

```

1 options nodate nonumber;
2 ods graphics on / width = 2.5 in height = 2.125 in;
3 ods pdf file = "/home/u64313648/Personal Work Folder/Progress Report - Final Project/Project.pdf" pdfnoc = 3 startpage :
4
5 proc import datafile="/home/u64313648/Personal Work Folder/Progress Report - Final Project/home_loan.csv"
6   out=home_raw
7   dbms=csv
8   replace;
9   guessingrows=max;
10  getnames=yes;
11 run;
12
13 /* Data Preparation: */
14
15 data loan;
16   set home_raw;
17   drop Loan_ID;
18
19   length Loan_Status_YN $3;
20   if upcase(Loan_Status) in ("Y","YES") then Loan_Status_YN = "Yes";
21   else if upcase(Loan_Status) in ("N","NO") then Loan_Status_YN = "No";
22   else Loan_Status_YN = "NA";
23
24   length DepCat $4;
25   if Dependents in ("0","1","2","3+") then DepCat = Dependents;
26   else DepCat = "3+";
27
28   length Gender2 $6;
29   if upcase(Gender) in ("MALE","M") then Gender2 = "Male";
30   else if upcase(Gender) in ("FEMALE","F") then Gender2 = "Female";
31   else Gender2 = "Other";
32
33   length Married2 $3;
34   if upcase(Married) in ("YES","Y") then Married2 = "Yes";
35   else if upcase(Married) in ("NO","N") then Married2 = "No";
36   else Married2 = "NA";
37
38   length Educ2 $12;
39   if upcase(Education) = "GRADUATE" then Educ2 = "Graduate";
40   else if upcase(Education) = "NOT GRADUATE" then Educ2 = "Not Graduate";
41   else Educ2 = "Other";
42
43   length SelfEmp2 $3;
44   if upcase(Self_Employed) in ("YES","Y") then SelfEmp2 = "Yes";
45   else if upcase(Self_Employed) in ("NO","N") then SelfEmp2 = "No";
46   else SelfEmp2 = "NA";
47
48   TotalIncome = ApplicantIncome + CoapplicantIncome;
49 run;
50
51 proc format;
52   value $loanfmt
53     "Yes" = "Approved"
54     "No" = "Not Approved"
55     "NA" = "Missing/Other";
56
57   value creditfmt
58     0 = "No / Poor"
59     1 = "Good";
60
61   value $depFmt
62     "0" = "0"
63     "1" = "1"
64     "2" = "2"
65     "3+" = "3+";
66
67   value $genderFmt
68     "Male" = "Male"
69     "Female" = "Female"
70     "Other" = "Other";
71
72   value $marFmt
73     "Yes" = "Married"
74     "No" = "Not Married"
75     "NA" = "Unknown";
76

```

```

77     value $eduFmt
78         "Graduate"      = "Graduate"
79         "Not Graduate"  = "Not Graduate"
80         "Other"         = "Other";
81
82     value $selfFmt
83         "Yes" = "Self-Employed"
84         "No"  = "Not Self-Employed"
85         "NA"  = "Unknown";
86 run;
87
88
89
90 /* Table 1: Loan_Status distribution */
91
92 title2 "Loan Status Prediction Project in SAS" justify = C;
93
94 proc freq data=loan;
95     tables Loan_Status_YN / nocum norow nocol;
96     format Loan_Status_YN $loanfmt.;
97     title2 "Distribution of Loan Approval Status";
98 run;
99
100 /* Table 2: Overall categorical distributions */
101
102 proc freq data=loan;
103     tables Married2 Credit_History Property_Area;
104     format Gender2      $genderFmt.
105            Married2     $marFmt.
106            DepCat       $depFmt.
107            Educ2        $eduFmt.
108            SelfEmp2     $selfFmt.
109            Credit_History creditfmt.;
110     title2 "Overall Distribution of Socio-Economic Categorical Predictors";
111 run;
112
113 data loan_adjusted;
114     set loan;
115     if (Married2= "NA" OR Gender2 = "Other") then delete;
116 run;
117
118 proc odstext;
119     p 'We removed values from the variables Married2 and Gender2 where we found that there were lack of observations for
120 run;
121
122 /* Table 3: Summary statistics overall (continuous predictors) */
123
124 proc means data=loan_adjusted n mean std median min max maxdec=1;
125     var ApplicantIncome CoapplicantIncome TotalIncome
126         LoanAmount Loan_Amount_Term;
127     title2 "Summary Statistics for Continuous Predictors (Overall Sample)";
128 run;
129
130 /* Histogram: ApplicantIncome */
131 proc sgplot data=loan_adjusted;
132     histogram ApplicantIncome;
133     density ApplicantIncome;
134     title2 "Distribution of Applicant Income";
135 run;
136
137 /* Histogram: LoanAmount */
138 proc sgplot data=loan_adjusted;
139     histogram LoanAmount;
140     density LoanAmount;
141     title2 "Distribution of Loan Amount";
142 run;
143
144 proc odstext;
145     p 'We can observe from the Distribution of Applicant Income vs Distribution of Loan Amount, that the first graph ter
146 run;
147
148 /* Table 4: Cross-tabs for key categorical predictors */
149
150 proc freq data=loan_adjusted;
151     tables Loan_Status_YN * Credit_History
152            Loan_Status_YN * Property_Area
153            Loan_Status_YN * Married2

```

```

154 / nocol norow;
155 format Loan_Status_YN $loanfmt.
156 Credit_History creditfmt.
157 Married2 $marFmt.;
158 title2 "Loan Approval Status by Selected Categorical Predictors";
159 run;
160
161 /* Fig. 2: Approval % by Credit History */
162
163 proc sgplot data=loan_adjusted;
164 vbar Credit_History / group=Loan_Status_YN stat=percent
165 groupdisplay=cluster datalabel;
166 format Credit_History creditfmt.
167 Loan_Status_YN $loanfmt.;
168 yaxis label="Percentage within Credit History Group";
169 xaxis label="Credit History";
170 title2 "Loan Approval by Credit History";
171 run;
172
173 proc odstext;
174 p 'For those who were not approved, No/Poor credit history vs Good seemed to not have a major impact on the proporti
175 run;
176
177 /* Fig. 3: Approval % by Property Area */
178
179 proc sgplot data=loan_adjusted;
180 vbar Property_Area / group=Loan_Status_YN stat=percent
181 groupdisplay=cluster datalabel;
182 format Loan_Status_YN $loanfmt.;
183 yaxis label="Percentage within Property Area";
184 xaxis label="Property Area";
185 title2 "Loan Approval by Property Area";
186 run;
187
188 /* Table 5: Continuous predictors by Loan_Status */
189
190 proc means data=loan_adjusted n mean std median min max maxdec=1;
191 class Loan_Status_YN;
192 var ApplicantIncome CoapplicantIncome TotalIncome
193 LoanAmount Loan_Amount_Term;
194 format Loan_Status_YN $loanfmt.;
195 title2 "Continuous Predictors by Loan Approval Status";
196 run;
197
198 /* Fig. 4: Boxplot of ApplicantIncome by Loan_Status */
199
200 proc sgplot data=loan_adjusted;
201 vbox ApplicantIncome / category=Loan_Status_YN;
202 format Loan_Status_YN $loanfmt.;
203 yaxis label="Applicant Monthly Income";
204 xaxis label="Loan Approval Status";
205 title2 "Applicant Income by Loan Approval Status";
206 run;
207
208
209 proc sgplot data=loan_adjusted;
210 vbox LoanAmount / category=Loan_Status_YN;
211 format Loan_Status_YN $loanfmt.;
212 yaxis label="Requested Loan Amount";
213 xaxis label="Loan Approval Status";
214 title2 "Loan Amount by Loan Approval Status";
215 run;
216
217 proc odstext;
218 p "In the Loan Amount by Loan Approval Status we see that the two boxplots look similar but that those who have the
219 run;
220
221
222 ods select none;
223 ods output ChiSq = chi_all;
224
225 proc freq data=loan;
226 tables Loan_Status_YN *
227 (Gender2 Married2 DepCat Educ2 SelfEmp2
228 Credit_History Property_Area)
229 / chisq;
230 run;
231

```

```

231
232 ods output close;
233 ods select all;
234
235
236 /* Keep only Pearson Chi-Square for each predictor */
237 data chi_summary;
238     set chi_all;
239     where Statistic = "Chi-Square";
240     length Predictor $20;
241     Predictor = scan(Table, 2, '');
242 run;
243
244 proc print data=chi_summary noobs;
245     var Predictor DF Value Prob;
246     format Prob pvalue8.4;
247     title2 "Chi-Square Tests: Association with Loan Approval (Categorical Predictors)";
248 run;
249
250
251 /* T-tests */
252
253 ods select none;
254 ods output TTests = t_all;
255
256 proc ttest data=loan plots=none;
257     class Loan_Status_YN;
258     var ApplicantIncome CoapplicantIncome TotalIncome
259         LoanAmount Loan_Amount_Term;
260 run;
261
262 ods output close;
263 ods select all;
264
265 data t_summary;
266     set t_all;
267     where Method = "Pooled";
268 run;
269
270 proc print data=t_summary noobs;
271     var Variable DF tValue Prob;
272     format Prob pvalue8.4;
273     title2 "T-Test p-values for Continuous Predictors by Loan Status";
274 run;
275
276
277 /* Wilcoxon rank-sum nonparametric */
278 ods select none;
279 ods output WilcoxonTest = w_all;
280
281 proc npar1way data=loan wilcoxon plots=none;
282     class Loan_Status_YN;
283     var ApplicantIncome CoapplicantIncome TotalIncome
284         LoanAmount Loan_Amount_Term;
285 run;
286
287 ods output close;
288 ods select all;
289
290 /* Keep only the Pr > |Z| line for each variable */
291 data w_summary;
292     set w_all;
293     *where Label1 = "Pr > |Z|";
294 run;
295
296 proc print data=w_summary noobs;
297     *var Variable nValue1;
298     *format nValue1 pvalue8.4;
299     title2 "Wilcoxon Rank-Sum p-values for Continuous Predictors by Loan Status";
300 run;
301
302
303
304 proc odstext;
305     p 'This dataset contains 480 loan applicants and includes socio-economic, demographic, and financial predictors rel
306
307 The continuous variables show the usual right-skewness seen in financial data. Applicant and co-applicant incomes vary v

```

```

308
309 To formally assess these patterns, we applied chi-square tests to each categorical predictor. Credit history, property :
310 run;
311
312
313
314 /* PROGRESS REPORT END */
315
316
317
318 title2 'Correlation Table Examining Collinearity Between Numerical Variables';
319
320 /* Reduced output from ods select */
321 proc corr data = loan_adjusted cov spearman plots =matrix(histogram);
322     var ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term TotalIncome;
323     ods trace on;
324     ods select Corr.VarInformation Corr.PearsonCorr
325 run;
326
327
328
329 proc odstext;
330     p 'We can see from this correlation matrix that TotalIncome and ApplicantIncome are both highly correlated with a P
331 run;
332
333
334 /* This will create a Loan_Status variable which is binary that can be used instead of the character version, but same i
335 Most up to date data that should be used is logistic_loan. No output created here*/
336 data logistic_loan;
337     set loan_adjusted;
338     if Loan_Status_YN = "Yes" then Loan_Status_Num = 1;
339     else Loan_Status_Num = 0;
340 run;
341
342 /* Uses Reg procedure to further show collinearity diagnostics from proc reg reduced output from ods select*/
343 title2 'VIF and Collinearity Diagnostics Displayed From the REG Procedure';
344
345 proc reg data = logistic_loan;
346     model Loan_Status_Num = ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term TotalIncome / VIF collin;
347     ods trace on;
348     ods select Reg.MODEL1.Fit.Loan_Status_Num.DependenceEquations Reg.MODEL1.Fit.Loan_Status_Num.ParameterEstimates Reg
349 run;
350
351 proc odstext;
352     p 'Using proc reg to see the VIF between our numerical continuous variables we see that TotalIncome is a linear comb
353     When looking at the collinearity diagnostics we see that TotalIncome has a condition index of 2125178 which is c
354 run;
355
356 /* Remember when fitting the models to not include TotalIncome in our model because of collinearity concerns */
357 Title2 'Stepwise Regression From LOGISTIC Procedure' justify=center;
358
359 /* Output Reduced*/
360 ods graphics on;
361
362 proc logistic data = logistic_loan;
363     class Credit_History (ref='0')
364         Property_Area (ref='Rural')
365         Married2 (ref='No')
366         SelfEmp2 Educ2 Gender2 DepCat
367     / param = ref;
368
369     model Loan_Status_Num(event='1') =
370         ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term
371         SelfEmp2 Educ2 Gender2 DepCat
372         Property_Area Credit_History Married2
373     / selection = stepwise details lackfit influence;
374
375     output out = logistic_loan2
376         cbar = cbar
377         DFBetas = DfBetas;
378
379     ods exclude
380         Influence
381         InfluencePlots
382         ROC
383         ROCCurve
384         AssociationPlot

```

```

385     EffectPlot
386     ParameterEstimates;
387
388     ods select
389         Logistic.ModelFit
390         Logistic.StepwiseSummary
391         Logistic.Step3.OddsRatios
392         Logistic.Step3.Association
393         Logistic.LackFit.LackFitChiSq;
394 run;
395
396 ods graphics off;
397
398
399
400 proc odstext;
401     p "Output reduced to step 3 fit statistics. We have an AIC value of 479.972, a c statistic or AUC value of .789 which
402         We see that our Pearson residuals have positive values as high as around 4, and an evident trend of cases above
403         When looking at our leverage we see that most of our higher leverage points tend to be cases where loan status is 1";
404 run;
405
406 title2 'Meaningful Observations';
407
408 /* Will only print a few rows that are influential*/
409 proc print data = logistic_loan2;
410     where cbar >= 1;
411 run;
412
413 proc print data = logistic_loan2;
414     where DfBetas >= (2 / sqrt(515));
415 run;
416
417 proc odstext;
418     p "These were influential observations for our model for one reason or another, observation 153 has a high applicant
419 run;
420
421 /* Genmod reduced output*/
422 title2 'Genmod Procedure Fit and Estimates';
423
424 proc genmod data = logistic_loan;
425     class SelfEmp2 Educ2 Married2 Gender2 DepCat Property_Area Credit_History;
426     model Loan_Status_Num(event = '1') = ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term SelfEmp2 Educ2 Married2 Gender2 DepCat Property_Area Credit_History;
427     ods trace on;
428     ods select Genmod.ModelFit Genmod.ParameterEstimates;
429 run;
430
431 proc odstext;
432     p "We're only displaying model fit statistics and parameter estimates, to show the AIC value (~490) from the genmod
433 run;
434
435 title2 'Final Model Selection and Interpretation';
436 proc odstext;
437     p "This shows that our stepwise logistic regression model is the best model for predicting loan status. It doesn't show
438         What this could mean overall is that the information in this training dataset is heavily reliant on based around
439         factor for determining loan status, and this dataset is reflective of that. Our second most influential variable is
440         If someone wanted to maximize their likelihood of getting a loan they would want to have a credit history, property area
441         As mentioned earlier no difference between urban and rural odds, however between semiurban and rural there is a difference
442         Finally for credit history we have a statistically significant result since 1 isn't within the interval of (19.6, 1.0)
443 run;
444
445
446
447
448 /* Additional analysis */
449
450 title2 "Classification Performance of Final Model";
451
452 ods select none;
453 ods output Classification = CT_05;
454
455 proc logistic data=logistic_loan;
456     class Credit_History (ref='0')
457         Property_Area (ref='Rural')
458         Married2 (ref='No')
459         / param=ref;
460
461     model Loan_Status_Num(event='1') =

```

```

462         Credit_History Property_Area Married2
463         / ctable pprob=0.5;
464 run;
465
466 ods output close;
467 ods select all;
468
469 proc print data=CT_05 noobs label;
470     title3 "Classification Table (Cutoff = 0.5)";
471 run;
472
473
474 ods select none;
475 ods output Classification = CT_CUT;
476
477 proc logistic data=logistic_loan;
478     class Credit_History (ref='0')
479           Property_Area   (ref='Rural')
480           Married2        (ref='No')
481           / param=ref;
482
483     model Loan_Status_Num(event='1') =
484           Credit_History Property_Area Married2
485           / ctable pprob=(0.3 0.5 0.7);
486 run;
487
488 ods output close;
489 ods select all;
490
491 proc print data=CT_CUT noobs label;
492     title3 "Classification Tables at Different Cutoffs (0.3, 0.5, 0.7)";
493 run;
494
495 proc odstext;
496     p "To further evaluate the performance of our final logistic regression model, we examined its classification accuracy. To better understand this tradeoff, we also examined alternative probability cutoffs of 0.3 and 0.7. When using a likelihood ratio test, the results show that the model with a cutoff of 0.5 performed best."
497 run;
498
499 proc odstext;
500     p "The code below has output that was created separate from the rest of the analysis above and serves to provide visualizations of the model's performance."
501 run;
502
503
504
505 /*
506 ods output FitStatistics = fitstats;
507
508 proc sgplot data=fitstats;
509     where Criterion = 'AIC';
510     series x=Step y=InterceptAndCovariates / markers;
511     xaxis label="Step in Selection Procedure";
512     yaxis label="AIC";
513     title "AIC Across Stepwise Logistic Regression";
514 run;
515
516 proc logistic data=loan;
517     class Credit_History (ref='0')
518           Property_Area   (ref='Rural')
519           Married2        (ref='No')
520           / param=ref;
521
522     model Loan_Status_YN(event='Yes') =
523           Credit_History Property_Area Married2;
524
525     output out=diag_out
526           pred=phat
527           reschi=pearson_resid
528           resdev=deviance_resid
529           hat=leverage
530           c=cooksd;
531 run;
532
533 proc sort data=diag_out;
534     by descending cooksd;
535 run;
536
537 title "Top 10 Observations by Cook's Distance (Final Logistic Model)";
538

```

```
539 proc print data=diag_out(obs=10);  
540     var Loan_Status_YN phat pearson_resid deviance_resid leverage cooksd;  
541 run;  
542 title;  
543  
544 proc logistic data=logistic_loan plots(only)=roc;  
545     class Credit_History (ref='0')  
546         Property_Area   (ref='Rural')  
547         Married2        (ref='No')  
548         / param=ref;  
549  
550     model Loan_Status_Num(event='1') =  
551         Credit_History Property_Area Married2;  
552 run;  
553  
554 ods graphics off; */  
555  
556  
557  
558
```