

---

# Project 4: Credit Card Approval Prediction

Team Data Mining Mavericks Members:

Jiahui Yang  
Saber Garibi  
Rajan Patel  
Ralf Welvers

---

# Agenda

- Project Overview
- Jiahui Yang: Logistic Regression
- Saber Garibi: Random Forest
- Rajan Patel: Support Vector Machines
- Ralf Welvers: K Nearest Neighbor
- Final Assessment

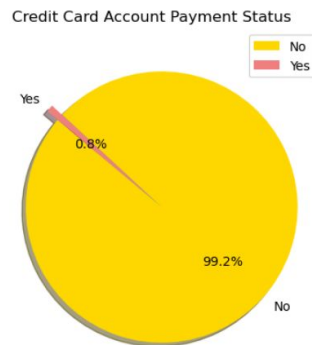
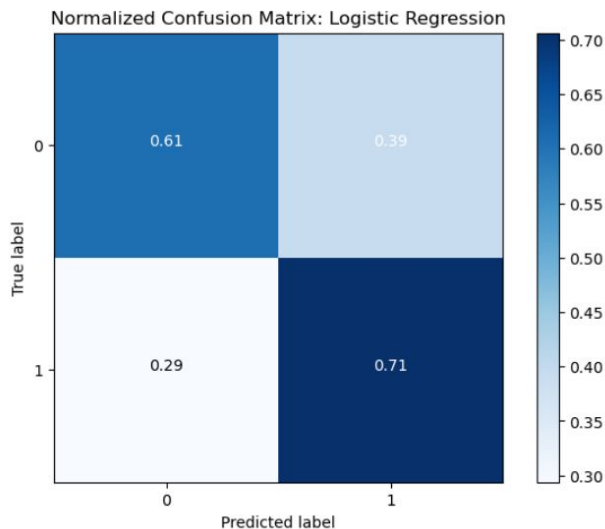
# Project Overview

- Using the following credit card data source from Kaggle:  
<https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction>
- Each team member to create a machine learning model from start to finish including data cleanup / pre-processing
- Our aim is to build a machine learning model to predict if an applicant is a 'good' or 'bad' client.
- Machine Learning models assigned as follows:
  - Jiahui Yang: Logistic Regression
  - Saber Garibi: Random Forest
  - Rajan Patel: Support Vector Machines
  - Ralf Welvers: K nearest neighbor
- All four models are reviewed and collectively team decided which one is most accurate

# Jiahui Yang: Logistic Regression

- It transforms the output of linear regression using a function called the "logistic function," mapping it to values between 0 and 1 for binary classification.

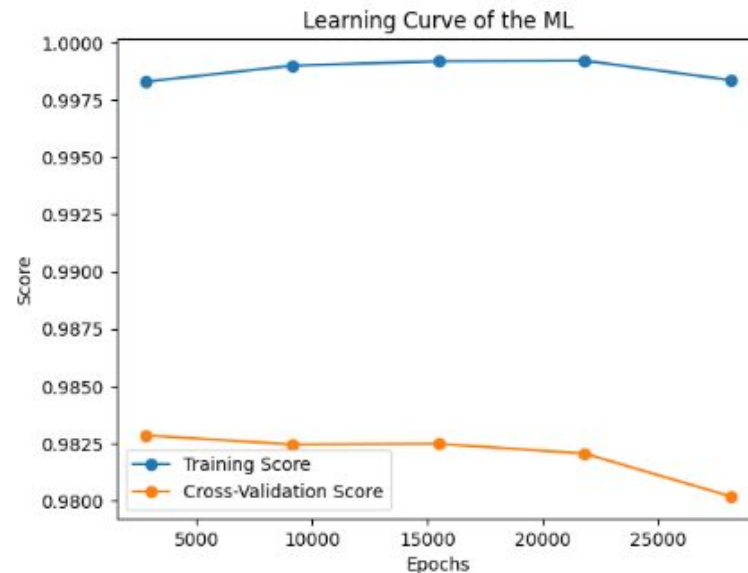
Accuracy Score is 0.65633



# Saber Garibi: Random Forest

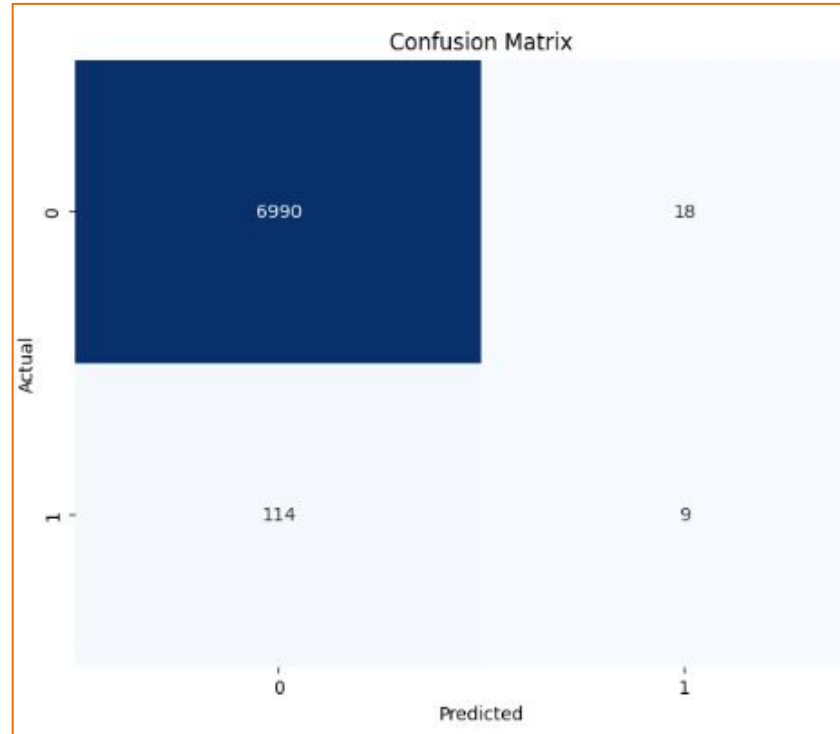
- **Features used:**
  - **Income**
  - **Experience**
  - **Family Status**
  - **Household Size**
  - **Length of Credit**
  - **Education**
- **Target: Missed Payments**
- **Test Size 0.2 - 0.3**

Accuracy: 0.9840976856453216  
Precision: 0.9769818557232648  
F1 Score: 0.9791290582399712



**Reflection:** Overall it was a straightforward and popular model to use. The imputer function was discovered here as RF doesn't like NaNs in the data. The most difficult part was understanding how to prepare data for machine learning.

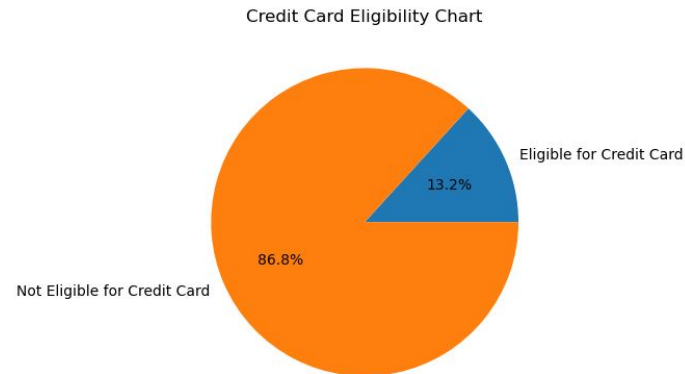
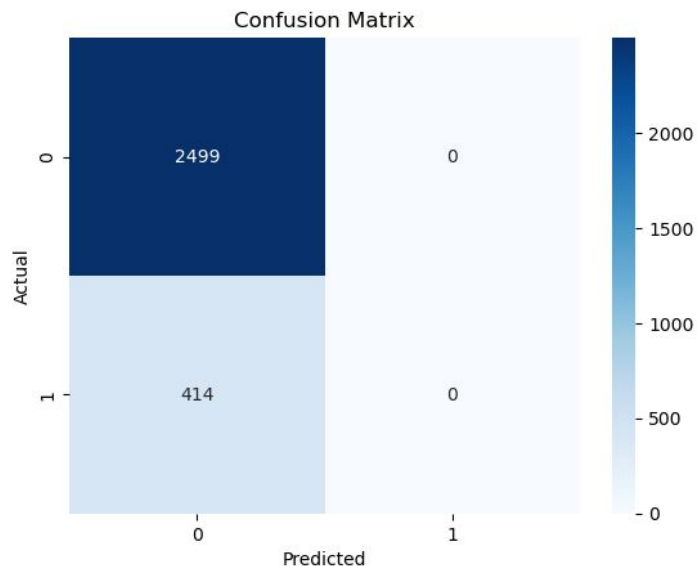
# Saber Garibi: Random Forest Cont'd



Model is good at identifying applicants who are deemed not creditworthy. High number of true negatives.

# Rajan Patel: Support Vector Machines

- Target: Late on payments by at least 30 days = High Risk
- Hyperparameter Tuning
- Best Accuracy: 85.7%



# Ralf Welvers: K Nearest Neighbor

- Tried several values for K. Difference in accuracy was negligible.
- K=3 provided best accuracy at 0.62.

Classification Report:				
	precision	recall	f1-score	support
0	0.59	0.65	0.62	60136
1	0.22	0.11	0.14	1970
2	0.09	0.04	0.05	177
3	0.00	0.00	0.00	44
4	0.04	0.03	0.03	37
5	0.61	0.52	0.56	330
C	0.68	0.68	0.68	68060
X	0.58	0.48	0.53	30547
accuracy			0.62	161301
macro avg	0.35	0.31	0.33	161301
weighted avg	0.62	0.62	0.62	161301



# Final Assessment

- Random Forest: 98%
- Logistic Regression: 65%
- K Nearest: 62%
- SVM: 85%
- The most accurate is SVM

**THANK YOU**