Consumer Loans in India

Team 2 - America

Cooper Atkins, Nusrat Nawshin, Ricardo Diaz, Varun Shah

Our dataset originally contained 13 variables and 252K observations

The dataset contains customer data including incomes (in rupees), job experience, home ownership, and other characteristics along with a **Risk Flag** signaling **default vs. non-default** loan status.

```
252,000 observations with 13 variables
```

```
 [1] "Id"               "Income"           "Age"              "Experience"       "Married.Single"
 [6] "House_Ownership"  "Car_Ownership"    "Profession"       "CITY"             "STATE"
[11] "CURRENT_JOB_YRS"  "CURRENT_HOUSE_YRS" "Risk_Flag"
```

Through our EDA, we were able to answer the majority of our original questions

Questions

1. Do customers who default on loans have statistically lower **incomes** than those who don't default?
2. Does **homeownership** correlate with lower rates of default?
3. Does being **married** decrease the likelihood of default?
4. Does an additional **year of homeownership** reduce the likelihood of default?
5. Does **job experience** or **age** show a larger impact on someone defaulting on their loan?
6. Are customers who default on loans **younger** than those who do not?

Answers

1. Not statistically significant
2. Statistically significant difference
3. Statistically significant difference
4. Defaulted is statistically significantly lower
5. Statistically significant difference
6. Defaulted is statistically significant lower

We developed new questions to be answered by our model

Questions we posed :-

1. Despite not being useful on it's own, is income statistically significant in a model when other variables are included?
2. Does current job experience or overall job experience yield a better model?
3. For each additional year, how does age impact likelihood of defaulting on a loan?
4. For each additional year, how does job experience impact likelihood of defaulting on a loan?
5. Does manual, exhaustive stepwise, or other modelling techniques produce a better model
6. Does each method for model selection produce a significant model as determined by ROC-AUC >= 0.8, and if so, which produces the best?
7. What are the most significant predictors for default, and do they appear across all of the "best" models?
8. Using a confusion matrix, which of our "best" models appears to perform best across the different metrics we care about (precision, recall-rate, etc.)
9. How do the different models fare when used on the test dataset?

Model Preparation

Train-Test Split: 75% and 25%

Target Variable: *Risk_Flag*

Predictor Variables: *Income, Age, Job Experience, Marital Status, Home Ownership, Car Ownership, Years in Current Home, Years in Current Job*

Modeling Techniques Used:

Logistic Models - Manual Selection, Backward Elimination, Forward Selection and Exhaustive (with AIC and BIC criterion)

Decision Tree and Random Forest

2

Logistic Regression

We started with a manual logistic regression based on EDA results to build a baseline

Step 1	Step 2	Step 3	Step 4	Step 5
First, we determine which of the job experience variables is most significant	Add a second variable to the model (House Ownership) and test interaction if it's significant	Add another variable to the model (Years in current house) and test interaction if it's significant	Continue adding variables to the model (removing insignificant ones along the way) and test interaction as you go	When there are no more variables to try and all remaining terms are significant, determine this as the final model

Manual Regression: Which experience variable to choose

Current Job Experience

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.87209    0.01470 -127.31  < 2e-16 ***
CURRENT_JOB_YRS -0.01519    0.00205   -7.41  1.3e-13 ***
---
```

Overall Job Experience

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.78876    0.01406  -127.3   <2e-16 ***
Experience  -0.01814    0.00124   -14.6   <2e-16 ***
---
```

Overall job experience is slightly more significant with a standard error of 0.00124 vs 0.00205

Manual Regression: Home Ownership and Interaction

No Interaction

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.76592    0.01418 -124.55  < 2e-16 ***
Experience          -0.01796    0.00124  -14.44  < 2e-16 ***
House_OwnershipOwning  -0.39694    0.03880  -10.23  < 2e-16 ***
House_OwnershipNeither -0.25606    0.04826   -5.31  1.1e-07 ***
---
```

Yes Interaction

```
Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -1.76713    0.01451 -121.78  < 2e-16 ***
Experience                          -0.01783    0.00128  -13.89  < 2e-16 ***
House_OwnershipOwning               -0.34248    0.07542   -4.54  5.6e-06 ***
House_OwnershipNeither              -0.28513    0.08996   -3.17   0.0015 **
Experience:House_OwnershipOwning    -0.00551    0.00660   -0.84   0.4034
Experience:House_OwnershipNeither    0.00304    0.00789    0.38   0.7004
---
```

Home Ownership baseline is “Renting” and interaction is insignificant

Manual Regression: Add New variables and test interactions

No Interaction

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.76592    0.01418 -124.55  < 2e-16 ***
Experience          -0.01796    0.00124  -14.44  < 2e-16 ***
House_OwnershipOwning  -0.39694    0.03880  -10.23  < 2e-16 ***
House_OwnershipNeither -0.25606    0.04826   -5.31  1.1e-07 ***
---
```

Yes Interaction

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -1.76713    0.01451 -121.78  < 2e-16 ***
Experience                        -0.01783    0.00128  -13.89  < 2e-16 ***
House_OwnershipOwning             -0.34248    0.07542   -4.54  5.6e-06 ***
House_OwnershipNeither            -0.28513    0.08996   -3.17   0.0015 **
Experience:House_OwnershipOwning  -0.00551    0.00660   -0.84   0.4034
Experience:House_OwnershipNeither  0.00304    0.00789    0.38   0.7004
---
```

Home Ownership baseline is “Renting” and interaction is insignificant

Manual Regression: Iterate until no more variables available to add

Coefficients:

	Estimate
(Intercept)	-1.62e+00
Experience	-1.31e-02
House_OwnershipOwning	
House_OwnershipNeither	

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.579042	0.025700	-61.44	< 2e-16	***
Experience	-0.017973	0.001243	-14.46	< 2e-16	***

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.551084	0.027433	-56.54	< 2e-16	***
Experience	-0.018001	0.001302	-13.82	< 2e-16	***
House_OwnershipOwning	-0.341465	0.120654	-2.83	0.0047	**
House_OwnershipNeither	-0.538463	0.152235	-3.54	0.0004	***
Age			.8	2.7e-16	***
Marriec			.2	0.0018	**
House_C			91	0.3641	
House_C			.2	0.0342	*
Experie			36	0.7206	
House_C			53	5.9e-06	***
House_C			78	0.0756	.
Age:Marriea.Singlemarriea		0.000544	0.001532	0.35	0.7227

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.72207	0.06516	-26.43	< 2e-16	***
Experience	-0.01794	0.00124	-14.43	< 2e-16	***
House_OwnershipOwning	-0.39660	0.03880	-10.22	< 2e-16	***
House_OwnershipNeither	-0.25619	0.04826	-5.31	1.1e-07	***
CURRENT_HOUSE_YRS	-0.00367	0.00533	-0.69	0.49	

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.55e+00	2.93e-02	-52.74	< 2e-16	***
Experience	-1.79e-02	1.24e-03	-14.36	< 2e-16	***
House_OwnershipOwning	-3.41e-01	1.21e-01	-2.83	0.00470	**
House_OwnershipNeither	-5.38e-01	1.52e-01	-3.53	0.00041	***
Age	-3.81e-03	4.51e-04	-8.45	< 2e-16	***
Married.SingleMarried	-2.46e-01	2.72e-02	-9.03	< 2e-16	***
Income	-1.62e-09	2.59e-09	-0.63	0.53057	
:Age	-2.07e-03	2.27e-03	-0.91	0.36208	
':Age	6.01e-03	2.85e-03	2.11	0.03501	*
:Married.SingleMarried	5.99e-01	1.33e-01	4.52	6.2e-06	***
d	-4.07e-01	2.27e-01	-1.79	0.07339	.

	Estimate	Std. Error	z value	Pr(>|z|)	
	-1.508202	0.026745	-56.39	< 2e-16	***
	-0.017803	0.001244	-14.31	< 2e-16	***
	-0.362277	0.120926	-3.00	0.00274	**
	-0.540712	0.152505	-3.55	0.00039	***
	-0.003801	0.000451	-8.43	< 2e-16	***
	-0.246429	0.027219	-9.05	< 2e-16	***
	-0.165359	0.016679	-9.91	< 2e-16	***
	-0.001647	0.002277	-0.72	0.46930	
	0.006078	0.002856	2.13	0.03330	*
d	0.602831	0.132633	4.55	5.5e-06	***
ed	-0.405479	0.227085	-1.79	0.07417	.

Manual Regression: Final Model

Final Model

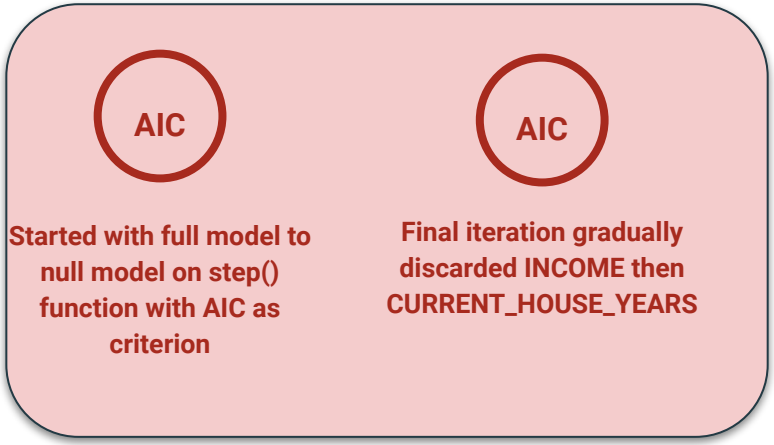
```
Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                 -1.508202   0.026745  -56.39  < 2e-16 ***
Experience                                  -0.017803   0.001244  -14.31  < 2e-16 ***
House_OwnershipOwning                       -0.362277   0.120926   -3.00  0.00274 **
House_OwnershipNeither                      -0.540712   0.152505   -3.55  0.00039 ***
Age                                         -0.003801   0.000451   -8.43  < 2e-16 ***
Married.SingleMarried                       -0.246429   0.027219   -9.05  < 2e-16 ***
Car_OwnershipYes                            -0.165359   0.016679   -9.91  < 2e-16 ***
House_OwnershipOwning:Age                   -0.001647   0.002277   -0.72  0.46930
House_OwnershipNeither:Age                   0.006078   0.002856    2.13  0.03330 *
House_OwnershipOwning:Married.SingleMarried  0.602831   0.132633    4.55  5.5e-06 ***
House_OwnershipNeither:Married.SingleMarried -0.405479  0.227085   -1.79  0.07417 .
---
```

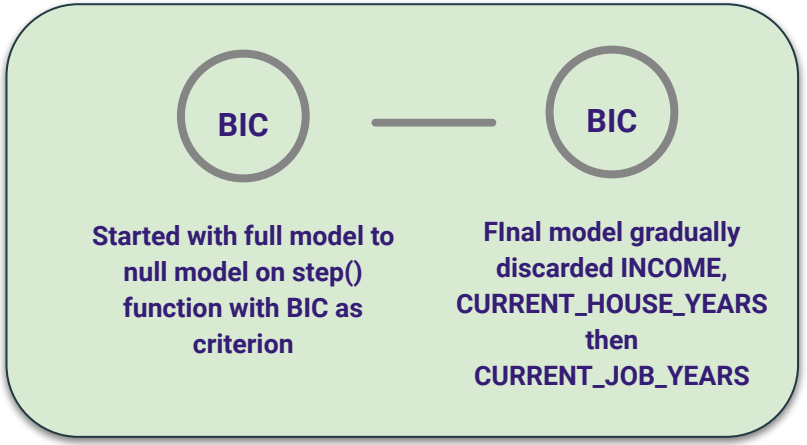
Accuracy	Precision	Recall Rate	Specificity	ROC-AUC
0.784	0.15	0.161	0.872	0.516

Interpretable Odds Ratio Coefficients

(Intercept)	0.221
Experience	0.982
House_OwnershipOwning	0.696
House_OwnershipNeither	0.582
Age	0.996
Married.SingleMarried	0.782
Car_OwnershipYes	0.848
House_OwnershipOwning:Age	0.998
House_OwnershipNeither:Age	1.006
House_OwnershipOwning:Married.SingleMarried	1.827
House_OwnershipNeither:Married.SingleMarried	0.667

Backward Elimination





As BIC could eliminated more, we are choosing this as our final model from **backward** elimination.

Final Model: "**Risk_Flag ~ Age + Married.Single + Car_Ownership + House_Ownership + Experience**"

Backward Elimination - Final Model

"Risk_Flag ~ Age + Married.Single + Car_Ownership + House_Ownership + Experience"

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.512612   0.026174  -57.79  < 2e-16 ***
Age                      -0.003720   0.000437   -8.52  < 2e-16 ***
Married.SingleMarried    -0.232842   0.026446   -8.80  < 2e-16 ***
Car_OwnershipYes         -0.165326   0.016674   -9.92  < 2e-16 ***
House_OwnershipOwning    -0.397992   0.038828  -10.25  < 2e-16 ***
House_OwnershipNeither   -0.262363   0.048282   -5.43  5.5e-08 ***
Experience               -0.017882   0.001244  -14.37  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Confusion Matrix

	0	1
0	64303	8729
1	9312	1656

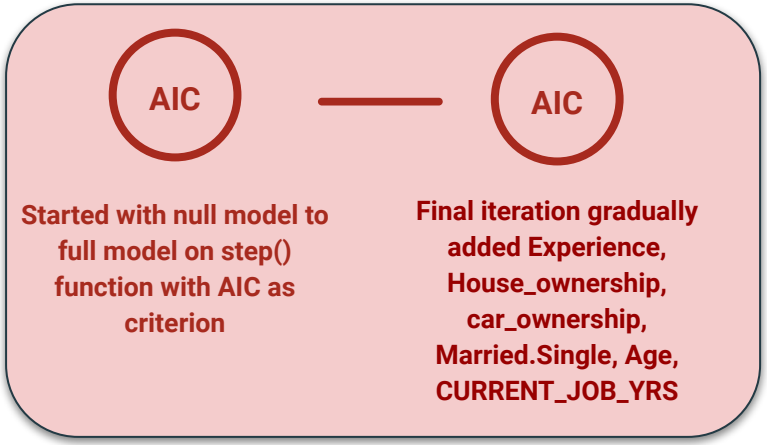
Accuracy	Precision	Recall Rate	Specificity	ROC-AUC
0.784	0.151	0.159	0.873	0.516

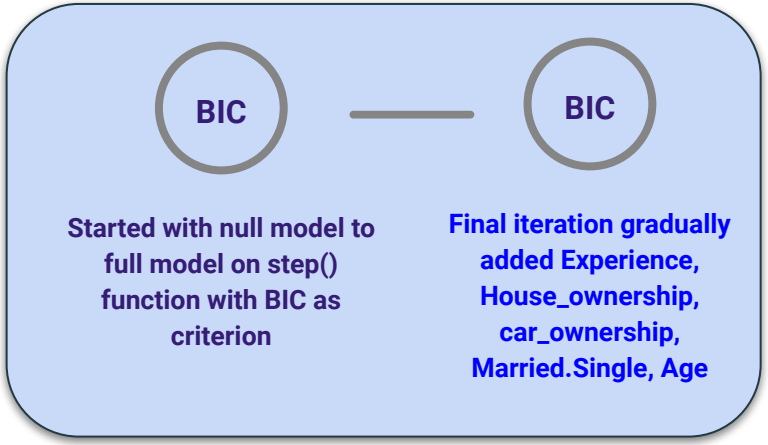
Odds Ratio Coefficients

Exponential of coefficients in Backward Logit Reg

	x
(Intercept)	0.220
Age	0.996
Married.SingleMarried	0.792
Car_OwnershipYes	0.848
House_OwnershipOwning	0.672
House_OwnershipNeither	0.769
Experience	0.982

Forward Selection





As BIC could eliminated more, we are choosing this as our final model from forward selection.

Final Model: "**Risk_Flag ~ Experience + House_Ownership + Car_Ownership + Married.Single + Age**"

Forward Selection - Final Model

"Risk_Flag ~ Experience + House_Ownership + Car_Ownership + Married.Single + Age"

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.512612   0.026174  -57.79  < 2e-16 ***
Experience               -0.017882   0.001244  -14.37  < 2e-16 ***
House_OwnershipOwning    -0.397992   0.038828  -10.25  < 2e-16 ***
House_OwnershipNeither   -0.262363   0.048282   -5.43  5.5e-08 ***
Car_OwnershipYes         -0.165326   0.016674   -9.92  < 2e-16 ***
Married.SingleMarried    -0.232842   0.026446   -8.80  < 2e-16 ***
Age                      -0.003720   0.000437   -8.52  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Confusion Matrix

	0	1
0	25536	2945
1	48079	7440

Accuracy	Precision	Recall Rate	Specificity	ROC-AUC
0.393	0.134	0.716	0.345	0.532

Odds Ratio Coefficients

Exponential of coefficients in Forward Logit Reg

	x
(Intercept)	0.220
Experience	0.982
House_OwnershipOwning	0.672
House_OwnershipNeither	0.769
Car_OwnershipYes	0.848
Married.SingleMarried	0.792
Age	0.996

Exhaustive - with AIC Criterion

The Top 5 Best Models

Income <lgl>	Age <lgl>	Experience <lgl>	Married.Single <lgl>	House_Ownership <lgl>	Car_Ownership <lgl>	CURRENT_JOB_YRS <lgl>	CURRENT_HOUSE_YRS <lgl>	Criterion <dbl>
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	124466
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	124468
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	124468
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	124469
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	124473

5 rows

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.531046   0.026883  -56.95  < 2e-16 ***
Age                     -0.003721   0.000437   -8.52  < 2e-16 ***
Experience              -0.021317   0.001690  -12.61  < 2e-16 ***
Married.SingleMarried   -0.232350   0.026448   -8.79  < 2e-16 ***
House_OwnershipOwning   -0.397442   0.038829  -10.24  < 2e-16 ***
House_OwnershipNeither  -0.263404   0.048283   -5.46  4.9e-08 ***
Car_OwnershipYes        -0.165946   0.016676   -9.95  < 2e-16 ***
CURRENT_JOB_YRS          0.008367   0.002758    3.03   0.0024 **
---
```

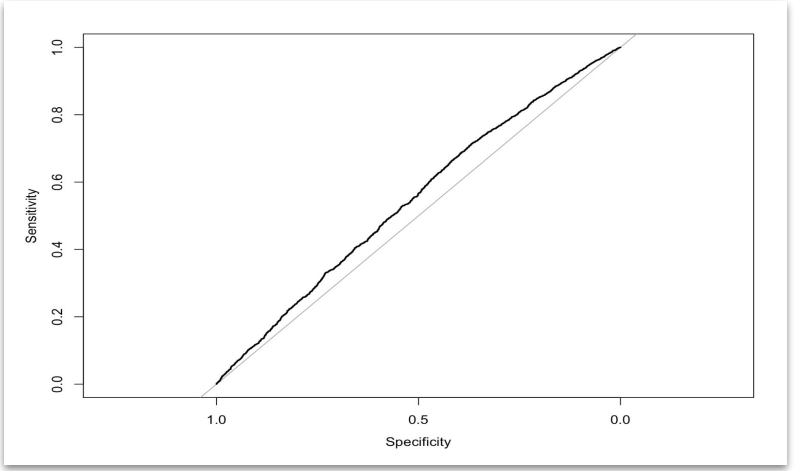
Baseline is ‘Single’, “Renting” and “No” Car ownership

	OR	2.5 %	97.5 %
(Intercept)	0.216	0.205	0.228
Age	0.996	0.995	0.997
Experience	0.979	0.976	0.982
Married.SingleMarried	0.793	0.752	0.835
House_OwnershipOwning	0.672	0.622	0.725
House_OwnershipNeither	0.768	0.698	0.844
Car_OwnershipYes	0.847	0.820	0.875
CURRENT_JOB_YRS	1.008	1.003	1.014

Exhaustive with AIC (Model Evaluation)

Confusion Matrix

Predicted \ Actual	0	1
0	67682	9317
1	5933	1068



Accuracy	Precision	Recall Rate	Specificity	ROC-AUC
81.8%	15.25%	10.28%	92%	0.548

Exhaustive - with BIC Criterion

The Top 5 Best Models

Income <lgl>	Age <lgl>	Experience <lgl>	Married.Single <lgl>	House_Ownership <lgl>	Car_Ownership <lgl>	CURRENT_JOB_YRS <lgl>	CURRENT_HOUSE_YRS <lgl>	Criterion <dbl>
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	124534
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	124536
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	124545
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	124545
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	124548

5 rows

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.512612   0.026174  -57.79  < 2e-16 ***
Age                      -0.003720   0.000437   -8.52  < 2e-16 ***
Experience               -0.017882   0.001244  -14.37  < 2e-16 ***
Married.SingleMarried    -0.232842   0.026446   -8.80  < 2e-16 ***
House_OwnershipOwning    -0.397992   0.038828  -10.25  < 2e-16 ***
House_OwnershipNeither   -0.262363   0.048282   -5.43  5.5e-08 ***
Car_OwnershipYes         -0.165326   0.016674   -9.92  < 2e-16 ***
---
```

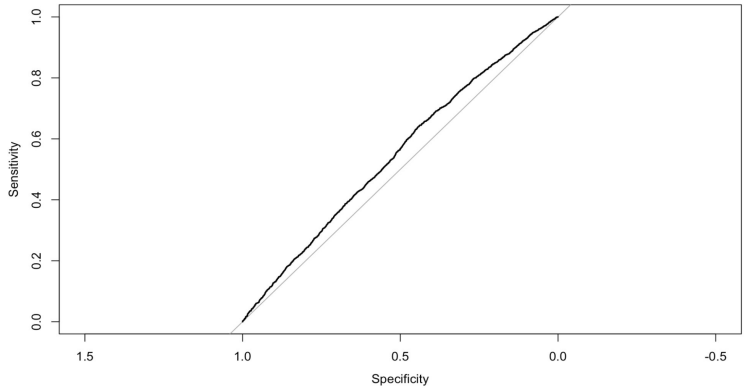
	OR	2.5 %	97.5 %
(Intercept)	0.220	0.209	0.232
Age	0.996	0.995	0.997
Experience	0.982	0.980	0.985
Married.SingleMarried	0.792	0.752	0.834
House_OwnershipOwning	0.672	0.622	0.724
House_OwnershipNeither	0.769	0.699	0.845
Car_OwnershipYes	0.848	0.820	0.876

Baseline is ‘Single’, “Renting” and “No” Car ownership

Exhaustive with BIC (Model Evaluation)

Confusion Matrix

Predicted	Actual 0	1
0	67655	9286
1	5960	1099



Accuracy	Precision	Recall Rate	Specificity	ROC-AUC
81.9%	15.6%	10.58%	92%	0.548

3

Decision Tree

Initial Results

Before Tuning

Accuracy	Sensitivity	Specificity
87%	100%	0%

Confusion matrix from Decision Tree

	0	1
0	73693	0
1	10307	0

Final Result

After Tuning

Accuracy	Sensitivity	Specificity	ROC-AUC
88.2%	56.9%	92%	53.2

Confusion matrix from Tuned Decision Tree

	0	1
0	68158	5457
1	4471	5914

Parameters used:

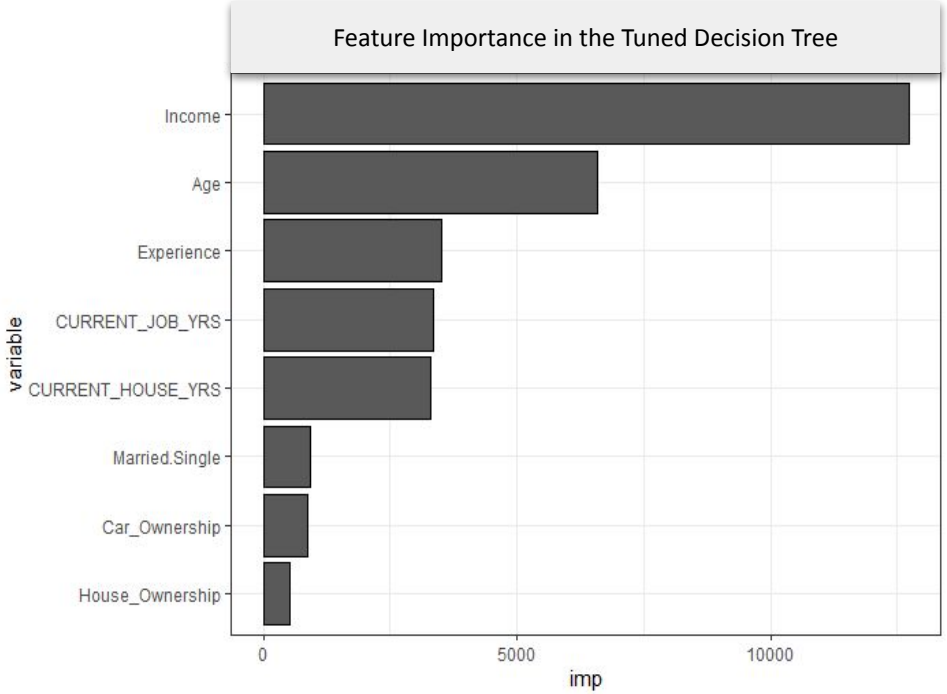
Min split : 4

Max depth: 30

MinBucket = 2

CP = 0

Feature Importance



4

Random Forest

Initial Results

Random Forest Accuracy:
0.88

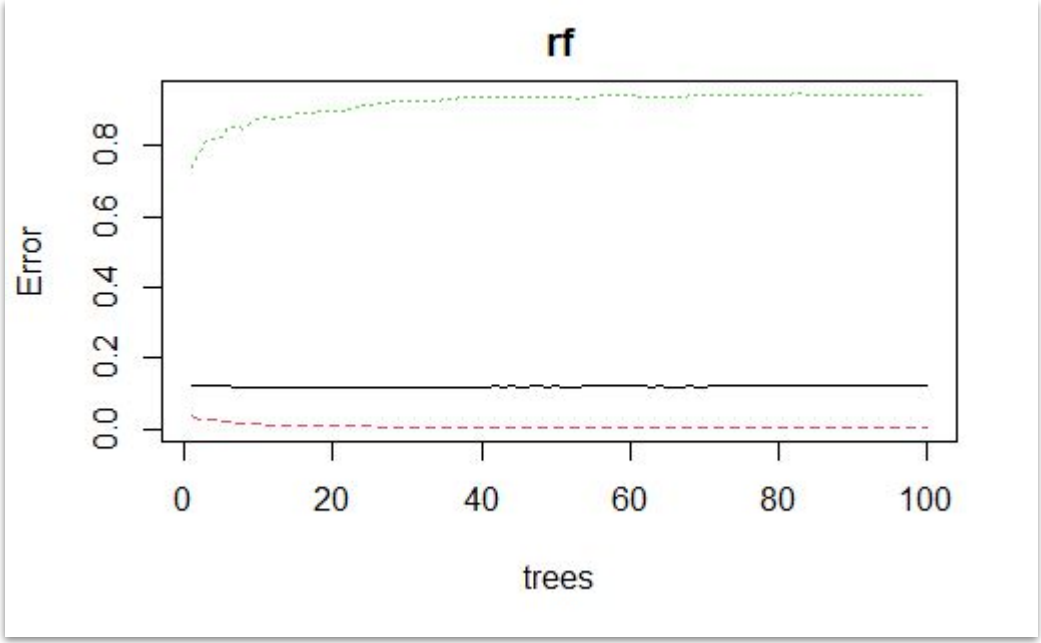
Area under the curve:
0.522

Sensitivity : **0.995**

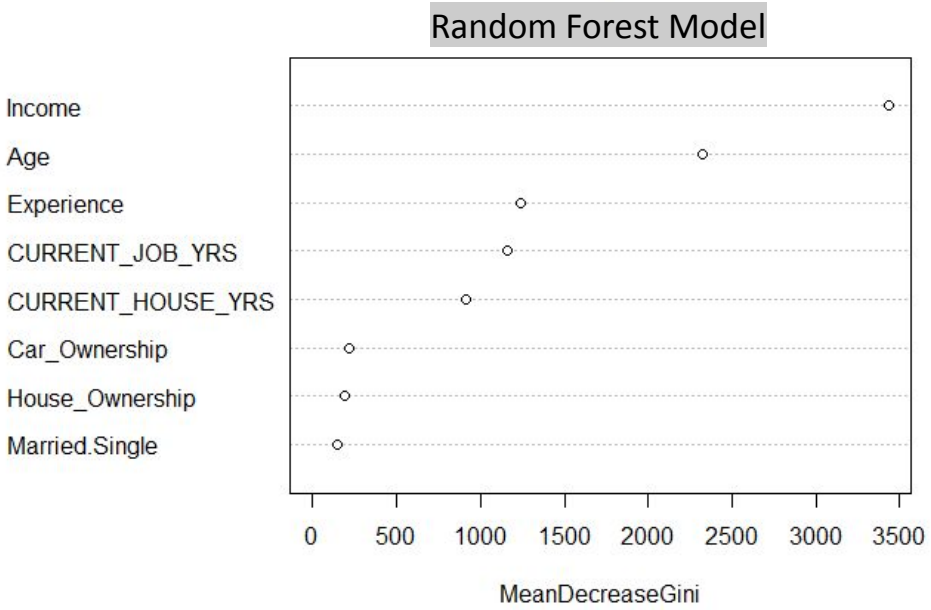
Specificity : **0.058**

Confusion matrix from Random Forest

	No Defaulted	Defaulted
No Defaulted	73319	374
Defaulted	9626	681



Feature Importance



5

Results & Answers

Comparing our models across a variety of metrics gives us an answer on which is best: **Decision Tree**

	Manual Regression	Backward Selection	Forward Selection	Exhaustive w/ BIC	Decision Tree	Random Forest
Accuracy	0.784	0.785	0.393	0.819	0.882	0.879
Precision	0.15	0.151	0.134	0.156	0.52	0.618
Recall Rate	0.161	0.159	0.716	0.106	0.569	0.058
Specificity	0.872	0.874	0.347	0.919	0.926	0.995
ROC-AUC	0.516	0.516	0.532	0.548	0.532	0.526

Note: Values calculated against test dataset

Answering SMART Questions About the Model Outputs

Questions	Answers
1. Despite not being useful on it's own, is **income** statistically significant in a model when other variables are included?	1. **None** of the logistics models could keep income 2. It was highly important in Decision Tree + Random Forest
2. Does current **job experience** or **overall job experience** yield a **better model**?	2. Overall job experience appears to be **better** in the models
3. For each additional year, how does **age** impact likelihood of defaulting on a loan?	3. Based on **exhaustive** selection w/ **AIC or BIC**, for each additional year, we expect the likelihood of default to **decrease** by **0.372%**
4. For each additional **year**, how does **job experience** impact likelihood of **defaulting** on a loan?	4. Based on **exhaustive** selection w/ **BIC**, for each additional year, we expect the likelihood of default to **decrease** by **1.78%**
5. What are the most **significant** predictors for default, and do they appear across all of the "**best**" models?	5. **Job experience**, **age**, **marital** status, **home-ownership** status, and **car ownership** status

Answering SMART Questions About the Model Process

Questions	Answers
1. Does **manual**, **exhaustive** stepwise, or **other** modeling techniques produce a better model ?	1. **Other** modeling techniques: **Decision tree** + **Random Forest**
2. Does each method for model selection produce a significant model as determined by **ROC-AUC >= 0.8**, and if so, which produces the best?	2. **None** of the models were significant via AUC-ROC
3. Using a **confusion matrix**, which of our "**best**" models appears to perform best across the different metrics we care about (**precision**, **recall**-**rate**, etc.)	3. **Forward** selection was the **worst** across these metrics, **random forest** was **ok**, but **decision tree** was the **least bad** across all confusion matrix based metrics
4. How do the different models **fare** when used on the test dataset?	4. They **fare** quite differently, but **overall** **not** **great**

References

Data sourced from Kaggle:
https://www.kaggle.com/subhamjain/loan-prediction-based-on-customer-behavior?select=Training+Data.csv