GERMAN CREDIT CASE:
A CLASSIFICATION SOLUTION

BACKGROUND

With the high costs that banks bear due to delinquency and default rates, accuracy on the predictions are needed.

This is the best solution based on the accuracy of the model and the economic benefits derived.

PROBLEM FORMULATION

- Credit risk evaluation using German Credit dataset results in net loss for the bank.
- Dataset - Credit Rating
  - 300 client records with GOOD CREDIT
  - 700 client records with BAD CREDIT
- Sample of 1000 clients awarded loans at a net loss of ($80,000)
- Credit approval cost of false positives ($500) outweighs benefit of true positives $100

OBJECTIVE

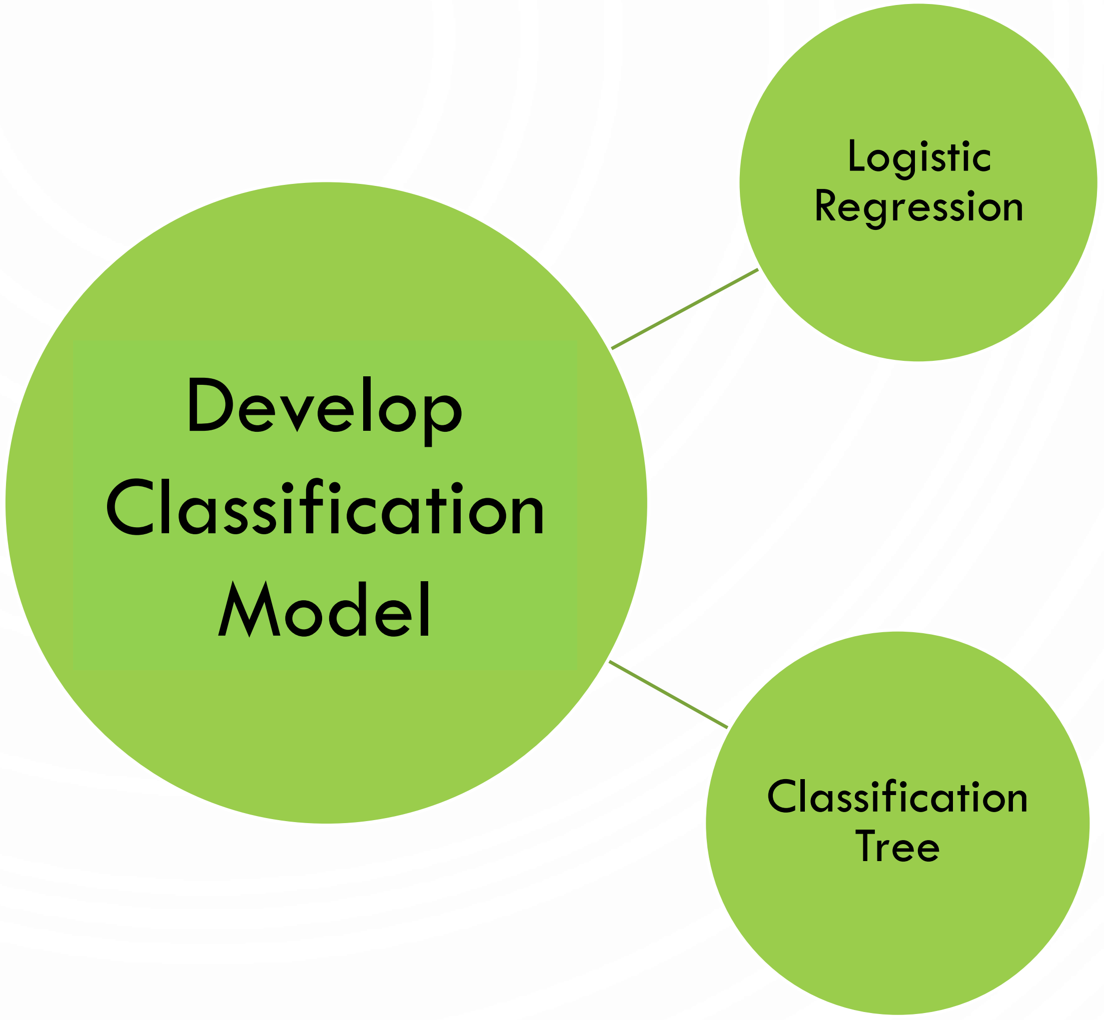
Develop a model that will minimize the cost of lending.

Classifying credit applicants as:

Good Credit

or

Bad Credit



THE DATASET EXPLORATION

RECORDS

Total 1000 RECORD

With 30 numeric Predictor VARIABLES

A binary RESPONSE variable rating

Good credit (300 Clients)

Bad Credit (700 Clients)

VARIABLES

DATA TYPE – All NUMERIC

Categorical

Ordinal

Binary

MODEL PREPARATION

Data was CLEANED

Data was EXPLORED

ACCESSED FOR

MISSING values

UNIQUE values

MODEL DEVELOPMENT

Logistic Regression Model 1 Confusion Matrix

```
Confusion Matrix (Accuracy 0.7917)

       Prediction
Actual   0   1
     0  99  86
     1  39 376
Confusion Matrix (Accuracy 0.7500)

       Prediction
Actual   0   1
     0  51  64
     1  36 249
```

26 out of 30 Predictors used

Logistic Regression Model 2 Confusion Matrix

```
Confusion Matrix (Accuracy 0.7983)

       Prediction
Actual   0   1
     0 107  78
     1  43 372
Confusion Matrix (Accuracy 0.7575)

       Prediction
Actual   0   1
     0  53  62
     1  35 250
```

All predictors used except RENT

Logistic Regression Model 3 Confusion Matrix

```
Confusion Matrix (Accuracy 0.7983)

       Prediction
Actual   0   1
     0 105  80
     1  41 374
Confusion Matrix (Accuracy 0.7650)

       Prediction
Actual   0   1
     0  53  62
     1  32 253
```

Backward Elimination reduced to 28 (AMOUNT and RADIO_TV are eliminated)

Classification Tree Model 1 Confusion Matrix

```
Confusion Matrix (Accuracy 0.7450)

       Prediction
Actual   0   1
     0 102  83
     1  70 345
Confusion Matrix (Accuracy 0.7575)

       Prediction
Actual   0   1
     0  65  50
     1  47 238
```

All predictors used

Classification Tree Model 2 Confusion Matrix

```
Confusion Matrix (Accuracy 0.7450)

       Prediction
Actual   0   1
     0 102  83
     1  70 345
Confusion Matrix (Accuracy 0.7575)

       Prediction
Actual   0   1
     0  65  50
     1  47 238
```

Variables' importance Method reduces features to 5

MODEL SELECTION

Logistic Model has higher accuracy, i.e., better ability to make true predictions

Classification Tree Model has more false predictions but lower number of False Positive

Which Model to roll out?

		Logistic Regression Model 3 (Accuracy 76.5%)		*Classification Tree Model 2 (Accuracy 75.75%)*	
		Predicted		*Predicted*	
		Bad Credit	*Good Credit*	*Bad Credit*	*Good Credit*
Actual	*Bad Credit*	True Negative 53	False Positive 62	True Negative 65	False Positive 50
	Good Credit	False Negative 32	True Positive 253	False Negative 47	True Positive 238

MODEL SELECTION

	True Positive	False Positive		
	Profit	Cost	Net Profit/Loss	Reduction in Loss
No Model	700 x 100 = 70,000	300 x 500 = 150,000	-80,000	-
Logistic Regression Model	627 x 100 = 62,700	142 x 500 = 71,000	-8,300	71,700
Classification Tree Model	583 x 100 = 58,300	133 x 500 = 66,500	-8,200	71,800

Best Economic Value

RECOMMENDED MODEL

Classification Rules:

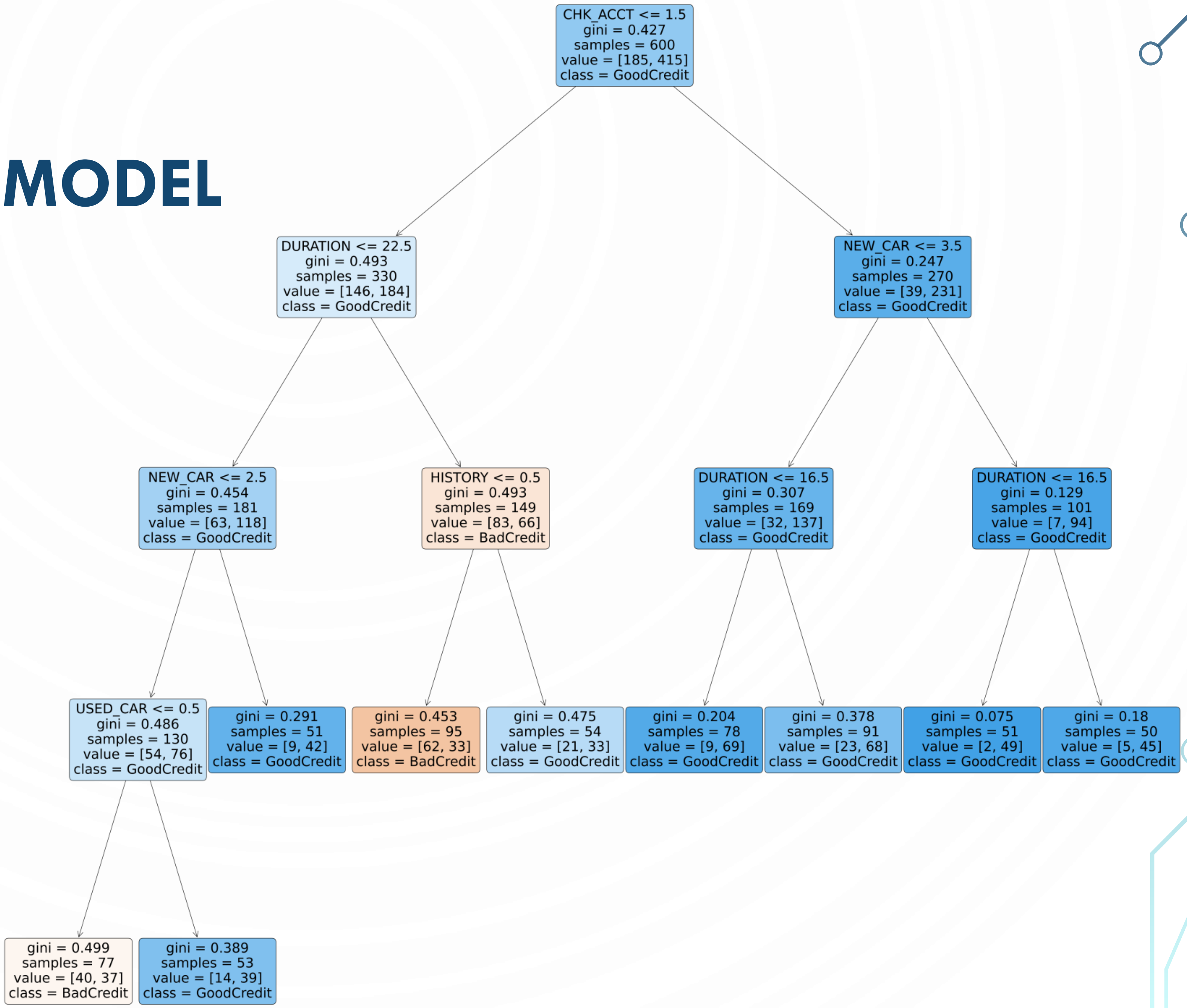
Amount in Checking Account

Duration of Credit in Months

Credit History

Purpose of Credit is New Car

Purpose of Credit is Used Car



CONCLUSIONS

The old mantra "No deal it's better than a bad deal" can prove to be wrong in our case.

Even if we cannot predict entirely if a new client will be a good or a bad client, we can implement a set of tools to reduce the incorporation of clients that will become delinquent on their loans or probably, default on.

Our decision tree model has a positive lift and might save the bank DM $71.800.

RECOMMENDATIONS

- Should be a joint effort from al the bank’s areas.
- Include more categories in the USE OF CREDIT. If this line of credit should be used in one of the present variables (NEW_CAR, USED_CAR, FURNITURE, RADIO/TV, EDUCATION, RETRAINING), the sum of those values should be 1. We observed 29 cases where no choice was made and 30% of those clients were part of the BAD_CREDIT statistic.
- We recommend identifying if other controls can be made with clients with INSTALL_RATE = 4. One in every three clients has been in default with the bank.
- There are three (3) variables related to gender and it’s not considering females. It’s important to include this in the model because certainly they will approach the bank as well and the model won’t fit properly if they are not contemplated in the algorithm.