Business Problem:

The client, a new card-issuing bank, is seeking to improve their applicant quality by reducing the number of individuals who are likely to default on their credit card accounts when they apply for a new account with the bank. A credit card default is defined as when an individual has become severely delinquent on their credit card payments, which is a result of consistently missing the required monthly payments.

Customers who default on their credit card accounts can have a strong negative impact on the business for a card-issuing bank. Cardholders who have overused their credit card consumption and accumulate heavy cash debt pose a challenge to the bank looking to increase market share, improve consumer confidence, and optimize their cash reserves just to name a few examples.

As a new card-issuing bank in the marketplace, customers who default may pose a higher risk than an already established and well-recognized card-issuing bank. Performing an analysis on past consumer payment data, which includes both individuals who defaulted and did not default on their credit account, can help towards decreasing the number of default accounts.

Problem Statement:

Create a binary classification model that can accurately predict whether a credit card applicant is likely to default on their credit card account.

Dataset Description:

The dataset used for this problem statement is the *Default of Credit Cards Clients Data Set* provided by researchers at Chung Hua University and Tamkang University, Taiwan and made available in the UCI Machine Learning Repository. This dataset is sourced from real customer data in the year 2016 of a Taiwan-based credit issuer. This dataset contains a total of 30000 records of row-data for each customer credit card account with 24 attributes recorded for each customer.

The attribute description as provided by the researchers of UCI dataset description:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = 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.
X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

The dataset comprises exclusively of numeric values, including categorical variables such as education, marriage, and sex of the account owner encoded as numeric.

A sample of the original dataset as seen in the first few rows:

```
data.head(10)
```

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
0	1	20000	2	2	1	24	2	2	-1	-1	...	0	0	0	0	689	0	0	0	0
1	2	120000	2	2	2	26	-1	2	0	0	...	3272	3455	3261	0	1000	1000	1000	0	2000
2	3	90000	2	2	2	34	0	0	0	0	...	14331	14948	15549	1518	1500	1000	1000	1000	5000
3	4	50000	2	2	1	37	0	0	0	0	...	28314	28959	29547	2000	2019	1200	1100	1069	1000
4	5	50000	1	2	1	57	-1	0	-1	0	...	20940	19146	19131	2000	36681	10000	9000	689	679
5	6	50000	1	1	2	37	0	0	0	0	...	19394	19619	20024	2500	1815	657	1000	1000	800
6	7	500000	1	1	2	29	0	0	0	0	...	542653	483003	473944	55000	40000	38000	20239	13750	13770
7	8	100000	2	2	2	23	0	-1	-1	0	...	221	-159	567	380	601	0	581	1687	1542
8	9	140000	2	3	1	28	0	0	2	0	...	12211	11793	3719	3329	0	432	1000	1000	1000
9	10	20000	1	3	2	35	-2	-2	-2	-2	...	0	13007	13912	0	0	0	13007	1122	0

Data wrangling steps included looking at the summary statistics, identifying recorded values that may have been encoded incorrectly and removing outliers. Some of the steps taken in the data wrangling:

- ID column can be dropped, this is likely the applicant's record id
- Bill amount shows a negative amount for minimum, this does not make sense as bill statement should always be a positive value. These rows were discarded.
- Features renamed for better clarity and consistency. I.e. all lowercase, 'default payment next month' to 'defaulted'
- Multiple redundant header columns, only kept named columns.
- The education variable had values not identified in the original dataset description: 1 = graduate school; 2 = university; 3 = high school; 4 = others. Values of 4, 5, 6, and 0 were re-encoded to 4.

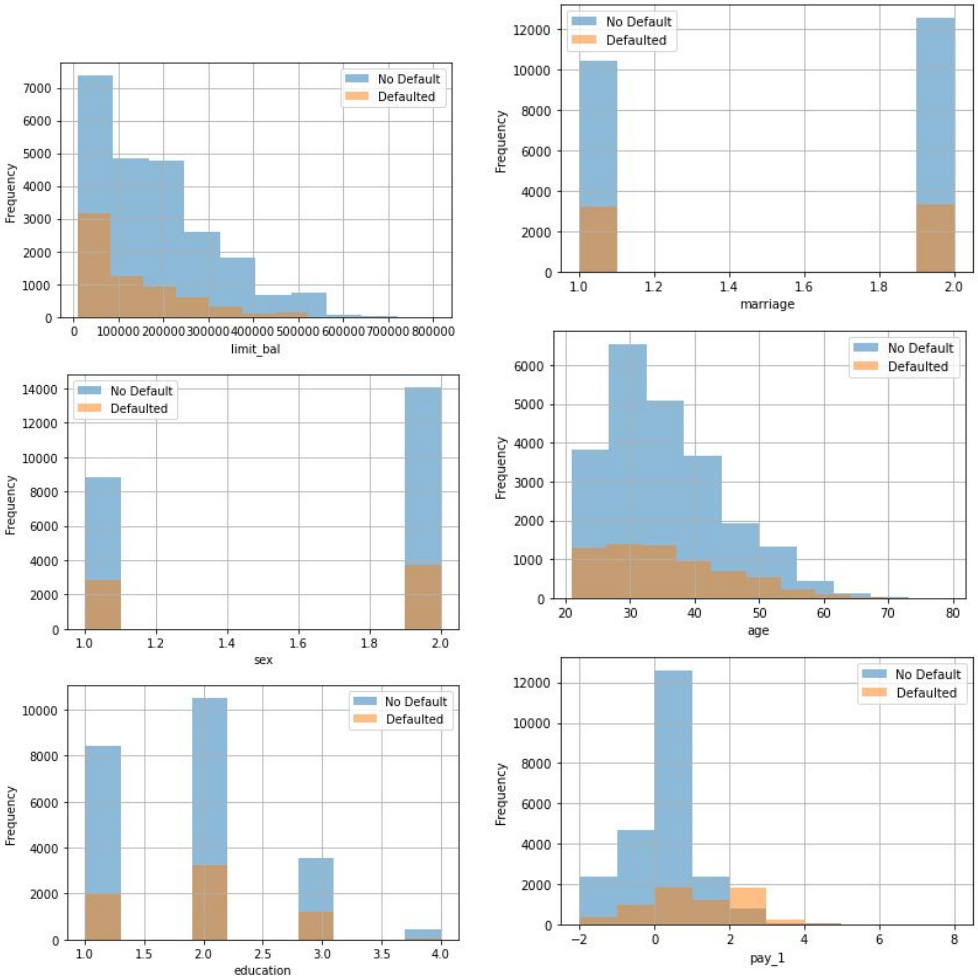
Initial findings from exploratory analysis:

Exploratory data analysis included looking at summary statistics of each variable and identifying any correlation among variables.

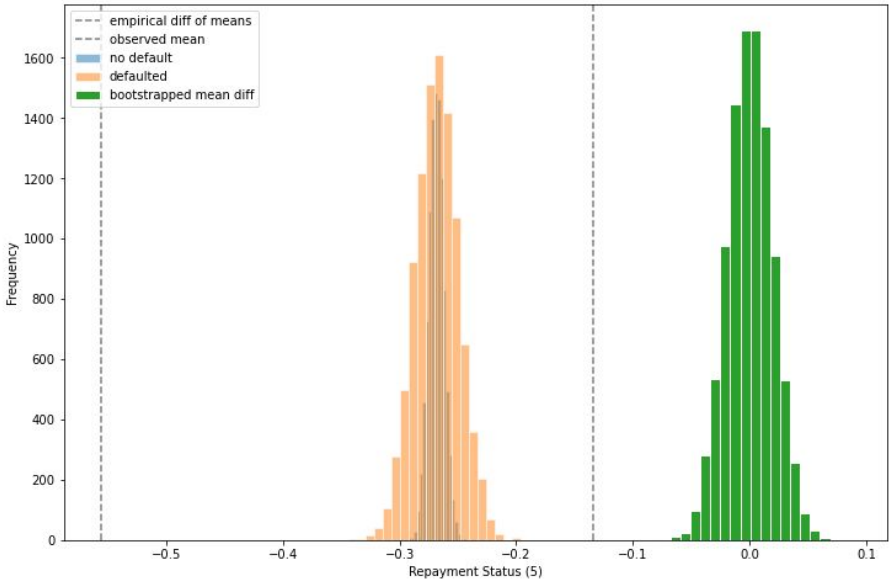
```
defaulted_group_corr = defaulted_group.corr()
defaulted_group_corr[ (defaulted_group_corr.iloc[:,:]>=0.5) & (defaulted_group_corr.iloc[:,:] < 1.0) ].dropna(how='all')
```

	limit_bal	sex	education	marriage	age	pay_1	pay_2	pay_3	pay_4	pay_5	...	bill_amt4	bill_amt5	bill_amt6	pay_amt1	pay_amt2	pay_amt3	pay_amt4	pay_amt5	pay_amt6	defaulted
pay_1	NaN	NaN	NaN	NaN	NaN	NaN	0.652074	0.550214	0.513598	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pay_2	NaN	NaN	NaN	NaN	NaN	0.652074	NaN	0.766482	0.629877	0.593002	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pay_3	NaN	NaN	NaN	NaN	NaN	0.550214	0.766482	NaN	0.777553	0.680952	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pay_4	NaN	NaN	NaN	NaN	NaN	0.513598	0.629877	0.777553	NaN	0.838514	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pay_5	NaN	NaN	NaN	NaN	NaN	NaN	0.593002	0.680952	0.838514	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pay_6	NaN	NaN	NaN	NaN	NaN	NaN	0.558582	0.631310	0.738156	0.841192	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bill_amt1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0.906817	0.877081	0.851740	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bill_amt2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0.929692	0.899672	0.874567	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bill_amt3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0.961397	0.931209	0.904864	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bill_amt4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	0.964750	0.937656	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bill_amt5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0.964750	NaN	0.971606	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bill_amt6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0.937656	0.971606	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

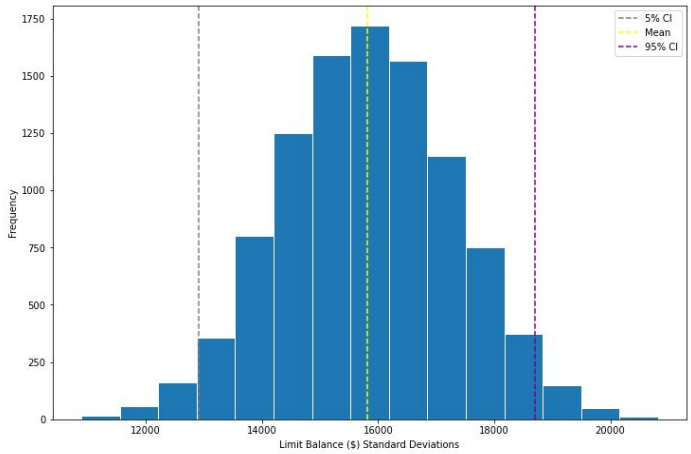
As seen above, a correlation matrix identified existence of multicollinearity among the bill amount variables (BILL_AMT*) and therefore could be considered to be excluded in the modeling stage. The dataset was identified to have a disproportionate number of accounts which were identified as non-default versus non-default (~80% vs 20%) which would need to be strongly considered for resampling of the dataset for certain data mining algorithms chosen in the modeling stage.



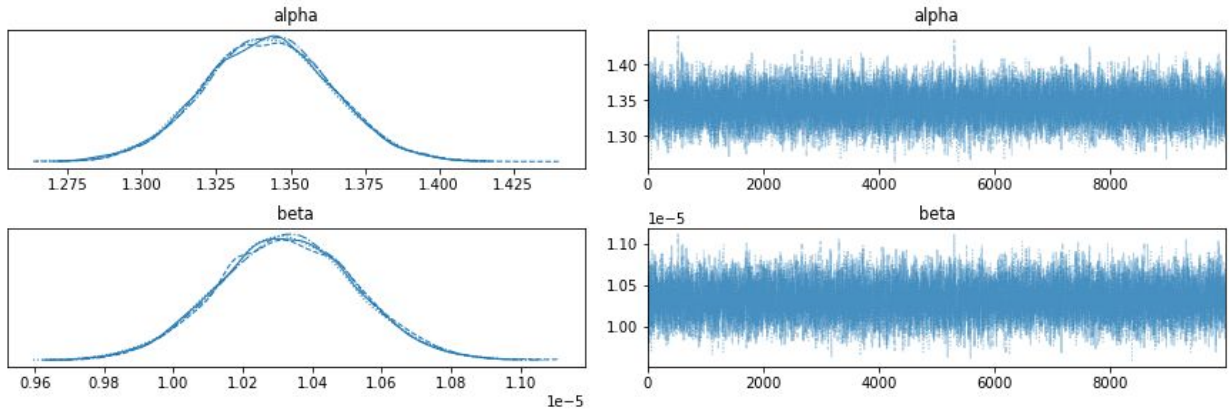
Looking at the frequency distribution of each variable among non-default versus non-default groups does not show a very strong apparent contrast in behavior. Repayment status (pay_*) and the account limit balance (limit_bal*) appear to have more stark contrast from histogram plots and were used for inferential testing.



Inferential tests including using the bootstrap sampling and Bayesian inference methods. A two-sample bootstrap test for repayment status (pay_1) using 10000 bootstrap samples using a mean adjustment confirms however that there is not a significant difference among non-defaulters and defaulters. As seen above, interestingly the bootstrapped samples show a deviation from the empirical mean, which may indicate there are multiple factors at play and we should consider other sampling methods considering the unbalanced dataset. A two-sample bootstrap test for limit balance (limit_bal), however, indicates there is a significant different among non-default and default groups by calculating a p-value of 0, where the bootstrapped samples indicate a 95% percentile for limit balance standard deviations in the range: [127200.48197576, 132730.17425986]) and the empirical standard deviation as near 16000:



Similarly, using the PYMC3 stats library, a Bayseian inference model indicates the 95% credible interval for defaulters had a limit balance between 130~140K, which is a comparatively larger difference than non-defaulters (mean of 177621). The resulting traceplot of taking 10000 samples of the alpha and beta posterior as show below:



A strong consideration as a result of the data wrangling and inferential statistics tests is to consider bucketing and grouping the dataset by certain education and limit balance brackets, as well as considering other sampling methods such as SMOTE due to the imbalance dataset. An interesting result of this preliminary analysis is that of the 24 variables, summary statistics and inferential statistics did not provide a strong contrast in behavior among the default and non-default groups. Looking at different subsets of the dataset, however, may be able to garner more apparent differences, but would be tedious to follow through and can be seen as a limitation of having a limited feature set and of inferential statistics.