Machine Learning Techniques for Predicting Credit Card Defaults

Laura Malovich, Marshal Will

Problem

Is it Possible to predict if a customer will default the next month ?			
Statistical Techniques	Generate a usable model from a M.L. algorithm	Make the model easy to use and intuitive	Useful at preventing unnecessary loss

Review of other Attempts

Analyzed which Machine Learning Method was the most accurate

Looked at what factors played the most in defaulting

Used as a practice set for commparing accuacy of other M.L. algorithms

As a basis to devlop other credit detection applications

Used Statistical Analsys to find better ways to predict when a client will default

Paper Analyzed

Yeh, I.-C., & Lien, C.-h. (2009). The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *ScienceDirect*, 2473-2480.

Classification accuracy

Method	Error rate		Area ratio	
	Training	Validation	Training	Validation
K-nearest neighbor	0.18	0.16	0.68	0.45
Logistic regression	0.20	0.18	0.41	0.44
Discriminant analysis	0.29	0.26	0.40	0.43
Naïve Bayesian	0.21	0.21	0.47	0.53
Neural networks	0.19	0.17	0.55	0.54
Classification trees	0.18	0.17	0.48	0.536

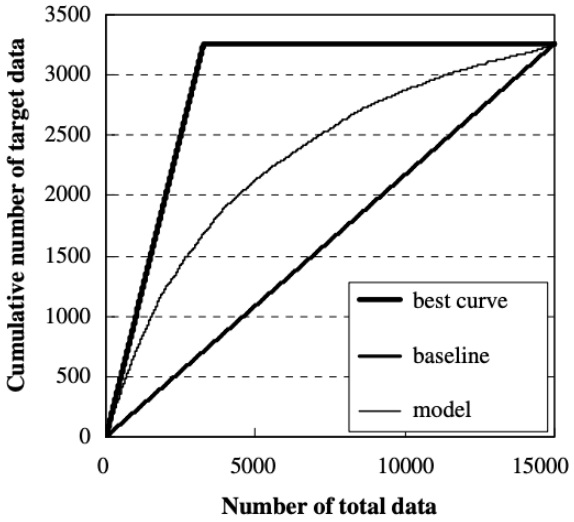


Fig. 6. Lift chart of artificial neural networks.

Initial Data Values

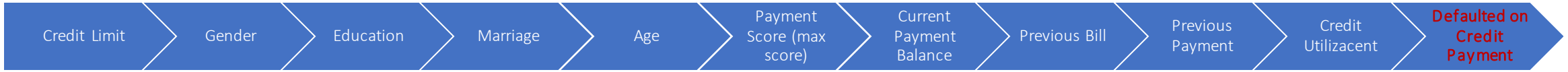
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	Y
ID	LIMIT_BA	SEX	EDUCATI	MARRIAC	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AM	BILL_AM	BILL_AM	BILL_AM	BILL_AM	BILL_AM	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	default paym
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0
7	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0
8	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0
9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0	0
11	200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	0
12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640	0
13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0	0
14	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0	1
15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	0
16	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100	0
17	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	0	1500	0	1650	0	1

Dataset Preparation

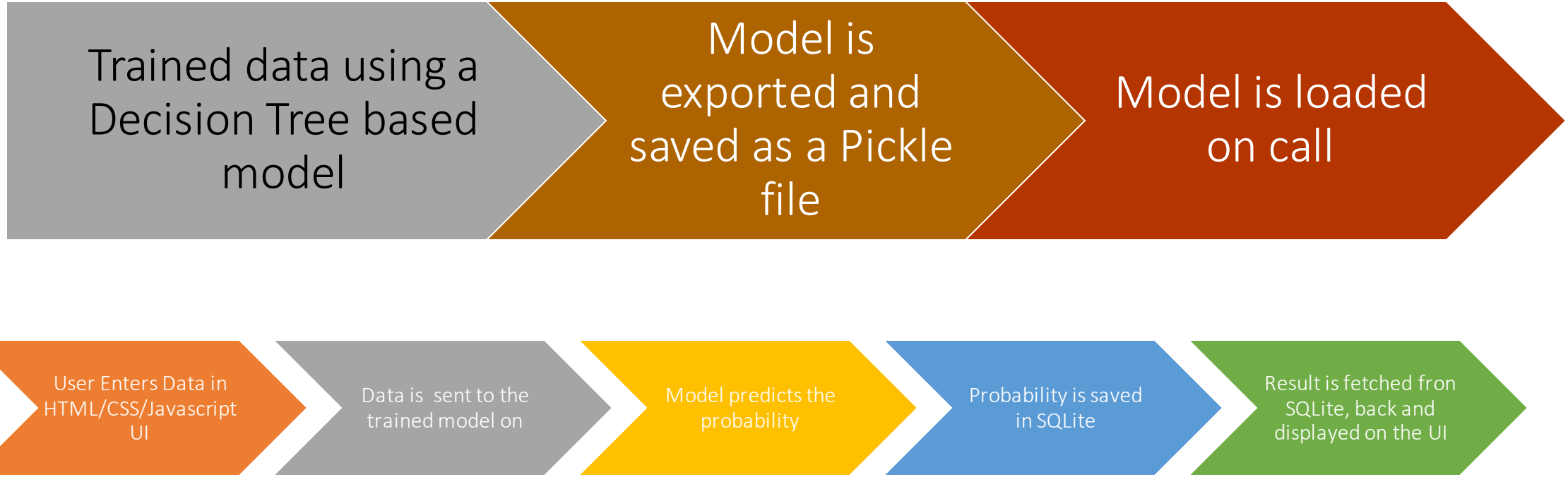
- Continuous variable types normalized and scaled
- Null values set to zero
- Calculate max payment score and credit utilization
- Drop unused columns
- Split Training 50%, Validation 25%, Testing 25%
- 30 000 rows and 10 columns

	B	C	D	E	F	G	H	I	J	K	L
1	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_MAX_SCORE	CURRENT_BAL	PREVIOUS_BAL	LAST_PAYMENT	CRED_UTILIZATION_PERCENT	default_payment_next_month
2	20000	2	2	1	24	2	3913	3102	0	20	1
3	120000	2	2	2	26	2	2682	1725	0	2	1
4	90000	2	2	2	34	0	29239	14027	1518	32	0
5	50000	2	2	1	37	0	46990	48233	2000	94	0

Dataset Features and Target



Design

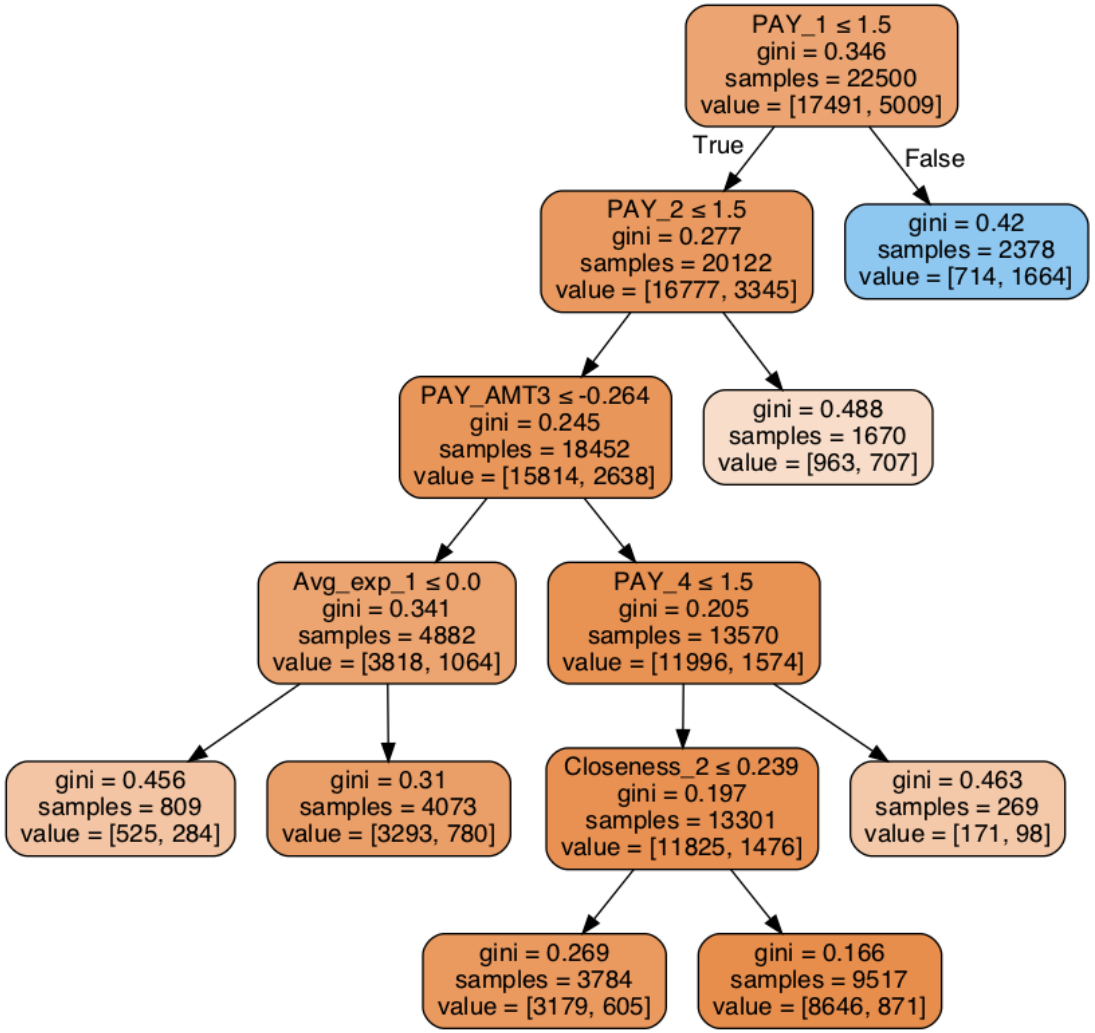

Trained data using a Decision Tree based model
Model is exported and saved as a Pickle file
Model is loaded on call
User Enters Data in HTML/CSS/Javascript UI
Data is sent to the trained model on
Model predicts the probability
Probability is saved in SQLite
Result is fetched fron SQLite, back and displayed on the UI

Statistical Techniques Tried

Model	Accuracy	Precision	Recall	F1 Score	ROC
Logistic Regression	.80	.65	.20	.31	.59
K-Nearest Neighbhors	.81	.63	.24	.35	.60
MLP	.83	.64	.39	.48	.66
Decision Tree Classifier	.83	.71	.31	.44	.65

Why Decision Tree?

- Highest Accuracy Score
- Best type for Financial related problems
- Know what the classifer values are



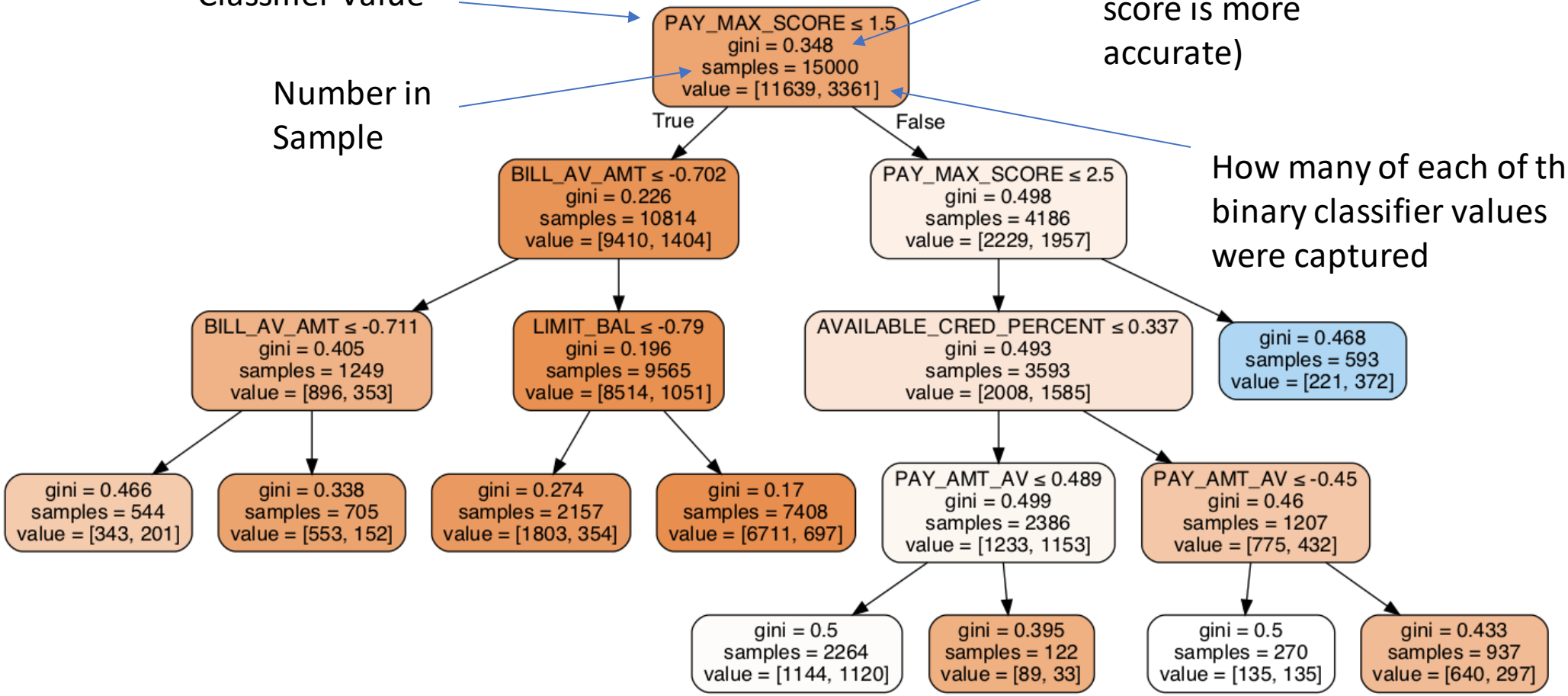
New Decision Tree

Classifier Value

Number in Sample

Measure of how well it can capture the target,(lower score is more accurate)

How many of each of the binary classifier values were captured



Training Process and Exporting Model

Feature Engineering

- Column Values Adjusted
- Preferred Value Types Used

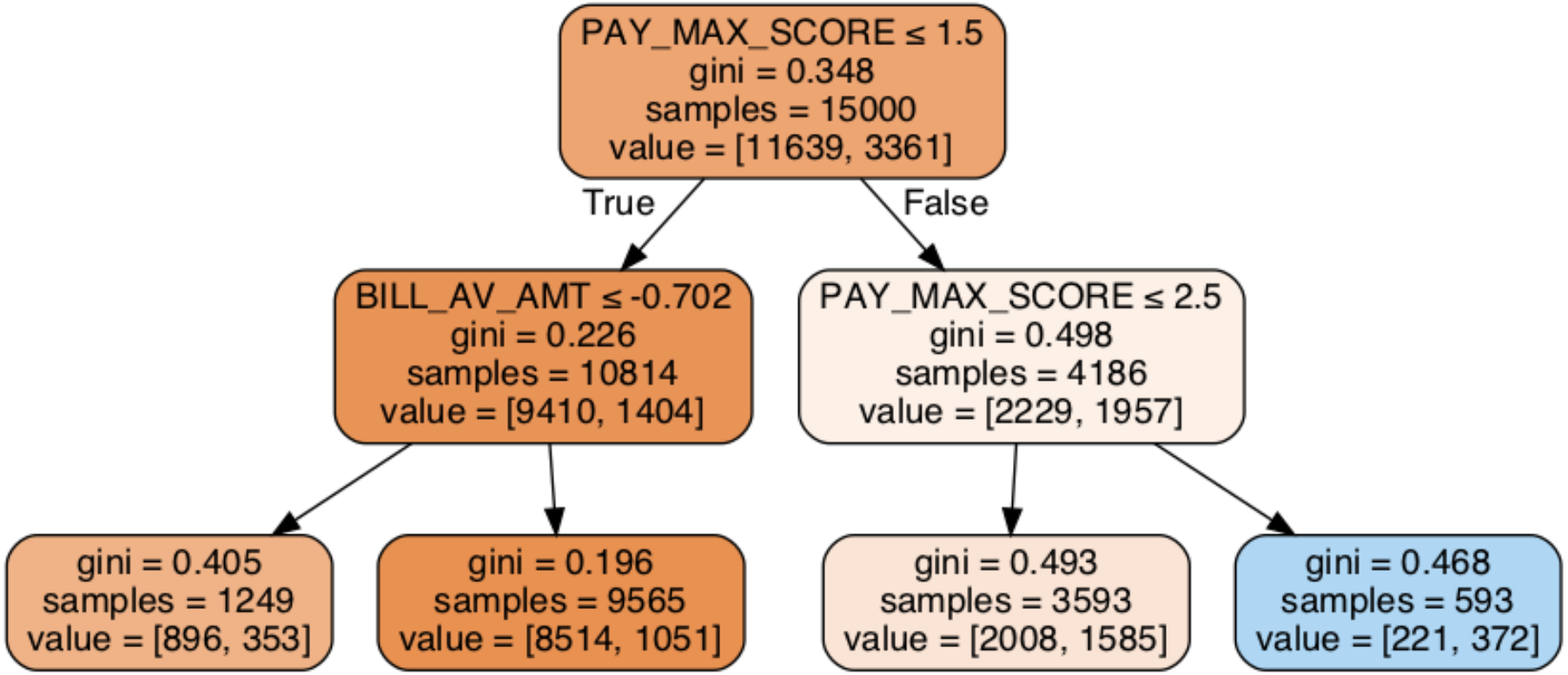
Training and Validation

- Library Searches for best Decision Tree Selectors
- Data is Tested with testing set getting 83% accuracy

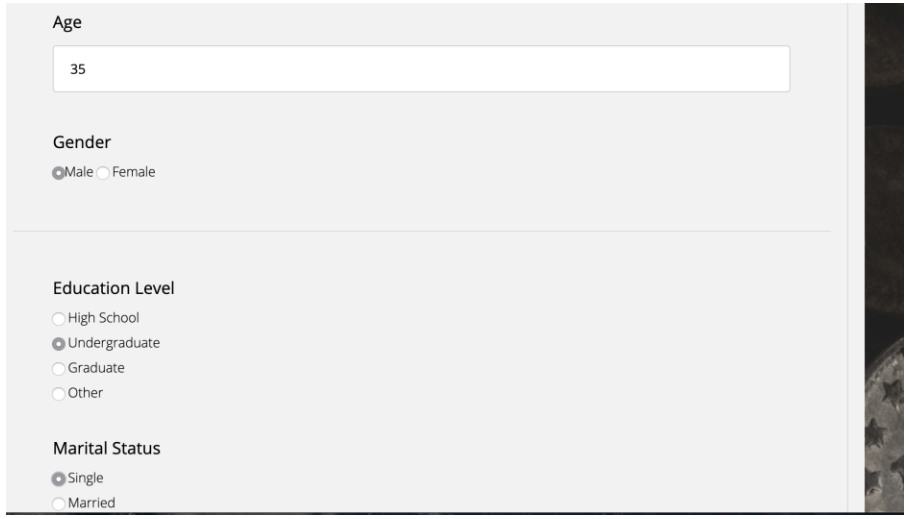
Model is exported to Pickle file

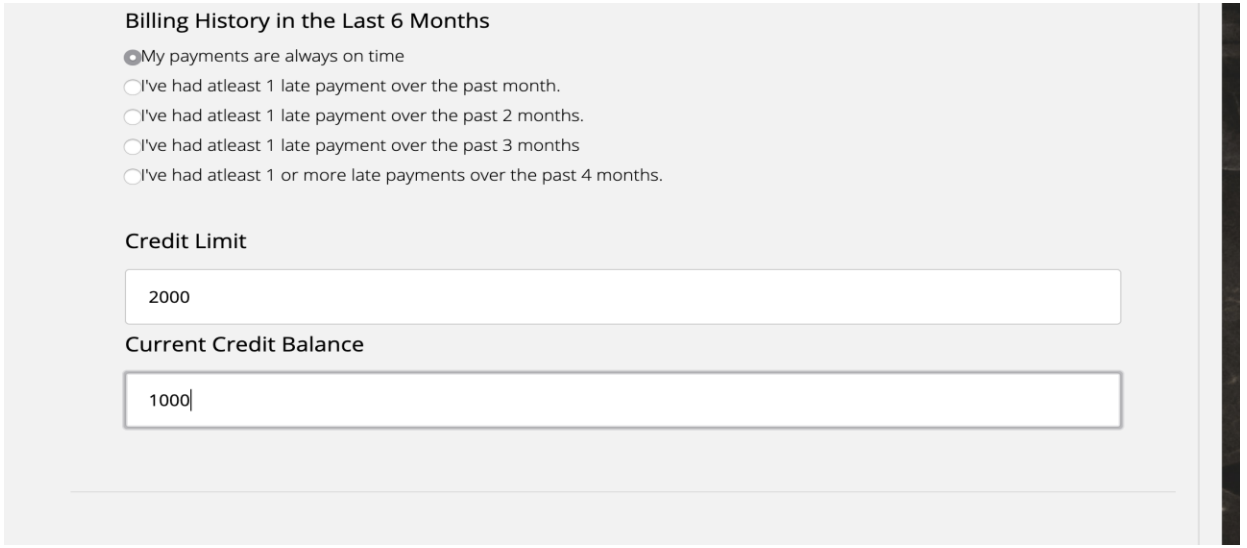
- Pickle file allows it to be called whenever it's needed
- File is used to calculate default probability

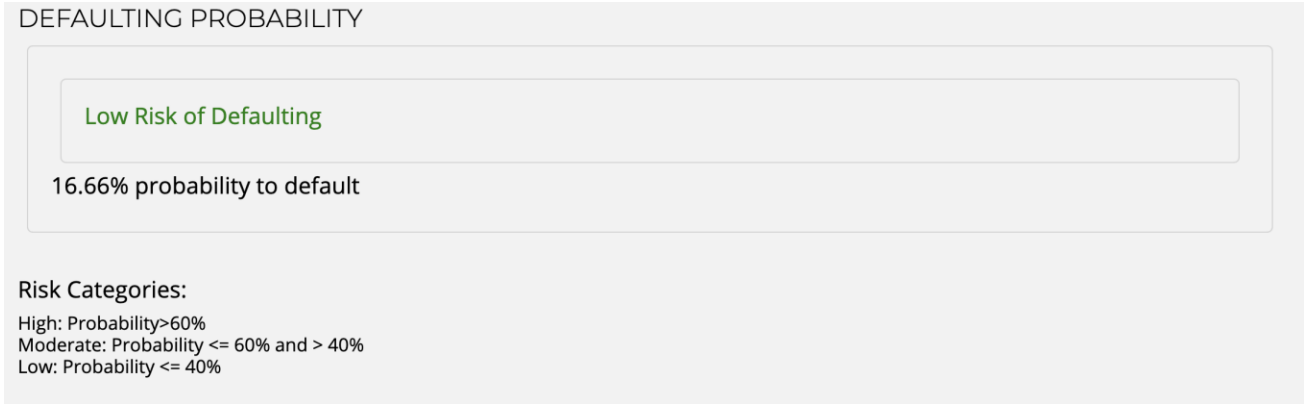
Other Tree Models Tried Grid Search



Test Case of Website

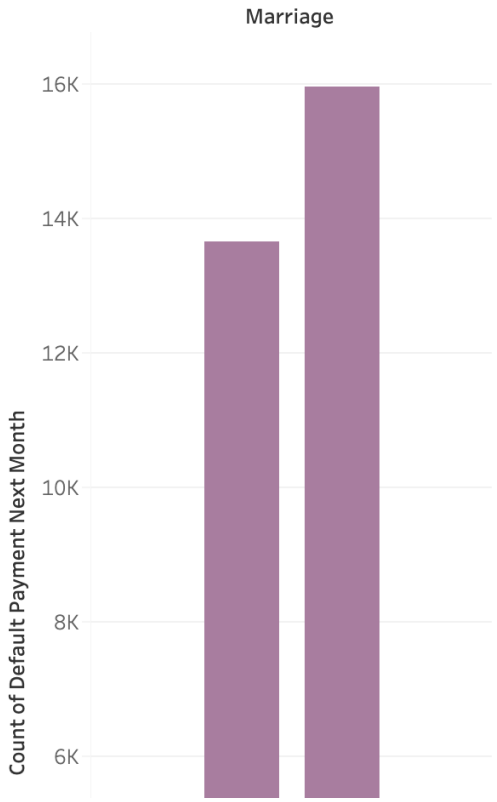


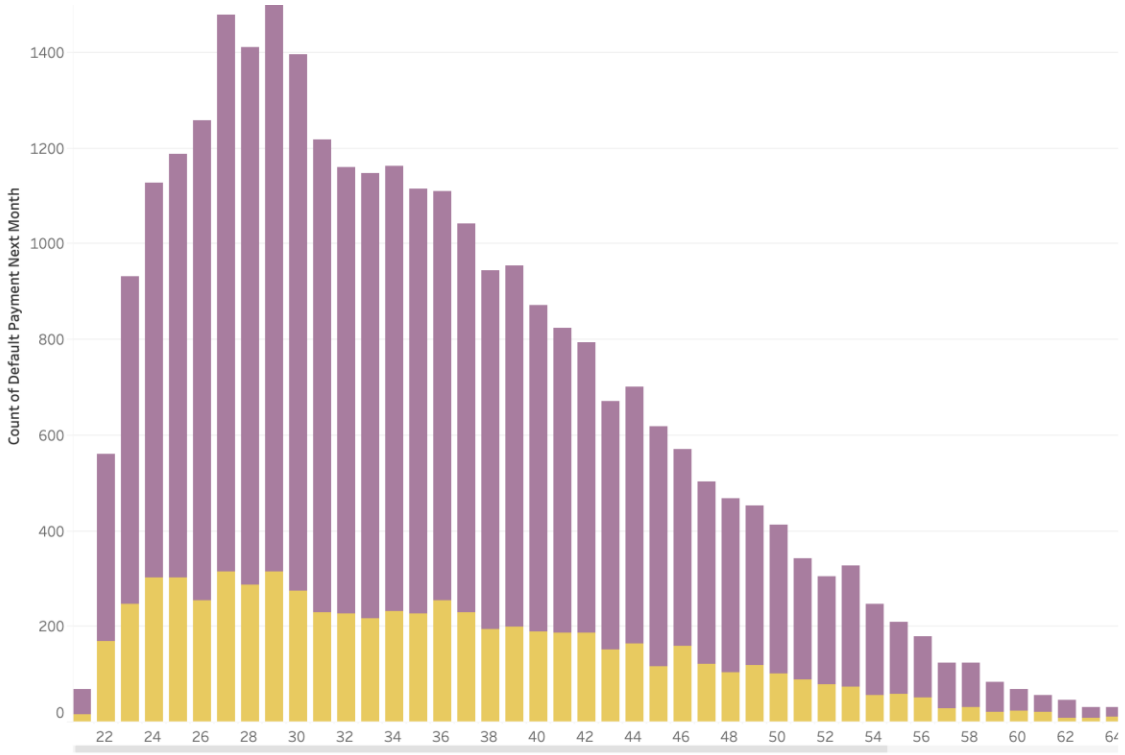




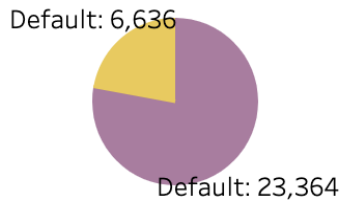
Other Output Statistics on Website

Default Status by Marital Status





Total Records by Default Status



Questions?