Business Problem

Knowing which customers are likely to default on their credit card payments is useful for the credit card company. The company may be able to intervene with certain customers to help increase repayment, thus ensuring revenue and minimizing losses. The company could try several tactics for reducing the default rate. For example, customers that are “at risk” of missing a payment could be sent a reminder(s) about their upcoming payments and information on how to keep their account on good terms.

Conversely, knowing which customers will be unlikely/unable to pay their upcoming bill may be useful as the company may be able to save time and money by not actively targeting these customers for personalized repayment campaigns.

Of course, customers that are highly likely to continue to pay their account as agreed upon do not need to be actively targeted for repayment or contacted. This can save the company time and money.

Dataset

The data to predict credit card defaults is from the UCI website. There are 30,000 records along with the following columns:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
				1=m		9/2005	8/2005	7/2005	6/2005	5/2005	4/2005	9/2005	8/2005	7/2005	6/2005	5/2005	4/2005	9/2005	8/2005	7/2005	6/2005	5/2005	4/2005	
	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_BAL	SEX	EDUC	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next mon
1	$20,000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
2	$120,000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
3	$90,000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
4	$50,000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
5	$50,000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0

- limit_bal: Amount of the given credit (New Tawianese dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- sex: Gender (1 = male; 2 = female).
- ed.: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- married: Marital status (0 = single; 1 = married; 2 = others).
- age: Age (year).
- p1 - p6: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:
    - p1 = the repayment status in September 2005
    - p2 = the repayment status in August 2005. . .
    - p6 = the repayment status in April 2005.
        - The measurement scale for the repayment status is:
        - -1 = pay duly

        - 1 = payment delay for one month
        - 2 = payment delay for two months…
        - 8 = payment delay for eight months
        - 9 = payment delay for nine months and above.
- b1 - b6: Amount of bill statement (NT dollar).
    - b1 = amount of bill statement in Sept. 2005
    - b2 = amount of bill statement in Aug. 2005. . .
    - b6 = amount of bill statement in Apr. 2005.
- pymt1 - pymt6: Amount of previous payment (NT dollar).
    - pymt1 = amount paid in Sept. 2005
    - pymt2 = amount paid in Aug. 2005. . .
    - pymt6 = amount paid in Apr. 2005.

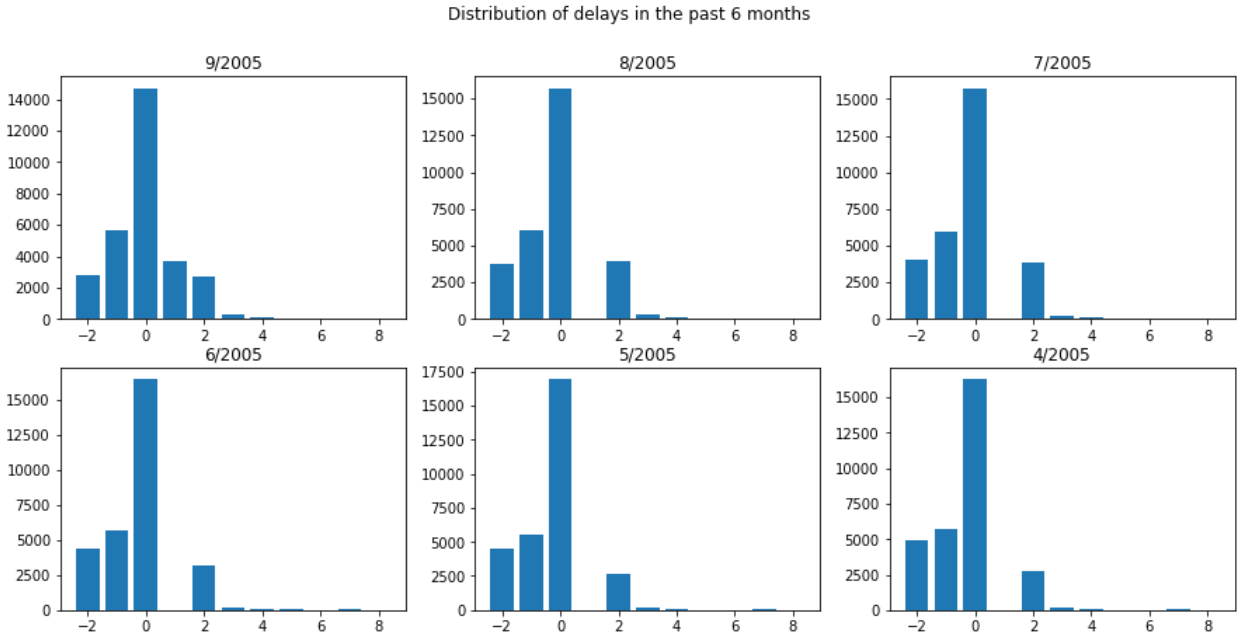
A quick rundown of things that I did with the data:

1. Plotted payment delays by month
2. Plotted bill amounts by month
3. Described bill amounts
4. Described payment amounts
5. Plotted late payments by marital status (total and percent)
6. Plotted late payments by age (total and percent)
7. Plotted late payments by sex (total and percent)
8. Plotted late payments by education (total and percent)
9. Plotted late payments by credit limit (total and percent)
10. Added a missed payment column
11. Plotted late payments based on balance amount (total and percent)

Initial Findings

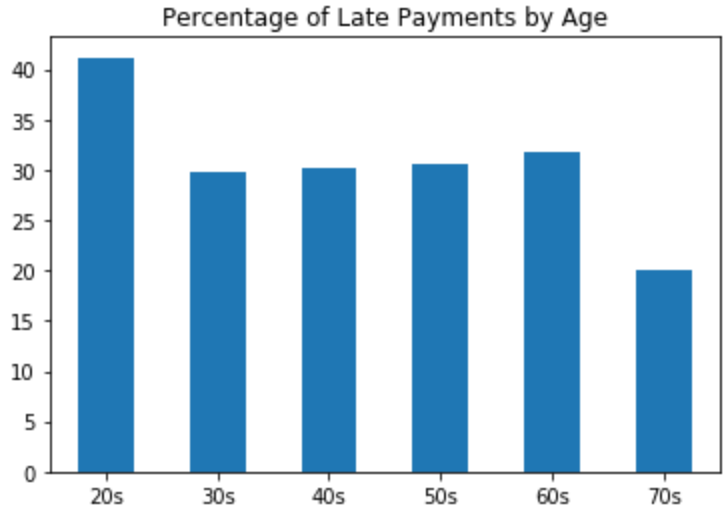
Exploratory Data Analysis has lead to several interesting findings. The data is somewhat unbalanced with roughly 20% of our customers predicted to default.

The payment delays is missing a lot of “1’s” values (see below). We also see that most customers pay on time (0) or ahead of schedule (negative values).



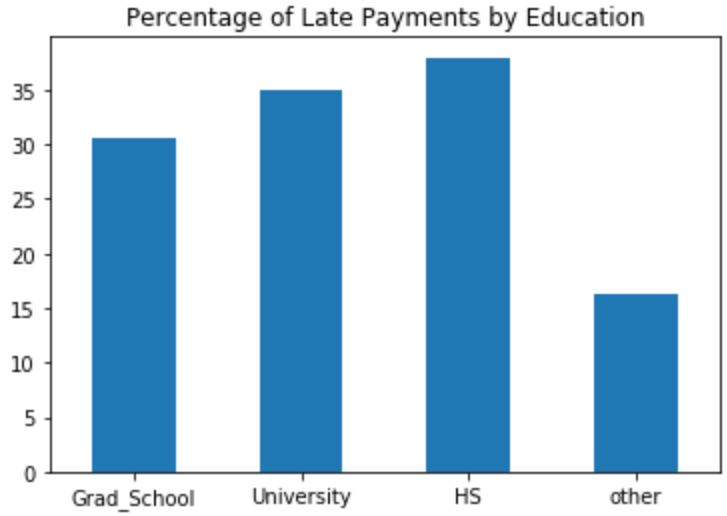
What percentage of people have late payments based on marital status?
Single: 33.5 %
Married: 33.6 %
Other: 36.3 %

What percentage of people have late payments based on age?

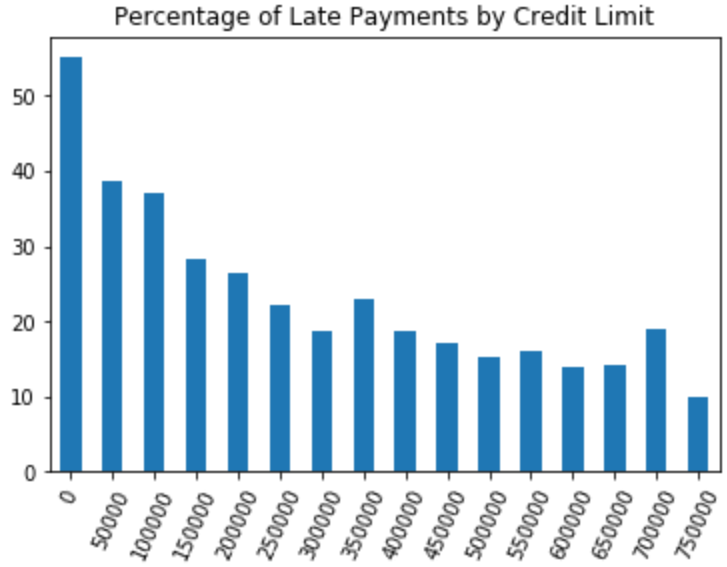


What percentage of people have late payments based on sex?
Male 35.2
Female 32.5

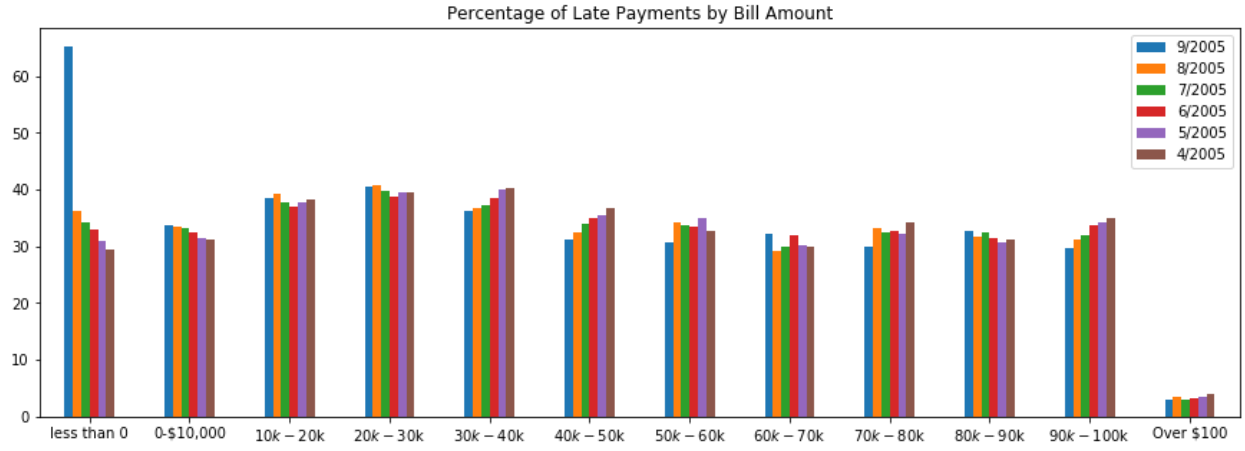
What percentage of people have late payments based on education?

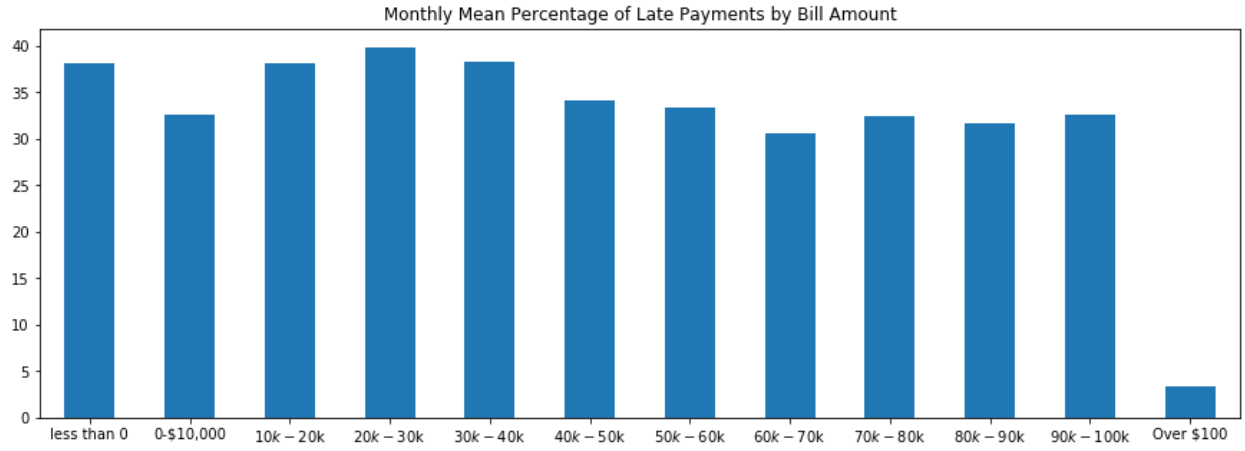


What percentage of people have late payments based on credit limit?



What percentage of people have late payments grouped by bill amount?





As we can see in the first plot, we have a pretty big outlier for 9/2005. This is likely due to incorrect data input. This value is roughly twice as big as the others which increases the mean in the category of “bill amount less than zero”. Otherwise we can see that bill amounts between 10K and 40K are where most of the late payments come from.

It would be interesting to see if there is a correlation between people’s bill amount and their bill amount as a proportion of their credit limit. This may give us some insight into whether these customers are able to pay their bill as agreed.

Further Analysis

I’ve considered calculating the mean of payment amounts, or mean of the balance, and then conducting bootstrapping on those variables, but after thinking about it I’m not sure this is a good approach. We aren’t trying to find the difference in the means, or whether or not means or

equal, or any sort of similar metric. What we have is a classification problem, which is best solved with predictive analytics and machine learning.

Next Steps

As we can see, we have many explanatory variables to use for building a model. Logistic regression and other machine-learning-based approaches will be used to predict which customers are most likely to default.

Additional features will also be created, specifically calculating the current bill amount as a proportion of the customer's credit limit may be a good indication of whether someone has enough money to make a payment.

Presentation Slides

https://docs.google.com/presentation/d/1xnHyM4ZIaiessxmYNi-VLllTJcg8yYjEUGaqLM89z88/edit?usp=sharing