Business Understanding

Credit cards are risky operations in the financial industry because the default rate is completely dependent on customers' behaviors. In this project, we aimed to predict credit risk on the customer level and decide whether a customer is 'good' or 'bad' based on how likely they are paying off on time. By using our model, the financial institution can plug in new applicant's information to determine whether the applicant is good, further to decide whether to approve the credit card application.

Other than customer payment history, we used customer demographic information in the modeling. However, we are aware that there are two issues for including such personal information. Firstly, there are strict regulations on the financial industry's modeling. Generally, it is not allowed to include sensitive information such as race, income, etc. So, we removed the sensitive information and categorized income as levels instead of the exact number, which could decrease the model performance. Secondly, the personal information is not updated most of the time. For example, if a customer joined two years ago who worked at company A, there is a large likelihood that the customer did not update with the bank even though he/she changed a job. This is a large limitation on the model performance.

Interpretability is another concern when building models for the financial industry. Besides powerful and complicated modeling such as Random Forest and CatBoost, we also built logistic regression because of its transparency and interpretability. Although the model performance may not be as good as complex modeling, it is important to convey our result to a non-technical audience.

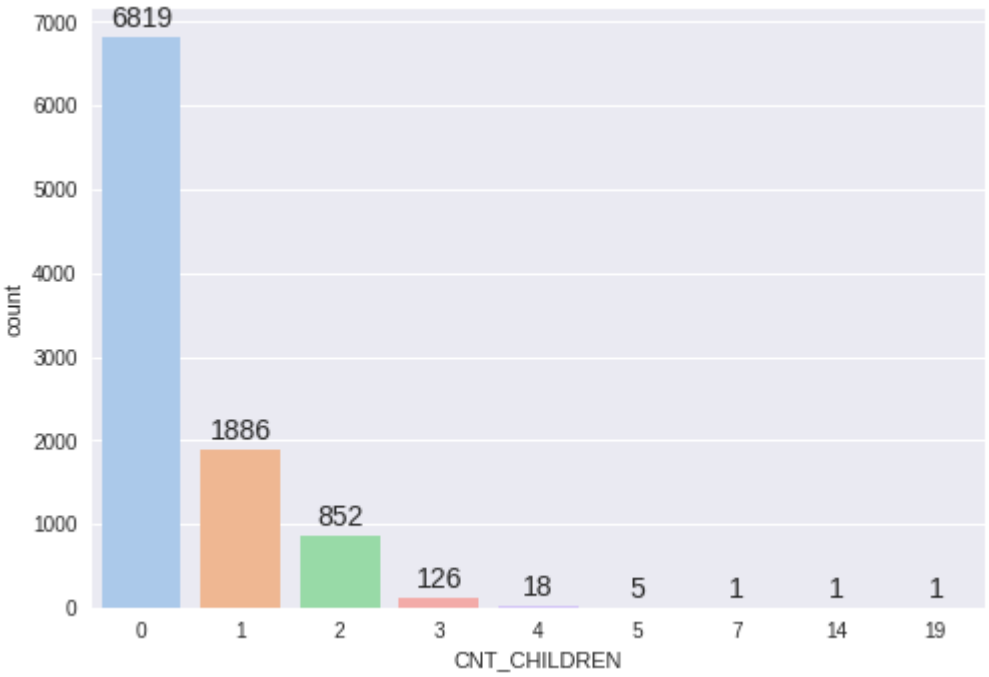
Data Understanding

Data Exploring (Variable vs. Target Variable)

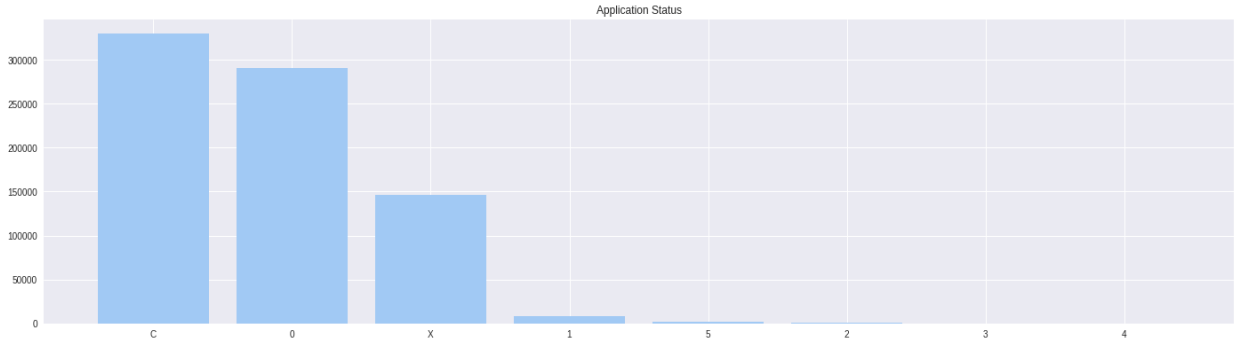
After we had clarification on the business problem, we used data from Kaggle on a credit card application to address the problem. Before the modeling, we did model-free exploratory data analysis to have a better understanding of the data. Our project contains two datasets, one contains various customer personal information like education level, occupation type and other typical demographic data, as well as financial information like income and types of assets the person has; the other dataset contains the credit records for each applicant that help us to understand customer historical payment balance records.

Since our business problem is related to credit card applications, we only extracted those who filled in the application form and analyzed them. After the extraction we started our model free data exploration on 9,709 distinct applicants.

Firstly, we took a look at the customer demographic information to find out what the major customer group looks like. We analyzed the binary and categorical variables such as gender, income type, education level, housing type, family status and occupations and found that the majority types of applicants are female (65%), married(67.2%), own realty (67%) and has zero or one child (89%). Below is one of the visualization that we created for understanding number of children for each applicants:



Since we decided to use the application status as our analysis label, we also wanted to have a look at the distribution of this variable. From the chart below we can find that most of the customers would pay their loan in a month. (X: No loan for the month; C: paid off that month; 0:1-29 days past due)



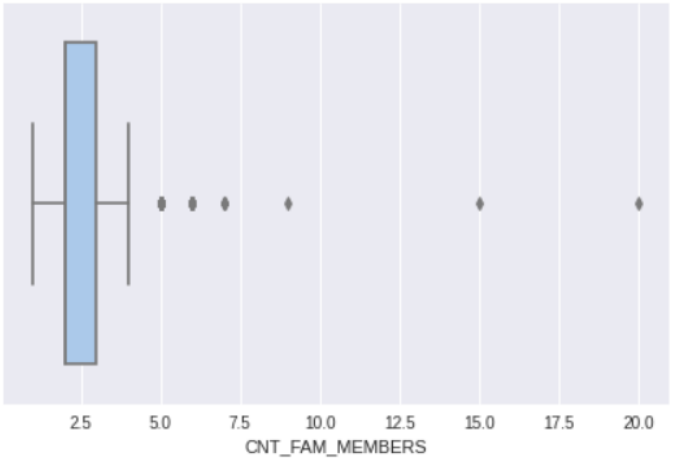
Next, we merged the two datasets and mapped the application status to a binary variable with 0 and 1 where ‘ 0’ are for bad customers who have passed due and ‘1’ indicated customers with great credit records to explore the relationship between target variable and other features. We found that for binary features, the bad customer rate seems to be similar for both types. For example, from the chart below we listed that the overall bad customer ratio for Male(Code_Gender=1) is 13.3 % where the Female applicants'(Code_Gender=0) bad customer ratio is 13.7%.

	feature	type	Bad customer in type	count	Bad_count
0	CODE_GENDER	0	0.137277	6323	868
1	CODE_GENDER	1	0.132605	3386	449

We also analyzed the numerical feature ‘number of children’ relation and found that applicants with over 2 children seem to have a higher chance of payment overdue. The bad customer rate for families with 0 or 1 children only has a 13% chance for having payment overdue, but 15% of applicants with 3 or more children have passed due and are categorized as bad customers in our analysis.

	children_count	bad_count	Good_count	Bad_rate
0	6819	921	5898	0.135064
1	1886	250	1636	0.132556
2	852	122	730	0.143192
3	152	24	128	0.157895

Besides, we checked the summary statistics and plotted out the distribution to detect outliers for the next stage. For example, we spotted that there are applicants having 20 family members that we should take into consideration for outlier processing.

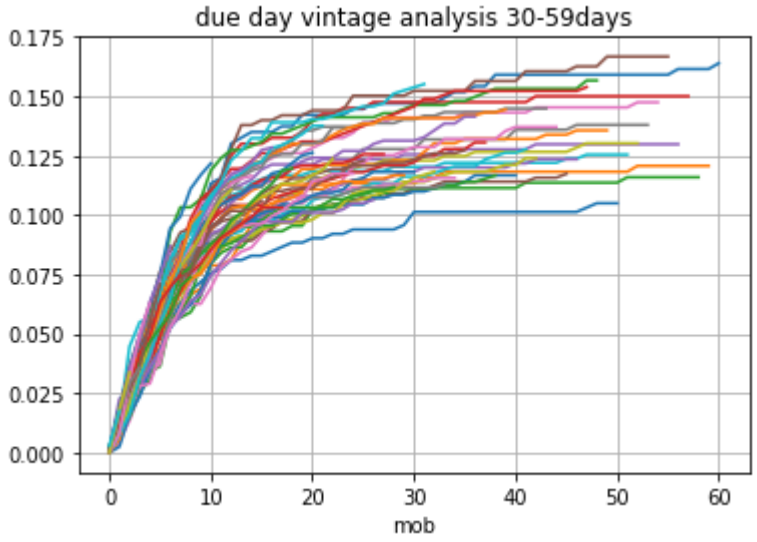


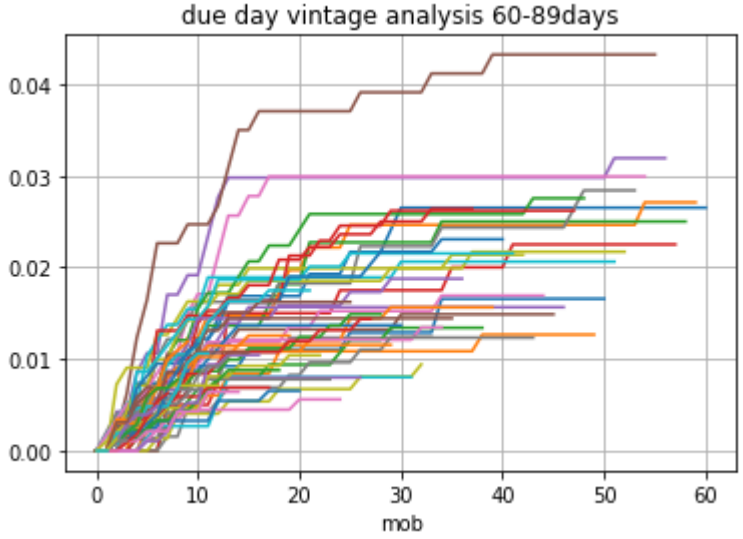
Vintage Analysis

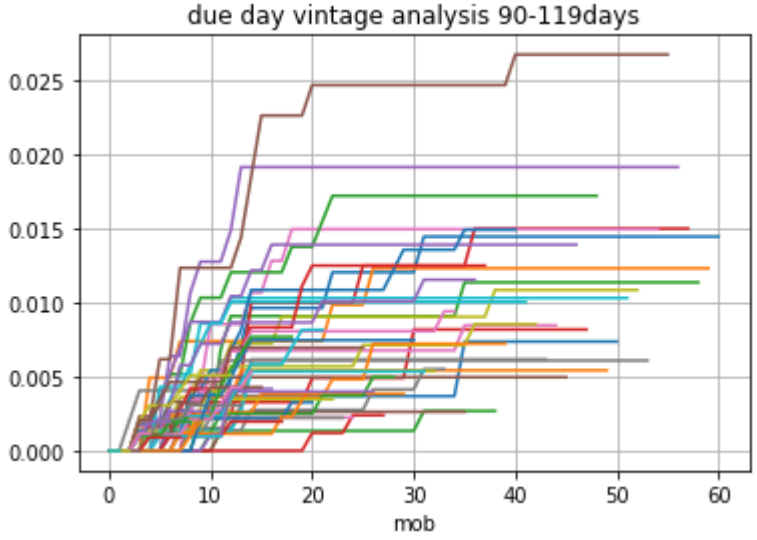
Vintage analysis is similar to cohort analysis that is commonly used in financial credit risk analysis. Vintage refers to the month in which the account was opened, and customers are being grouped based on the different vintages. Performance of the credit (e.g. overdue rate, y-axis) could be measured by each vintage (different lines) and by each month (x-axis).

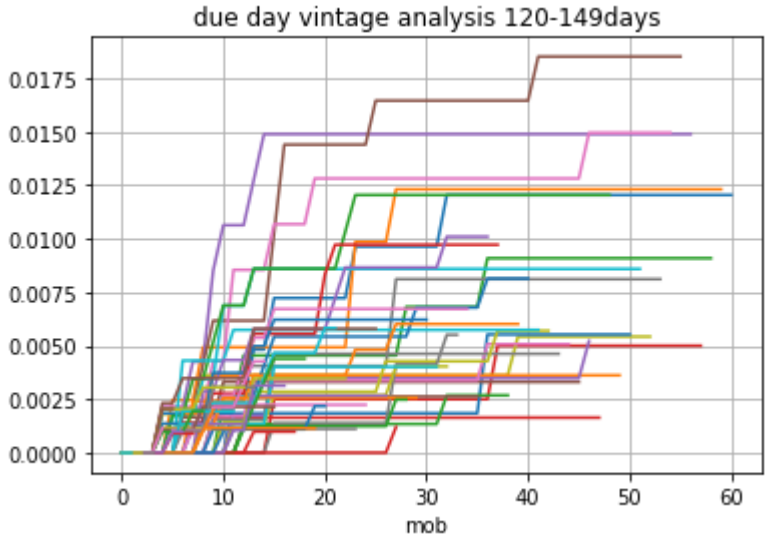
Multiple charts are displayed below. Each chart represented different levels of overdue periods, starting from 30-59 day to >150 days. Several insights could be drawn.

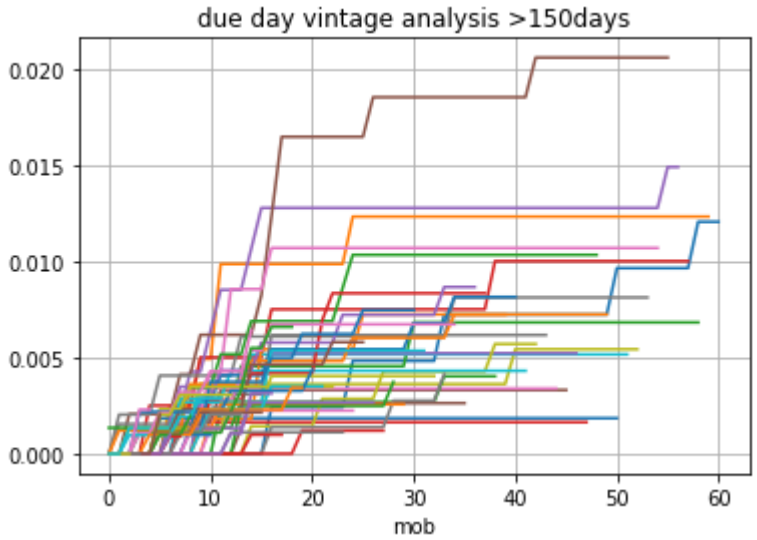
1. Firstly, customers will most likely overdue their payment in the first 20 months of business, followed with a relatively stable payment behavior.
2. Secondly, 30-59days overdue has the highest overdue rate of 17%, while >150days overdue has the least likelihood of ~2%. The number of days of overdue measures how bad the customer is and will be and it makes business sense that most customers may payback the overdue amount shortly.


due day vintage analysis 30-59days
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
10
20
30
40
50
60
mob


due day vintage analysis 60-89days
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
60
mob


due day vintage analysis 90-119days
0.025
0.020
0.015
0.010
0.005
0.000
0
10
20
30
40
50
60
mob





Note: vintage analysis is inspired by Kaggle project [EDA & Vintage Analysis | Kaggle]

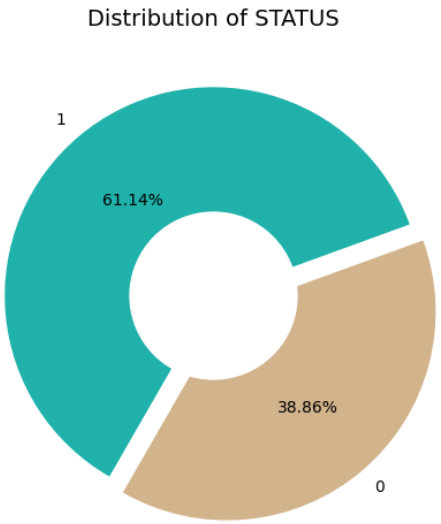
Data Preparation

Before going into the modeling, we did some data preprocessing and feature engineering to transform all features in a way that the model can take.

Target Variable Conversion:

Target variable in the origin dataset is application status and it has 9 values. For example, 0 means 0-29 days past due; 1 means 30-59 days past due. However, we are only separating customers into two categories here. 1 represents good customers, who have never passed due, while 0 for bad customers who have passed due.

We checked the percentage of each class in our dataset to detect potential data imbalance issues. Imbalanced data could decrease the model accuracy as the model will not have enough data points to learn the smaller class.



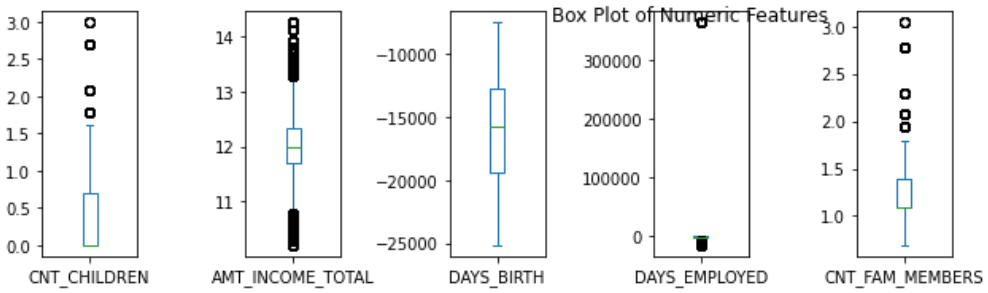
We can tell the data is pretty balanced and thus we don't need to do further steps to deal with it.

Missing Value

From the summary statistics we had before, we found there is a relatively high percentage of missing value in the occupation_type column. Around 30% data points are missing occupation types. Thus, we conducted missing value imputation. We don't want to drop this column because we think this could be an important feature from our domain knowledge. Instead, we created a new binary variable to indicate whether the value is missing. At the same time, we replaced the missing value with unassigned to indicate a special category in occupation_type.

Numerical Variable transformation

The distribution plot from EDA suggests that some columns are highly skewed. We would prefer a normal distribution for numerical variables and we did log transformation on those columns.



We can tell from the graphs that most columns successfully changed to a more normal distribution, however, there are some columns that need more preprocessing because of outliers.

Outlier

Outliers generally have a bad impact on the model as it may suggest some unusual pattern, which further confuses the model. Based on the previous distribution plot, we applied the

three sigma statistical method to filter out data points that are higher than 99.7 percentile or lower than 0.03 percentile.

Binning

Binning is a way to transform the numerical data into categorical variables that have similar size bins. Take the income feature as an example. $10,000 and $10,001 is not a big difference, but if we use 10,000 as a cut point the two people will belong to different levels, which is not reasonable. Instead, we used a statistical method to find the most reasonable threshold for categorizing. We found the quantile of the distribution and used it as a threshold to separate the numerical variable to a categorical variable with four levels.
One advantage of binning is that it enables the model to be more stable.
We first found the quantile.

Then used 25%, 50%, and 75% points as thresholds to map corresponding columns. For the DAYS_BIRTH(age) we divide all ages into 6 categories according to the threshold usually used in the marketing industry.

	count	mean	std	min	25%	50%	75%	max
CNT_CHILDREN	777715.0	0.3	0.4	0.0	0.0	0.0	0.7	1.7
AMT_INCOME_TOTAL	777715.0	12.0	0.5	10.8	11.7	12.0	12.3	13.2
DAYS_BIRTH	777715.0	-16124.9	4104.3	-25152.0	-19453.0	-15760.0	-12716.0	-7489.0
DAYS_EMPLOYED	777715.0	-1589.8	3065.6	-7583.5	-3292.0	-1682.0	-431.0	3860.5
CNT_FAM_MEMBERS	777715.0	1.1	0.3	0.7	1.1	1.1	1.4	1.8

Indexing/Encoding

To be able to run it for models like logistic regression that can't handle categorical variables, we would do one hot encoding for the applicable variables (creating a new binary column for every possible category value of the column). We had a total of 12 categorical variables we encoded. For models like CatBoost and random forest that are robust and can handle categorical variables, we would just index the categorical variables (replacing the category value with a number).

Weight of Evidence

The weight of evidence(WOE)and information value(IV) are two terminologies that appear often in the credit card related business.They have been used as the benchmark to see variables in the credit risk modeling. Here, we are using WOE to calculate IV in order to get the feature importance. 0.3-0.5 IV indicates strong predictive power, while IV less than 0.02 means not useful for prediction. The following table is the features with relatively high IV.

It gives us that MONTHS_BALANCE has the most predictive power among all the independent variables.

MONTHS_BALANCE
OCCUPATION_TYPE
days_birth_category
days_employed_category
NAME_EDUCATION_TYPE
AMT_INCOME_TOTAL_category
DAYS_BIRTH
AMT_INCOME_TOTAL
DAYS_EMPLOYED
NAME_FAMILY_STATUS
NAME_INCOME_TYPE
FLAG_EMAIL

Modeling

We ran three three main different types of models. We used random forest, CatBoost, and logistic regression models. Random forest and CatBoost models are part of meta-modeling techniques and will potentially provide higher accuracy, while logistic regression model is still being considered as it offers higher transparency and interpretability to financial regulators and business executives. We provide a balanced scorecard approach to better illustrate the pros and cons for three selected models.

For each model, we used nested cross validation, where we used the inner cross validation to optimize their applicable parameters with randomized grid search, and secondly ran for generalization performance. Using randomized grid search allowed us to yield a relatively better performance and find the best combination of hyperparameters for each model in the least possible time.

	CatBoost	Random Forest	Logistic Regression
Model Brief	This model would improve and support the business by providing the most accurate predictions for if a customer is more likely to default.	This model would improved the business's accuracy by giving a very quick, overall understanding of how likely a person is to default on average	This model is very interpretable so it would help employees and decision makers to know what key traits to be looking at for a person applying for a credit card
Interpretability	Harder to interpret because results from one tree affect the results of the next tree	Generally easy for business stakeholders to understand. Decision trees are more intuitive and based on rules/thresholds that you can explain rather than from a mathematical algorithm.	Most interpretable
Time	Took a significantly long time to fit model because	Took the quickest time to fit the model because it	Was able to fit the model relatively quick

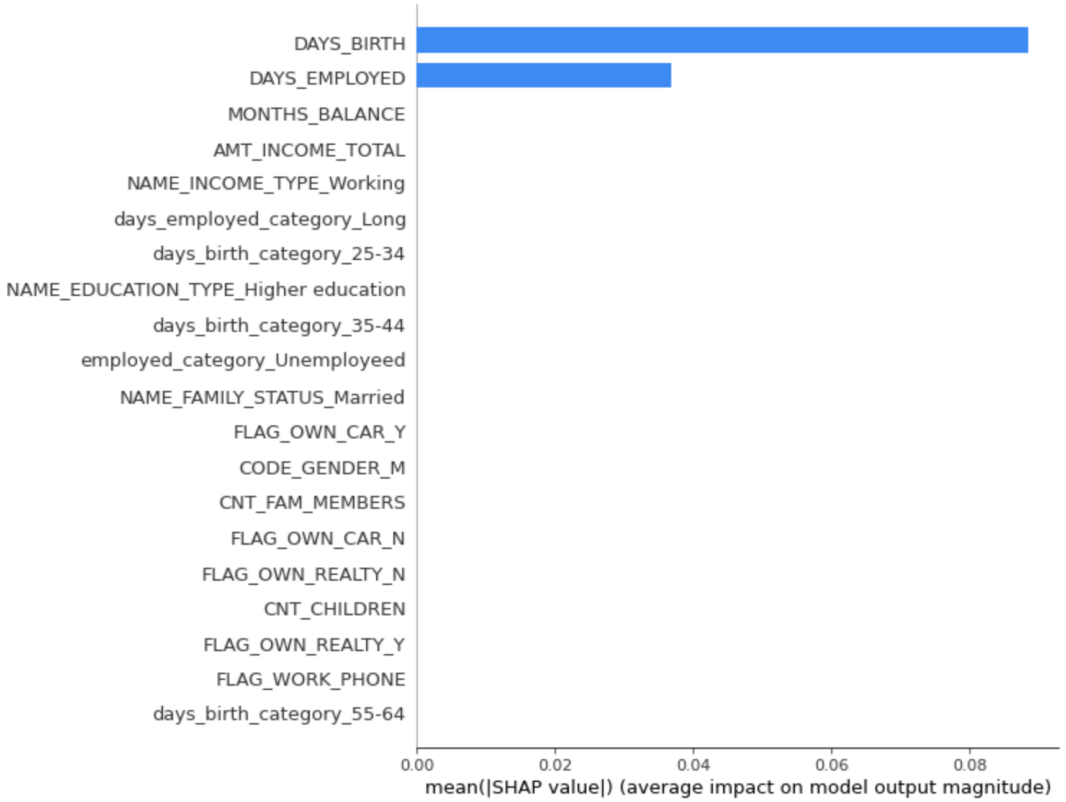
	cannot run in parallel	can run trees in parallel	but not as fast as random forest
Accuracy	Very robust and able to capture very complex patterns with high accuracy from boosting (putting more importance for correcting misclassification from previous model)	Is a pretty robust and complex algorithm. It fits multiple decision trees and takes the average of the results for making a prediction.	Not as accurate in this situation. Not able to capture such complex trends

Evaluation

	CatBoost	Random Forest	Logistic Regression
F1	.778	.763	.758
ROC/AUC	.73	.64	.5

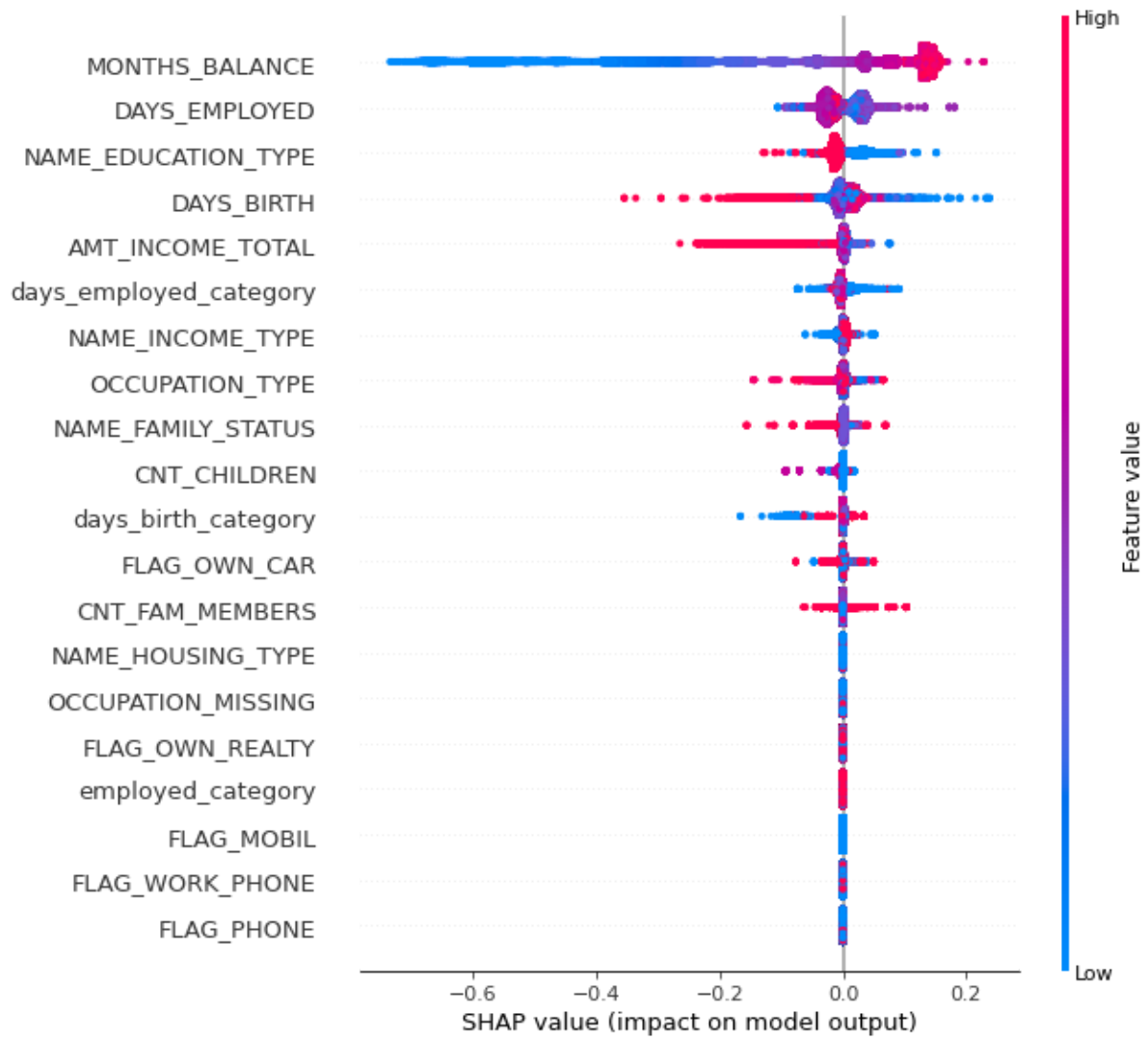
Several approaches can evaluate the performance of the modes.

Evaluating based on the F1 and AUC score, CatBoost is proved to be the best model while logistic regression seems to have the lowest performance. Logistic regression's AUC of .5 suggests that the model performance is no better than randomly guessing. Looking deeper into the model predictions, we can see that it predicted that every customer is a good customer. Such under-performance might be due to the fact that customer demographic information is not the most significant for predicting credit card payments with an algorithm not as complex as logistic regression. If we had more features that dealt with spending patterns like average amount spent per transaction or types of items frequently bought, the logistic regression model would produce much better results.

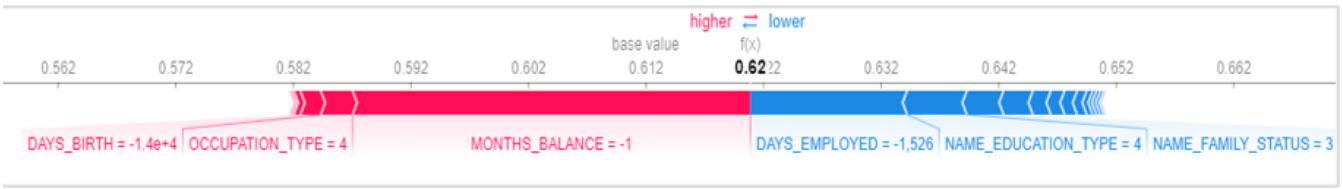


We then use shapley value to conduct a deep-dive analysis on the feature importance of the CatBoost model. Shapley value tells us the average marginal contribution of a feature value for prediction. For the chart below, we could learn that

1. The most important features are months_balance, days_employed, name_education_type, days_birth, and amt_inome_total. The top 1 feature, months_balance is also aligned with weight of evidence analysis.
2. For each feature, we could clearly see how the feature value is correlated with shapley value. For example, for months_balance, a newly registered customer is more likely to become a good customer with no overdue payment.
3. Days_employed_category is a feature we created by using the binning method, and this feature has a relatively high predicting power.



Furthermore, we also conducted individual-level shapley value analysis, which allows us to have a more granular view of how different features impact each customer's propensity to overdue. From the chart below, we randomly select a customer, and we could learn that this person has been employed for a little over four years and has a college education which supports that they're more likely to make their credit card payments. On the other hand, they are also about 38 years old and have a balance from last month they still haven't paid which are signs they might end up defaulting on their credit card payments in the future. Overall, this customer's prediction threshold from the model for defaulting on their payment is .62.



Lastly, we conducted model bias analysis on gender. It has been a rising issue in credit card scoring practice that the model will yield biased results towards male and females, and we work backwards to further investigate if our model produces significantly different outcomes for male and female

customers. It turns out that the F1 score for male and female customers have no significant discrepancy.

Deployment

There are several deployment strategies we can pivot towards:

1. Firstly, the model can be used to evaluate applications during the new customer acquisition stage. Customer demographic information such as employment and income show strong predictive power and the model will be a tool to help credit card firms separate good and bad customers.
2. Secondly, the model can be used for continuous monitoring. Existing customers will behave differently during their lifetime, as reflected by the vintage analysis, and the model will inform firms of who will be potentially overdue or not.

However, while the credit record information is constantly updated each month, customer information seems static and does not get updated. This might not generate a major issue for features such as gender and number of children, but for features including employment and income that is dynamic over the years, this will create bias and decrease predictive power of the model.

It is proposed that firms should collaborate with third party providers to ethically collect updated customer information or motivate customers to report their personal information update.

In addition, we should be aware of the pros and cons of customer profiling. It is true that collecting more customer demographic data will improve the model performance, but firms should avoid overuse and act within the ethical and legal standard. For example, race and ethnicity should normally be avoided. Also, firms should always regularly test if the model yields gender bias or not.

Team Contribution

Please evaluate the contribution of all team members of your team project. The evaluation should be objective and precise.

Luke He

Team Member Name	Contribution (%)	Comment
1. Michelle Wan	20	Modeling and evaluation. Troubleshoot on advanced modeling.

2. Yushan Yang	20	Feature engineering and data preparation. Research on advanced feature engineering techniques.
3. Ellie Wang	20	EDA and business understanding. Take lead on sliding and reporting.
4. Luke He	20	Modeling and evaluation. Research on vintage analysis.
5. Yufei Lu	20	Take lead on sliding and reporting. Business understanding and modeling.

Michelle Wan

Team Member Name	Contribution (%)	Comment
1. Michelle Wan	20	Did modeling and evaluation. Also contributed to report and slides
2. Yushan Yang	20	Did data cleaning and feature engineering. Also contributed to report and slides
3. Ellie Wang	20	Did early EDA and contributed to report and slides
4. Luke He	20	Also did modeling and evaluation with me. Contributed to the report and slides as well. Really helped to structure our slides professionally.
5. Yufei Lu	20	Took the lead on the report. Helped to clarify business problem and contributed to slides

Ellie Wang: We had clear task assignments and everyone did their part perfectly.

Team Member Name	Contribution (%)	Comment
1. Luke He	20%	Model & Evaluation
2.Yufei Lu	20%	Report lead and Model approach
3.Yushan Yang	20%	Feature Engineering
4.Michelle Wan	20%	Model & Evaluation
5.Ellie Wang	20%	EDA

Yufei Lu

Team Member Name	Contribution (%)	Comment
1. Luke He	20	Modeling and Evaluation
2. Yufei Lu	20	Report and modeling approach
3. Yushan Yang	20	Feature Engineering
4. Michelle Wan	20	Modeling and Evaluation
5. Ellie Wang	20	EDA

Yushan Yang

Team Member Name	Contribution (%)	Comment

1. Luke He	20%	Model & Evaluation
2.Yufei Lu	20%	Report lead and Model approach
3.Yushan Yang	20%	Feature Engineering
4.Michelle Wan	20%	Model & Evaluation
5.Ellie Wang	20%	EDA