Harshaanth Thiyagaraja Kumar

Loan Eligibility Predictions:

Problem Statement:
Dream Housing Finance company deals in all home loans. They have a presence across all urban, semi-urban, and rural areas. Customer first applies for a home loan after that company validates the customer eligibility for a loan.

The company wants to automate the loan eligibility process (real-time) based on customer details provided while filling out the online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and others. To automate this process, they have given a problem to identify the customer's segments, those are eligible for loan amounts so that they can specifically target these customers. Here they have provided a partial data set.

Dataset Key Information.

Key Name	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)

Key Name	Description
Self_Employed	Self-employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of a loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi-Urban/ Rural
Loan_Status	Loan approved (Y/N)

Harshaanth Thiyagaraja Kumar

Missing Data Handling:

Imputer Parameters													
Variable	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncom	CoapplicantIncome	LoanAmount	Loan_Amount_Te	Credit_History	Property_	Loan_Statu
Reduction Type	DELETE RE	DELETE RECOR	DELETE REC	DELETE RECORD	DELETE RECORD	DELETE RECORD	DELETE RECORD	DELETE RECORD	DELETE RECORD	DELETE RECORD	DELETE RECOR	DELETE RE	DELETE RE
# Records Treated	0	13	3	15	0	32	0	0	22	14	50	0	0
Missing Value Code													
# Output Records	480												
#Records Deleted	134												

Logistic Regression:

Numerical Variables: Loan_Amount_Term, LoanAmount, CoapplicantIncome, ApplicantIncome

Categorical Variables: Property_Area, Credit_History, Self_Employed, Education, Dependents, Married, Gender

Target Variables: Loan_Status

Created Dummies for the Categorical Variables, And Set Baseline as Gender_Female, Married_No, Dependents_0, Education_Not_Graduate, Self_Employed_Yes, Property_Area_Rural

Best Subsets Details				
Subset ID	#Coefficients	RSS	Mallows's Cp	Probability
Subset 1	1	331.9927383	43.48587516	2.36272E-06
Subset 2	2	303.9297766	17.63481572	0.00973577
Subset 3	3	294.3779927	10.15515695	0.092244823

For variable selection, I Chose Subset 3 as it has the highest probability and mallows cp closest to the number of variables. And chose the significant variables based on P-Value < 0.05.

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-2.549495532	-3.662752746	-1.436238318	0.078121	0.567998811	20.1471517	7.17E-06
Credit_History	3.673082177	2.548549027	4.797615327	39.37307	0.573751946	40.98386825	1.53E-10
Property_Area_Semiurban	1.278173389	0.519422084	2.036924693	3.590076	0.387125126	10.90126315	0.000961

Logit (Loan Status=1)= b0+b1x1+b2x2+b3x3…bqxq

Logit (Loan Status=1)= -2.549+3.673Credit_History+1.278Property_Area_Semiurban

Odds(Loan Status=1)=e^(b0+b1x1+b2x2+b3x3…bqxq)

Odds(Loan Status=1)=e^(-2.549+3.673Credit_History+1.278Property_Area_Semiurban)

Odds(Loan Status=1)=Odds0*(Odds1)^x1*(Odds2)^x2*(Odds3)^x3*……*(Oddsq)^xq

Odds(Loan Status=1)=0.078*(39.373)^Credit_History *(3.59)^Property_Area_Semiurban

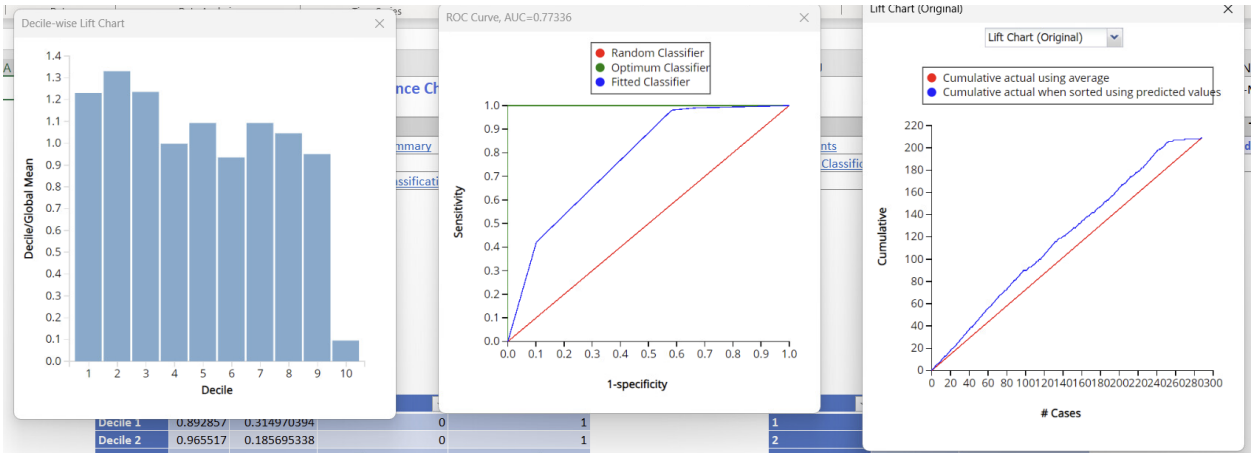
P(Loan Status=1)=1/(1+ e^-(b0+b1x1+b2x2+b3x3…bqxq)

P(Loan Status=1)=1/(1+ e^-(-2.549+3.673Credit_History+1.278Property_Area_Semiurban))

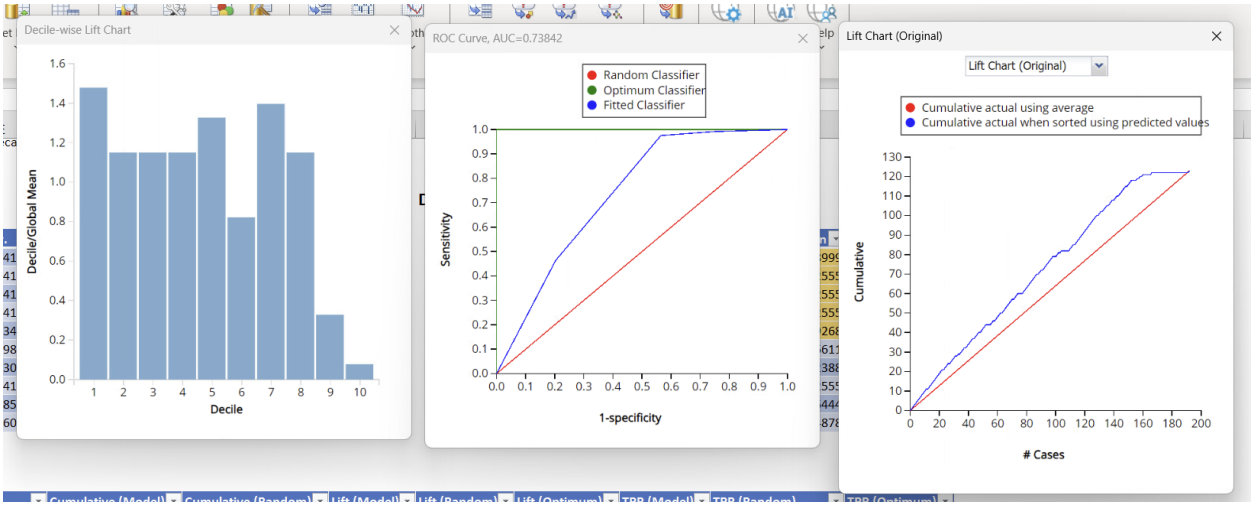
Interpretation of coefficients:

- For Credit_History: Holding all other variables constant, if an applicant has a credit history (1), the log odds of the loan status being 1 (approved) are expected to increase by a factor of 39.373 compared to an applicant without a credit history.
- For Property_Area_Semiurban: Holding all other variables constant, being in a semiurban property area increases the log odds of the loan status being 1 by a factor of 3.59.

Training Lift charts:



Validation Lift Charts:



This model demonstrates robust performance, with an AUC larger than 0.5, indicating superiority over the benchmark. Additionally, the model's performance remains consistent across both the training and validation datasets, as evidenced by the training AUC value of 0.773 and the validation AUC of 0.738, which are quite close without any significant difference. This consistency suggests the absence of overfitting issues and underscores the model's strong predictive capability and generalizability to new data.

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	30	39
1	3	120

Error Report			
Class	# Cases	# Errors	% Error
0	69	39	56.52173913
1	123	3	2.43902439
Overall	192	42	21.875

Metrics	
Metric	Value
Accuracy (#correct)	150
Accuracy (%correct)	78.125
Specificity	0.434782609
Sensitivity (Recall)	0.975609756
Precision	0.754716981
F1 score	0.85106383
Success Class	1
Success Probability	0.5

The model summary shows that the model has a high recall value of 0.97, but a low specificity of 0.43 and a high error rate for class 0 of 57%. So, we would like to explore and choose a new cutoff probability based on the Decile chart. Deciles 1-5 score over 1. Thus, choosing the probability of records in the 50th percentile as the new cutoff probability, 0.75. It had an improved specificity of 0.80, but the recall is reduced to 0.45.

However, since our business question is to identify class 1 loan approvals, we will go with a cutoff probability of 0.5, which has the best recall of 0.97, meaning that the model correctly identifies 97% of applicants whose actual loan status is 'Y'.

- **How do the results answer your questions?**

The results from the logistic regression model provide insights into the likelihood of an individual purchasing the product based on various factors, such as credit history and property area. Based on the model we have identified that customers with a credit history and who live in Property_Area_Semiurban are more likely to get their loan status approved.

- **Make an example of one new record and make prediction/classification**

If an applicant is a married male who is self-employed, with an applicant income of $6000, co-applicant income of $5000, has 2 dependents, and has a credit history(1) from a semi-urban area, with a graduate-level education, the loan amount would be $500 with a tenure of 120 months.

Odds(Loan Status=1)=Odds0*(Odds1)^x1*(Odds2)^x2*(Odds3)^x3*……*(Oddsq)^xq

Odds(Loan Status=1)=0.078*(39.373)^Credit_History *(3.59)^Property_Area_Semiurban

Odds(Loan Status=1)=0.078*(39.373)^1 *(3.59)^1

Odds(Loan Status=1)= 11.34

P = Odds/1+Odds = 11.34 /1+11.34 = 0.9192

Since predicted Probability (0.9192) > Success class cutoff probability (0.5), Classification 1

So, the probability of the applicant's loan status being approved as 'Y' is 0.9192. Since this probability is greater than the cutoff probability of 0.5, the applicant's loan status is predicted to be 'Y'.