Machine Learning Final Project

BAX 452

Prasad, Yashas, Harsh

Credit Card Customer Approval

Executive Summary

We have been hired by a fintech (internet banking) startup, FreeMoney Inc., based out of New York City. Since FreeMoney is an online bank, most of their business comes from credit card users. Recently, they have been dealing with a very high rate of default. This has started to worry the investors, and hence, FreeMoney has decided to hire us to solve this problem. We have been given data for around 45000 users with approximately 20 features for every user. We finally decided to use a simple logistic regression approach to predict as it has good accuracy and is simple and highly interpretable.

Problem Formulation

We will be utilizing the techniques we learned in the course to come up with a model to predict whether we should give someone the credit card or not, in other words, whether the user will default (to pay their credit card debt) or not. The final decision on which model to choose can depend on various factors. In our case, it is more important for us to detect accurately the users who are going to default rather than the users who are not going to default. This can also be explained as we care more about the true negative rate.

Background Information

Internet banking can be considered a product of the internet age. They have no physical branches and everything is done digitally. Credit cards form a big portion of the revenue of these companies. Discover is one of the most the biggest internet banks. It is a highly competitive market and companies have to frequently come up with new products and schemes. It is widely known that it is always cheaper to retain an old customer than to acquire a new one. This is why it is a very important problem for FreeMoney to solve and quickly. Since this is a matter of urgent importance, we will have to come up with a simple and quick solution.

Traditional solutions and Improvements

This business problem is not a new problem for the fintech industry. Data scientists have been trying to solve this problem for a long period of time. The simplest solution that has been implemented is to look at a combination of their credit report and monthly income. FreeMoney has provided us with information other than financial metrics. We believe that by just taking into account we are missing a lot of latent features which might have a strong impact on a person's ability to pay back the loans. We used various techniques to figure out the simplest model to improve the default rate.

There are various steps involved in this process:

1. Initial Exploratory Data Analysis (EDA)
2. Removal/Imputation of Null Values
3. Feature Engineering
4. Secondary EDA
5. SHapley vAlue exPression (SHAP) value for feature importance (To decide on the first few variables for the stepwise model, helps keep the process simple)
6. Run different models and decide on the best model for the use case
    a. Logistic regression (with stepwise regression)
    b. Random Forrest
    c. XGBoost

1. Initial EDA

This step involved looking at the data from a very broad perspective. We realized that there are 4 columns with Null values. We dealt with those in the next section.

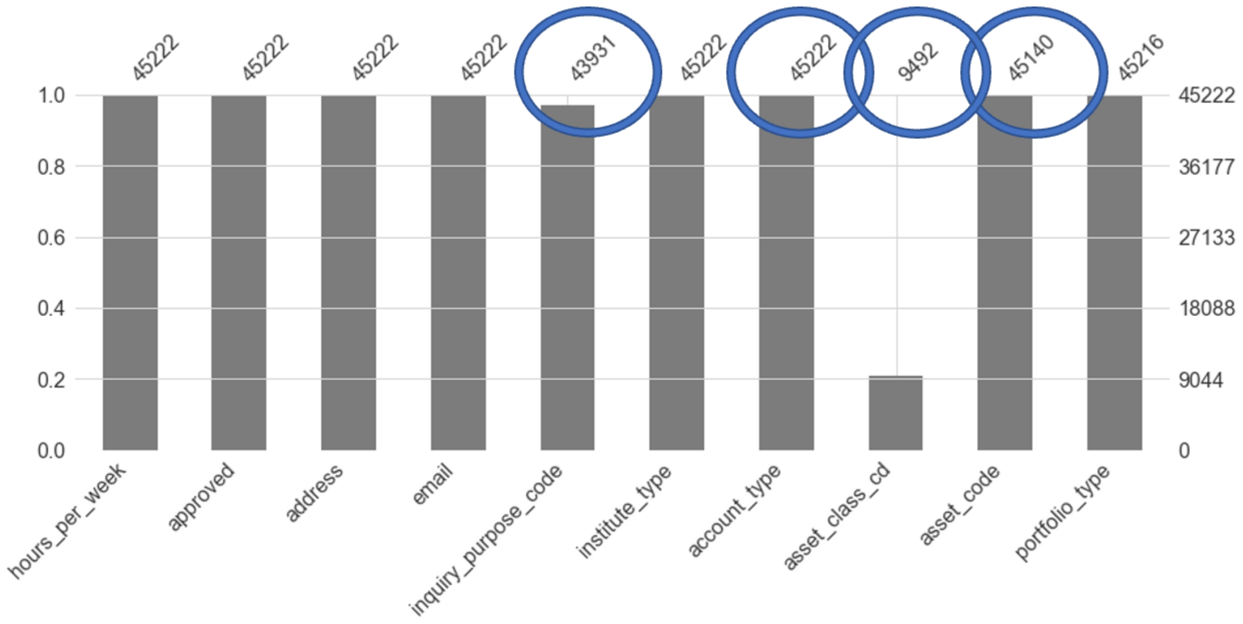


Figure 4: EDA part 1 charts

The above-marked columns are the ones with null values.

2. Removal/Imputation of Null values

a. asset_class_cd: This column has over 80% of the values as null values. The rule of thumb is that if a column has over 60% values as Null values, we can directly remove that column from our analysis.

b. inquiry_purpose_code: This column only has 3% of values as null and hence, we can remove all the rows with inquiry_purpose_code as null without creating any bias in the code even if the data was not missing at random.

c. asset_code and portfolio_type have relatively few null values and hence, these rows can also be removed without creating any bias in the dataset.

The picture below clearly shows that the null values are distributed at random in the dataset.

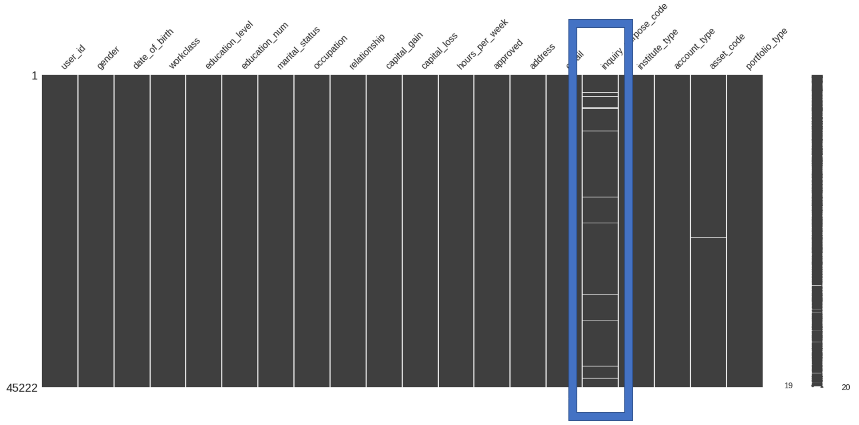


Figure 2: Randomness of Null values

3. Feature Engineering

This is the step in which we modify the different columns to be used in our model. The different steps involved are as follows:

a. Whole addresses are difficult to use in our model so we will extract the zip code from the data and see if they have any effect on the model.

b. Email addresses as a whole are difficult to use in the model, we tried to extract the domain name of the email address if they have any major effect on the predictive power.

c. We converted the date of birth in the required format to calculate a user's current age.

d. We will also remove some of the users who have duplicate data in the dataset. This is done to remove any biases in the data.

e. We will also remove any possible outliers in the dataset.

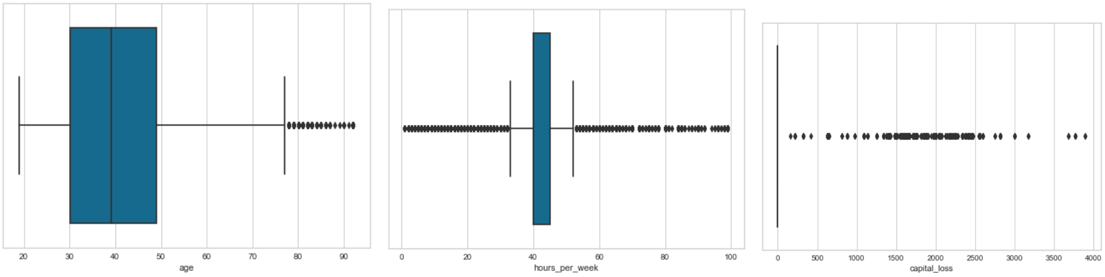


Figure 3: Figuring out the outliers for different columns

4. Secondary EDA

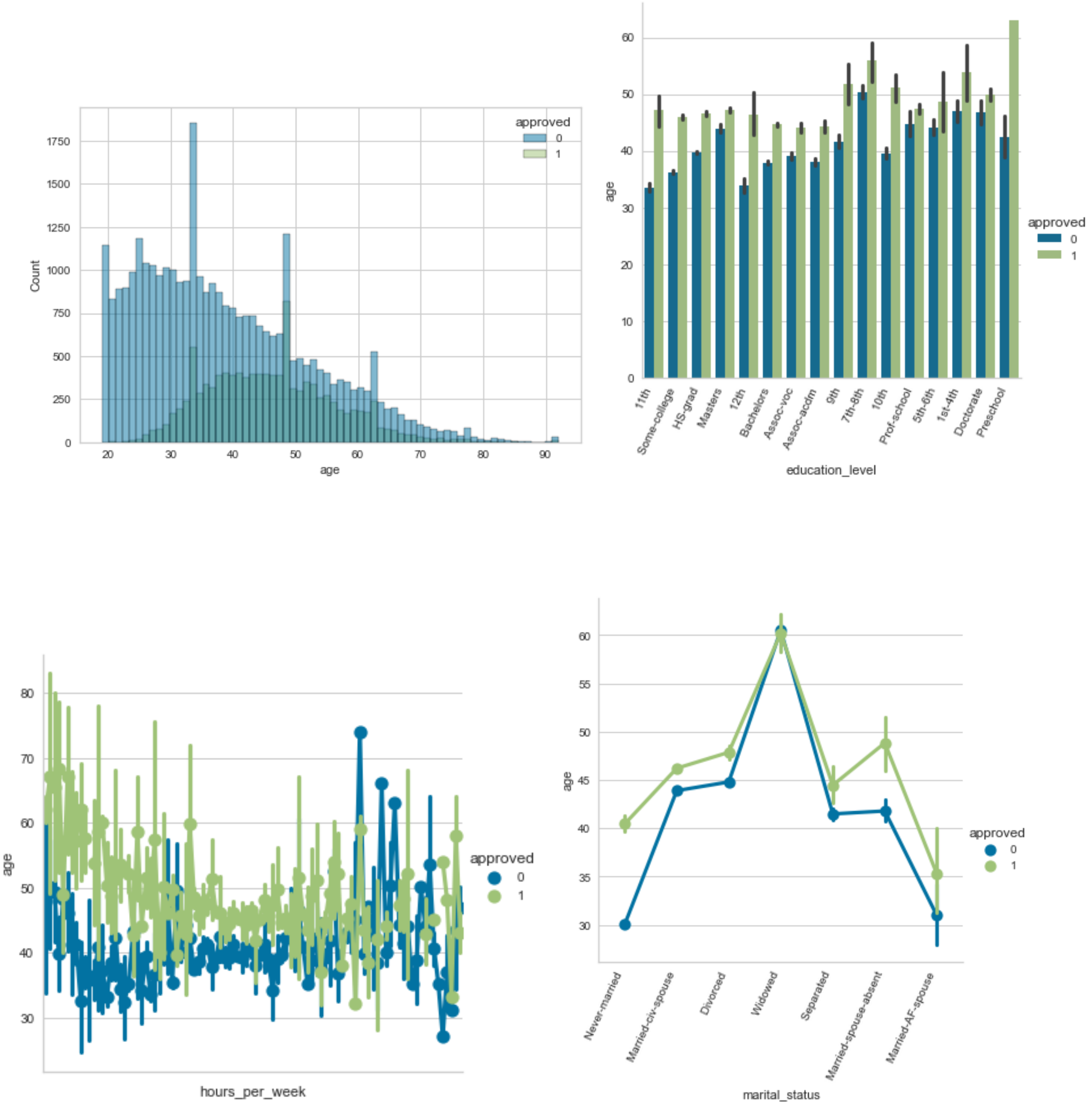


Figure 4: EDA part 1 charts

Observations:

a. Older users are more likely to get approved, this also holds true for each individual education level too.

b. As the number of hours per week for which the users work increases, the effect of age decreases on whether a user is going to default or not.

c. Users who get approved have a higher average age for all marriage levels.

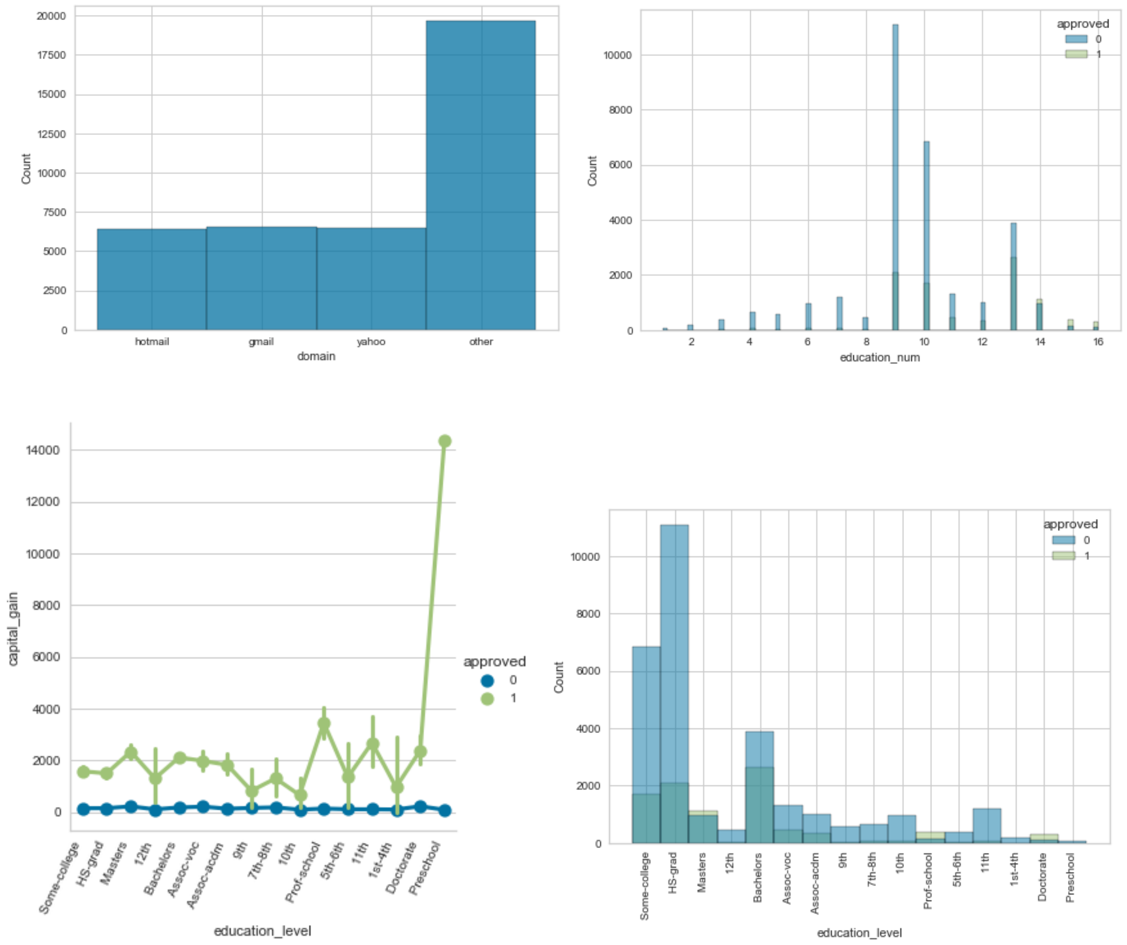


Figure 5: EDA part 2 charts

Observations:

a. Most users sign in from a domain not among the top three big companies. This is also a possible indicator that most users use their work email addresses.

b. The percentage of users who get approved increases with the increase in the education level of users.

c. The average capital gains also increase with education level. Average capital gain is also higher for approved users.

5. SHAP

SHapley vAlue exPressions or SHAP is a concept derived from the game theory approach to calculate the marginal importance of different features. This is an excellent way to figure out the initial set of important features, which can act as a jumping-off point for the step-wise regression or if we wanted a simple way to figure out the feature importance.

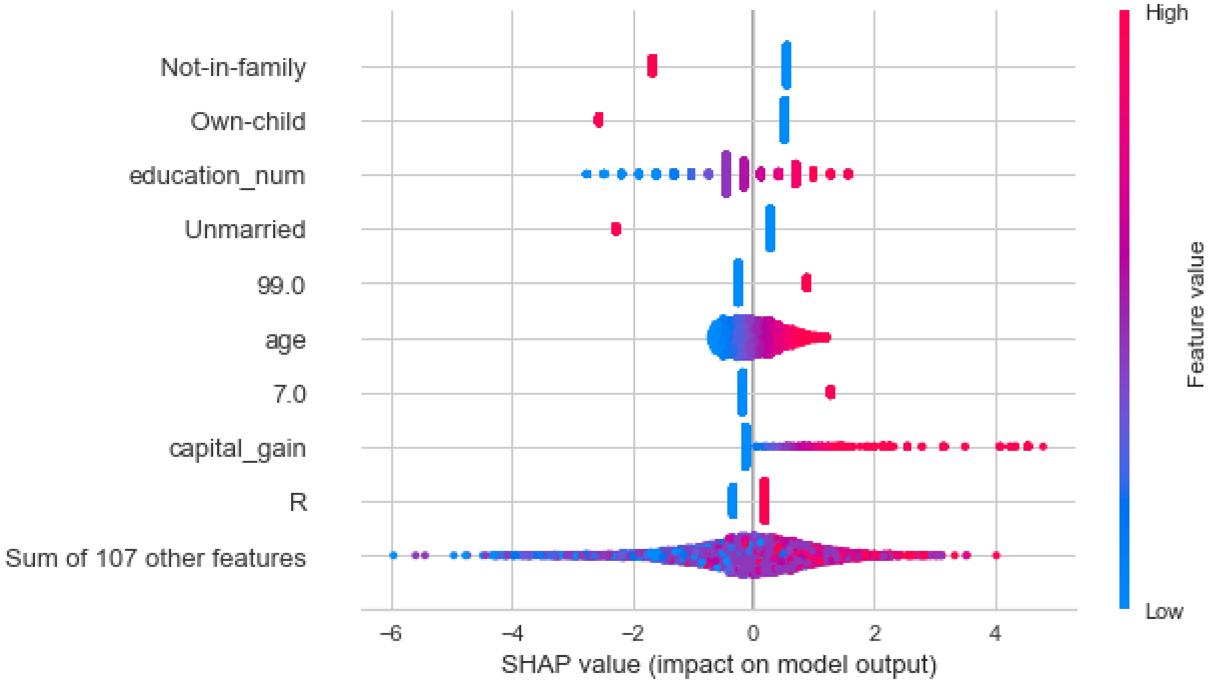


Figure 6: SHAP explaining marginal importance

From the above SHAP chart, we can clearly see that relationships, education_num, marital_status, age, capital gains play the most major role in deciding what features to include. We will go from there and add variables in sort of step-wise regression.

Note: It is always a good practice to standardize the data before applying any machine learning techniques to remove any possible biases that might creep into our prediction, hence, we will perform standardization before this step.

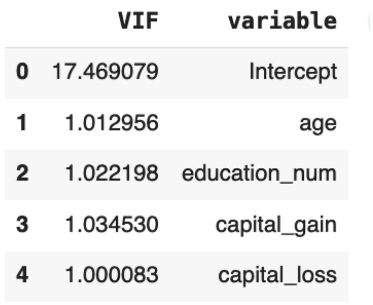
6. Machine Learning Techniques

a. Logistic Regression with step-wise modeling

Final logistic regression model:

```
approved ~ age + education_num + capital_gain +
C(inquiry_purpose_code) + C(relationship) + C(occupation) +
C(institute_type) + C(marital_status)
```

Since we have already looked at the correlation between features, the extent of multicollinearity is very low. This is proved by the figure below:



	VIF	variable
0	17.469079	Intercept
1	1.012956	age
2	1.022198	education_num
3	1.034530	capital_gain
4	1.000083	capital_loss

Figure 7: Variance Inflation Factor

As discussed above, we will be optimizing the true negative rate (TNR), the TNR for this case is 86.5%.

b. XGBoost

Applying XGboost on the same dataset leads to a true negative rate of 92%.

c. Random Forrest

The true negative rate for Random Forrest is 59% which means it's not working efficiently for our dataset, possibly because of the presence of a high number of dummy variables.

Hence, out of the 3 models, we will go with the simplest model as it is still giving us a high accuracy of users who are going to default so based on the business problem where we have to give a simple solution, we will use model 1.

Additional Problem

The result of the logistic regression problem can also be used to further enhance the analysis. We can use this to calculate the initial credit limit for all the users.

Since capital gains are an important part of our model, we can use that to calculate the credit limit. As we can see that we are predicting with high accuracy the TNR rate. We also have a continuous probability distribution function, which we can use to calculate the credit limit that should be given.

credit limit = capital gains * f(probability to be approved) + const.

The above formula states how I propose the credit limit to be calculated.

References:

1. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
2. https://www.bankrate.com/glossary/i/internet-bank/
3. https://en.wikipedia.org/wiki/Shapley_value
4. https://www.kaggle.com/