithms train faster.

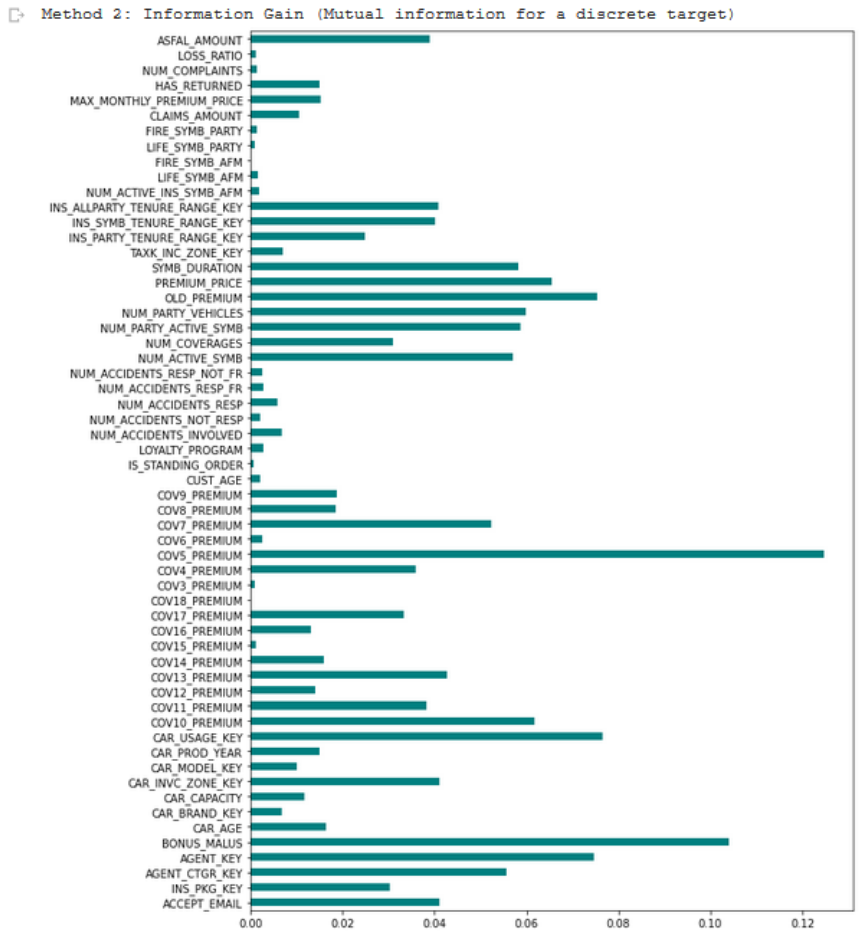
Pearson Correlation Matrix

- Gives an overview of how the features are related to each other or the target variable.



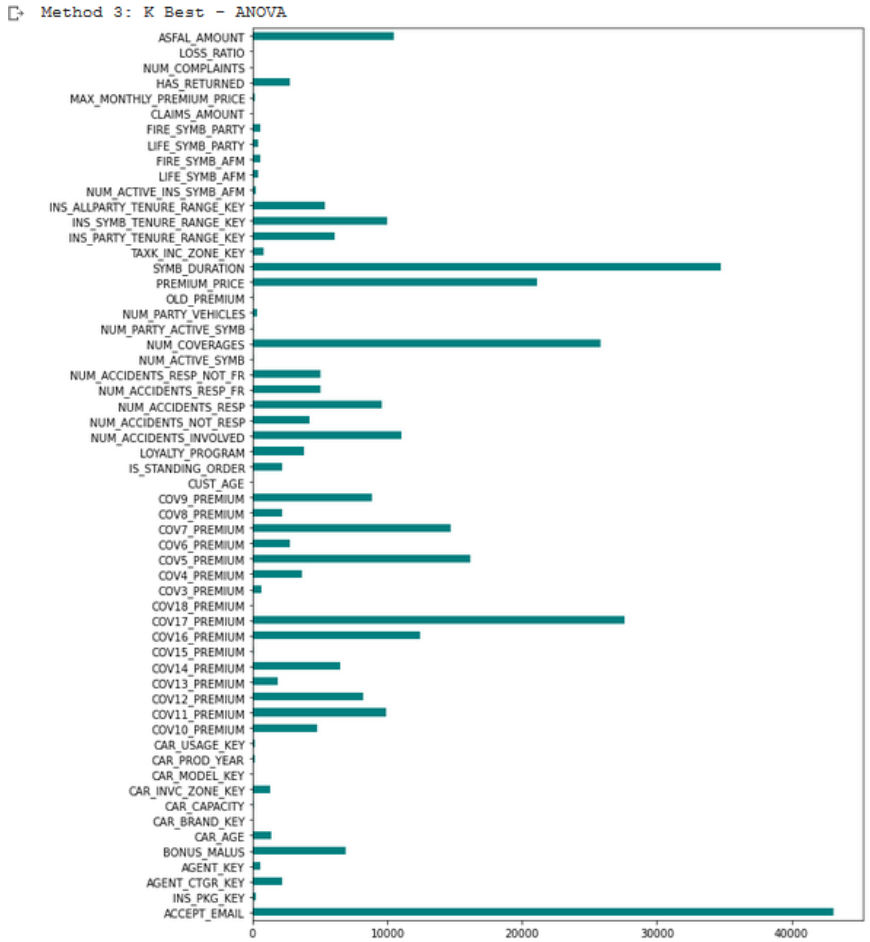
Mutual Info

- Calculates mutual information value for each independent variable with respect to the dependent variable.
- Select those with most information gain.



K-Best

- A more general approach compared to the above-mentioned methods → allows the selection of the function to use in feature selection.
- Use mutual information based feature selection, which can capture any kind of statistical dependency.



Final feature selection

- Combine scores of the methods presented above and conclude to 15 features to be used by the algorithm.
- Later, during model training these features will also be evaluated using a wrapper-style feature selection method (RFE).

Model Selection & Evaluation

- Decision Trees
- Random Forests
- Logistic Regression
- Linear Discriminant Analysis
- k-Nearest Neighbors *
- Support Vector Machine *

K-Fold Cross Validation

Compared using different scoring functions such as accuracy, balanced accuracy, precision, recall, F1 score.

* In the final version several models were eliminated either due to performance issues in terms of training time (SVM, k-NN)

Model Selection & Evaluation

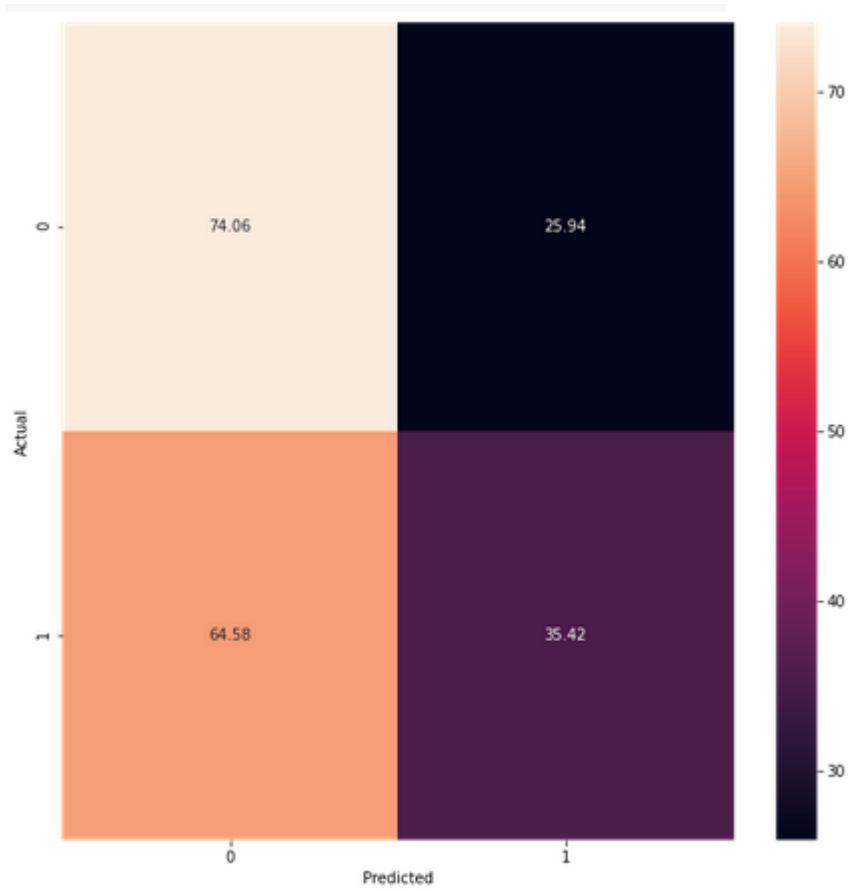
- Final results → select 3 best models
- Each model is tuned (RFE feature wrapper)
- Fit the training dataset and save model

	accuracy	balanced_accuracy	precision	recall	f1_score	roc_auc_score
LRB	0.673997	0.669209	0.587028	0.644503	0.614416	0.669209
LR	0.674990	0.638415	0.637026	0.449866	0.527329	0.638415
LDA	0.674113	0.636897	0.636983	0.445016	0.523958	0.636897
RF	0.729063	0.717831	0.667394	0.656613	0.661571	0.717551
DTC	0.696793	0.683716	0.624010	0.619598	0.622309	0.684129

Model Test & Predictions

- Model A (Random Forest)

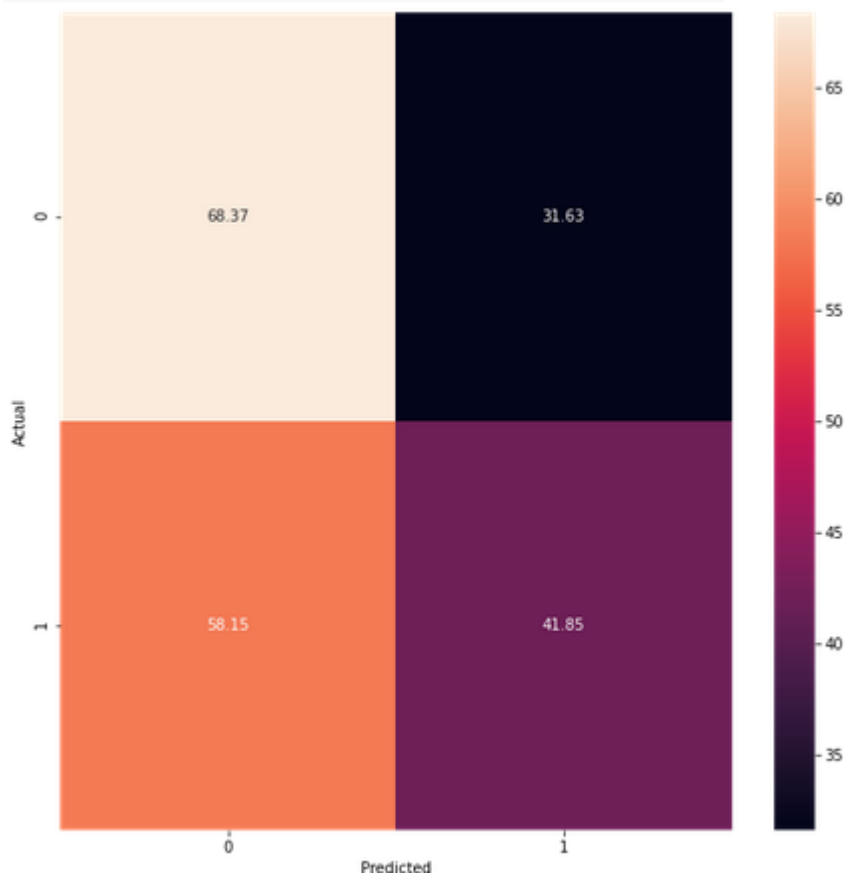
	precision	recall	f1-score	support
0	0.150248	0.740611	0.249815	11423.000000
1	0.898545	0.354196	0.508103	74089.000000
accuracy	0.405814	0.405814	0.405814	0.405814
macro avg	0.524396	0.547403	0.378959	85512.000000
weighted avg	0.798585	0.405814	0.473600	85512.000000



Model Test & Predictions

- Model B (Decision Tree Classifier)

	precision	recall	f1-score	support
0	0.153462	0.683708	0.250662	11423.000000
1	0.895638	0.418510	0.570459	74089.000000
accuracy	0.453936	0.453936	0.453936	0.453936
macro avg	0.524550	0.551109	0.410560	85512.000000
weighted avg	0.796496	0.453936	0.527739	85512.000000



Model Test & Predictions

- Model C (Logistic Regression - Balanced)

	precision	recall	f1-score	support
0	0.124495	0.789810	0.215086	11423.000000
1	0.815917	0.143638	0.244273	74089.000000
accuracy	0.229956	0.229956	0.229956	0.229956
macro avg	0.470206	0.466724	0.229680	85512.000000
weighted avg	0.723554	0.229956	0.240374	85512.000000