NEW ALLIANCE BANK

# Case Competition Report

"Fraud Activity Identification"

*By*

Team 2

Choukry, Kenza

Eisenman, Dana

Kommareddy, Krithik

Nukala, Sriram

Zhang, Hao

*Prepared for: MSBA 635*

*Prof: Dr. Green*

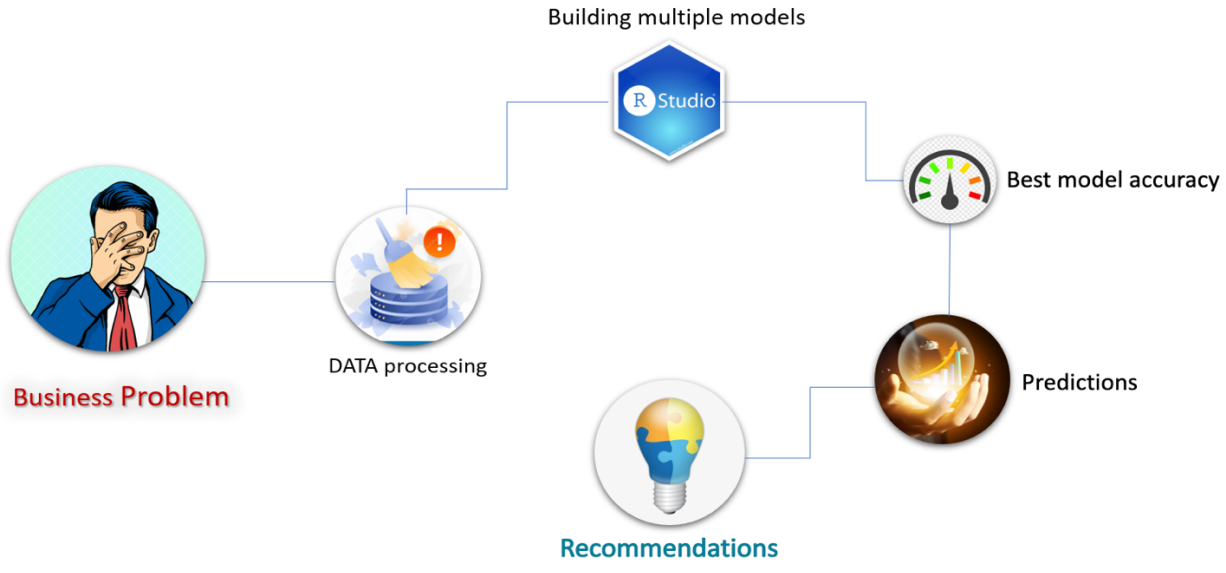
# Case Background/Business Problem:

Alliance Bank is an international bank and financial services company that provides highly integrated banking services. Since the start of the pandemic, consumer's use of online banking has skyrocketed as banks have limited in-person services or visits. As a result, more New Alliance Bank customers (especially in the older and more vulnerable population) are using online banking, which means that more customers are susceptible to financial exploitation (scamming and fraud). The older population of customers are also less likely to report any exploitation or fraud transactions. Currently digital payment data is not analyzed.

The aim of this analysis is to help the New Alliance Bank to detect if a transaction is fraudulent or not and depending on these results, the bank will choose what actions need to be taken. To better understand the study we are working on, it is worthwhile looking at some global statistics. The banking industry has known a multi-billion-dollar fraud problem, which means that banks or financial institutions spend a lot of money on fraud management tools or procedures to minimize fraudulent transactions. As of 2020, there was a total of $20 billion in losses for U.S. banks and financial institutions according to a new report from a  software company FiVerity (ref: Banking journal "Source").

As it might be known, real-time payments sometimes are settled instantly, and if they are not stopped, the money is then lost. The problem we are tackling is supervised and requires a categorical model as we are attempting to predict our character variable (fraud or not fraud) of interest based on known customer and transactional data. To do so, we have come up with our study flow with the different steps taken during the analysis:

**Figure 1: Study Flow**



**Business problem:** correctly defining the business problem is a big step towards a more accurate analysis.

**Data processing:** this starts with data cleaning and better understanding the New Alliance bank data. Also, this includes adding features (variables) to the data to take a closer look at the fraudsters' target.

**Building multiple models**: building multiple models using the R statistical computing software then picked the model that best fits the model

**Prediction accuracy:** from that model we are getting number prediction if this is a fraud or not fraud

And depending on these predictions we came up with our own set of recommendations that will help the Bank detect fraudulent transactions among digital payment and present the bank with the right solutions to help in different aspects (creating a smooth and safe experience to the customers, minimizing lost funds…).

# Recommendations:

All the recommendations should be first implemented on customers in a city then the state and then the whole group of customers. As we start small, we can get to know whether the recommendations are working efficiently and then implement those on a large scale. This methodology will help in identifying the better recommendations to proceed further. The recommendations are:

**Model to Identify customers at higher risk for fraud:**

As a future consideration of the model, the model should be developed to identify customers who are at a higher risk of fraud. These customers should be informed, and the bank should provide tips and tricks to help reduce their vulnerability. Some of the tips and tricks are as follows:

Set strong passwords: A strong password is at least eight characters in length and includes a mix of upper and lowercase letters, numbers, and special characters.

Keep your computers and mobile devices up to date.: Having the latest security software, web browser, and operating system are the best defenses against viruses, malware, and other online threats. Turn on automatic updates so you receive the newest fixes as they become available.

Secure your internet connection: Always protect your home wireless network with a password. When connecting to public Wi-Fi networks, be cautious about what information you are sending over it.

Change password regularly: The passwords should be changed in a span of 60-90 days as this can reduce the potential risk if the password was leaked or known to someone

**Personal fraud protection officer:**

The customers who are identified to be at higher risk for fraud should be assigned an individual fraud protection officer who closely observes their transactions, which will make the customer feel less pressurized.

**Alert on transactions flagged as FRAUD:**

When the model detects any fraudulent activity, an automated text message should be sent to the customer to verify the purchase.

**Multiple FRAUDS:**

If some types of fraudulent activity are being noticed on a set of people, then a fraud protection officer who observes their transactions should be assigned for them.

**Biometrics:** Biometry is the physical or behavioral characteristics that can be used to digitally identify a person, as a security measure. Biometrics work as the identification part of the payment flow to authenticate the shopper to confirm that he is the legit user of the payment method. Biometrics can be physical, like a fingerprint, iris scanner, or facial ID; or behavioral, like navigation / keystroke patterns or physical movements unique to the individual.

### Enhanced Knowledge-based Authentication:

Knowledge-based authentication technology validates identifies against outside sources. Soon it will be easier to quickly determine if an address truly belongs to the caller. The technology incorporates third-party data and tools from providers like LexisNexis, which features access to data on 400 million people from over 10,000 sources. It looks to see whether personal information like phone numbers, addresses, and other data provided by a caller appears in association with the same individual in other records. It can also combine data from multiple sources to create extremely hard-to-guess challenge questions—such as the color of a vehicle the caller owned in 2010.

Finally, Federal Law plays a large part in the actions of fraudsters. The New Alliance Bank should join forces with other financial institutions to lobby US Representatives and convince them to reduce the dollar limit at all levels (ex: make it a Class B Felony for any fraudulent transaction up to $250 instead of $500.)

### Table 1: Punishments by Level of Fraud

| Fraud Amount (USD) | Crime | Punishment in Kentucky |
|---|---|---|
| **Up to $500** | Class A Misdemeanor | Up to 30 days in prison |
| **$501 to $10,000** | Class D Felony | 1 to 5 years |
| **Over $10,000** | Class C Felony | 5 to 10 years |

# Value:

From the given dataset, New Alliance Bank had a loss of $1.6m in 2021 due to fraud transactions. The model developed has an accuracy of 99% to predict fraud transactions. Once predicted, if the Bank can stop the transaction, it can prevent the New Alliance Bank from huge loss. The model developed is automated to predict fraud transactions. So, this lower Fraud operation cost which the New Alliance Bank spends on Data analysts to analyze the data manually and reduces the work of the data analysts.

When businesses do not have to conduct manual reviews, they have more time to focus on strategic business objectives. For example, without the weight of managing manual reviews and chargebacks, fraud analysts can make recommendations on how to manage risks associated with new products and expanded offerings. The model reduces stress among the customers who are worried about fraud transactions as there will high security and in turn increases the trust in the bank.

# Areas for Further Analysis:

There are three areas that could be further explored to improve the model. First, there are many missing data. Adding those data to train our model could help to improve the model's accuracy. We could explore what are the reasons caused by the lost data and recover them if it is lost through switching systems and can use data mapping to gather the lost data to our model. Secondly, adding some new variables would help us to improve the model as well. For example, adding a variable such as payment method would help us to see whether mobile or online transactions have a higher chance of fraud. If one type has a higher chance of fraud, the developers should add more safety features to the web application or the mobile app. And adding a variable such as payment category could help us to see if shopping for gift cards, or shopping on certain websites would relate to a higher chance of fraud. Finally, we could also improve our model by utilizing more of the existing data. This includes a better understanding of the available fields. In this model, we did not consider the internet carrier as a part of building our model. There could be a relationship that switch carrier would relate to a chance of fraud. These possibilities could be further explored.

# Data Exploration:

There are a total of 12601 observations and 24 predictor variables. The predictor variables can be grouped into four main categories:

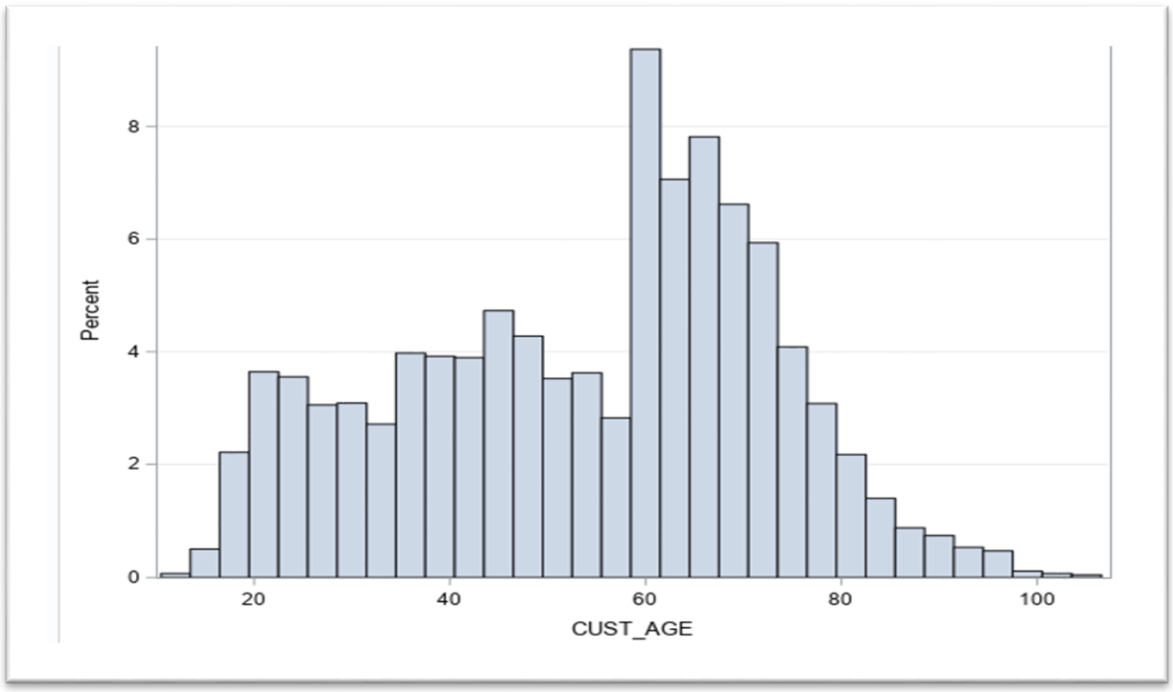
- **Transaction History:** (Transaction ID, Amount of transaction, transaction's region, transaction's state, transaction date, transaction time stamp, carrier's name, activity date of transaction)
- **Customer's profile:** (Customer's age, Customer's zip code, Customer's state, bank's joining date, phone number update timestamp, customer's password update timestamp, available account balance before transaction, customer's device age).
- **Transaction's alert authentication:** (device type alerted, device type used in transaction, bank account authentication result, primary authentication transaction type, secondary authentication results)
- **Internal new Alliance variables:** (Internal new Alliance account information, internal new Alliance identification code, internal new Alliance transaction code, account action code)

Some of the categorical variables have missing values and we also noticed outliers and inconsistencies. Like device age has a negative value. Some of the timestamp's variables are defined as characters which makes R think that it is a categorical variable.

### Table 1: Summary statistics

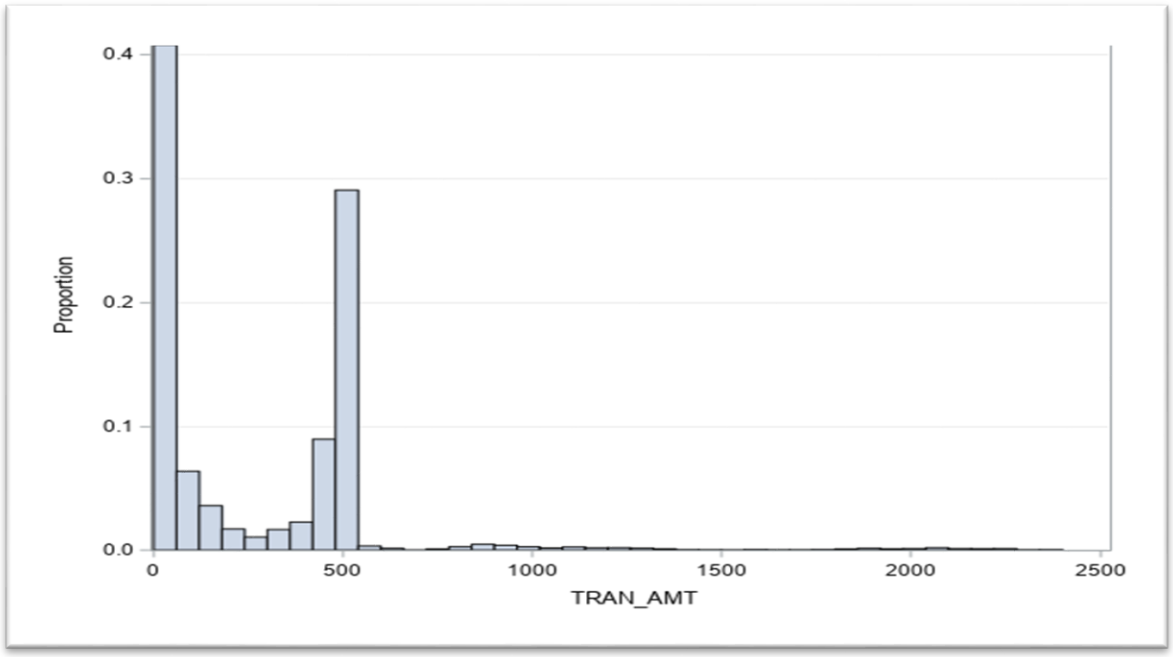
```
                           Values
Name                       Alliance_data
Number of rows             12601
Number of columns          25
_________________________
Column type frequency:
  character                18
  numeric                  7
_________________________
Group variables            None

-- Variable type: character -----------------------------------------------------------------------------------------------------------
# A tibble: 18 x 8
   skim_variable         n_missing complete_rate   min   max empty n_unique whitespace
 * <chr>                     <int>         <dbl> <int> <int> <int>    <int>      <int>
 1 PWD_UPDT_TS                2802         0.778    13    18     0     9588          0
 2 CARR_NAME                  2447         0.806     3    78     0      537          0
 3 RGN_NAME                   2447         0.806     7    19     0       19          0
 4 STATE_PRVNC_TXT            2447         0.806     4    20     0      124          0
 5 ALERT_TRGR_CD                 0         1         4     4     0        2          0
 6 DVC_TYPE_TXT               1576         0.875     5     7     0        4          0
 7 AUTHC_PRIM_TYPE_CD            0         1         6     8     0        5          0
 8 AUTHC_SCNDRY_STAT_TXT        59         0.995     5    17     0        3          0
 9 CUST_STATE                   32         0.997     2     2     0       48          0
10 PH_NUM_UPDT_TS             6354         0.496    13    18     0     5841          0
11 CUST_SINCE_DT                0         1         8    10     0     8024          0
12 TRAN_TS                      0         1         5    15     0    11863          0
13 TRAN_DT                      0         1         5     9     0      344          0
14 ACTN_CD                      0         1         6     6     0        1          0
15 ACTN_INTNL_TXT               0         1        10    10     0        1          0
16 TRAN_TYPE_CD                 0         1         3     3     0        1          0
17 ACTVY_DT                     0         1         5     9     0      344          0
18 FRAUD_NONFRAUD               0         1         5     9     0        2          0

-- Variable type: numeric -------------------------------------------------------------------------------------------------------------
# A tibble: 7 x 11
  skim_variable           n_missing complete_rate     mean       sd      p0     p25    p50    p75    p100 hist
* <chr>                       <int>         <dbl>    <dbl>    <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl> <chr>
1 TRAN_ID                         0             1  6301     3638.      1     3151    6301   9451   12601  ▇▇▇▇▇
2 TRAN_AMT                        0             1   276.      318.      0.01    11.4   162.    489.   2376. ▇▂▁▁▁
3 ACCT_PRE_TRAN_AVAIL_BAL         0             1 10205.    29804.      0        0    2417.   4777. 361519. ▇▁▁▁▁
4 CUST_AGE                        0             1    54.0     18.8    13       39      59      68     105  ▁▃▅▇▃
5 OPEN_ACCT_CT                    0             1     6.67     8.76    0        3       5       7     227  ▇▁▁▁▁
6 WF_dvc_age                      0             1   610.      668.    -117      74     361     962    2783  ▇▂▂▁▁
7 CUST_ZIP                        0             1 73819.    25069.      0    60656   85029  92139   99835  ▁▂▂▂▇
> |
```

### Figure 2: Histogram of Customer's age



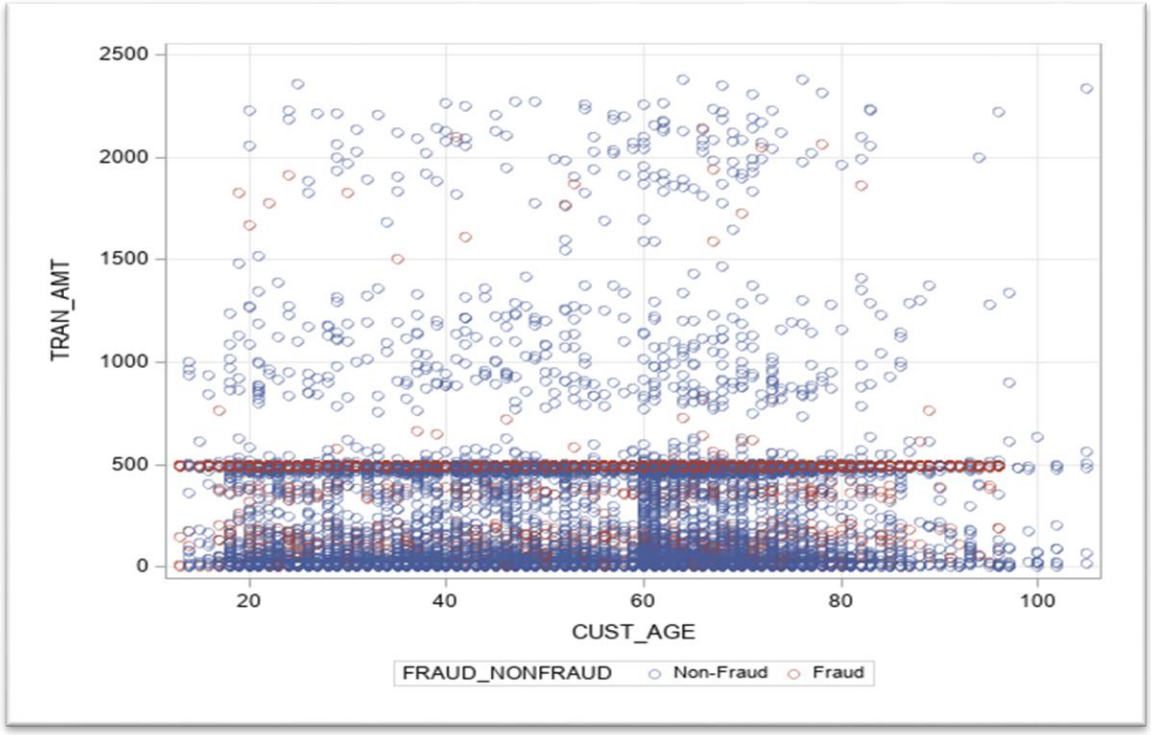

Based on this distribution, about 10% of the customers are 60 years old.

### Figure 3: Histogram of Transaction amounts



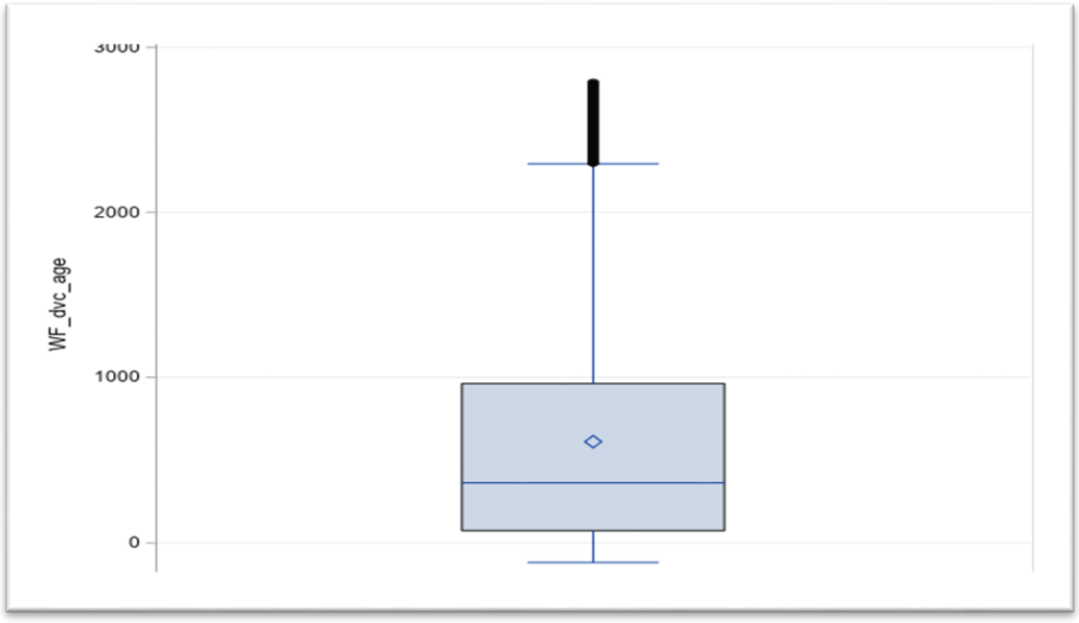

Based on this distribution, 40% of the transaction are below 30$ and 40% are in between $400 to $500.

**Figure 4: Scatter plot, Customer's age vs Transaction amount**




Scatter plot showing customer's age vs transactions amount grouped by fraud or not.

**Figure 5: Box plot for customer device age**




This plot shows that most of the customer's device age associated with their accounts is 600 days and shows the outliers.

# Pre-processing, and Feature Engineering

If we omit rows that have one or more missing values, we will delete ~10,000 observations. To prevent this, we replaced missing values with appropriate entries:

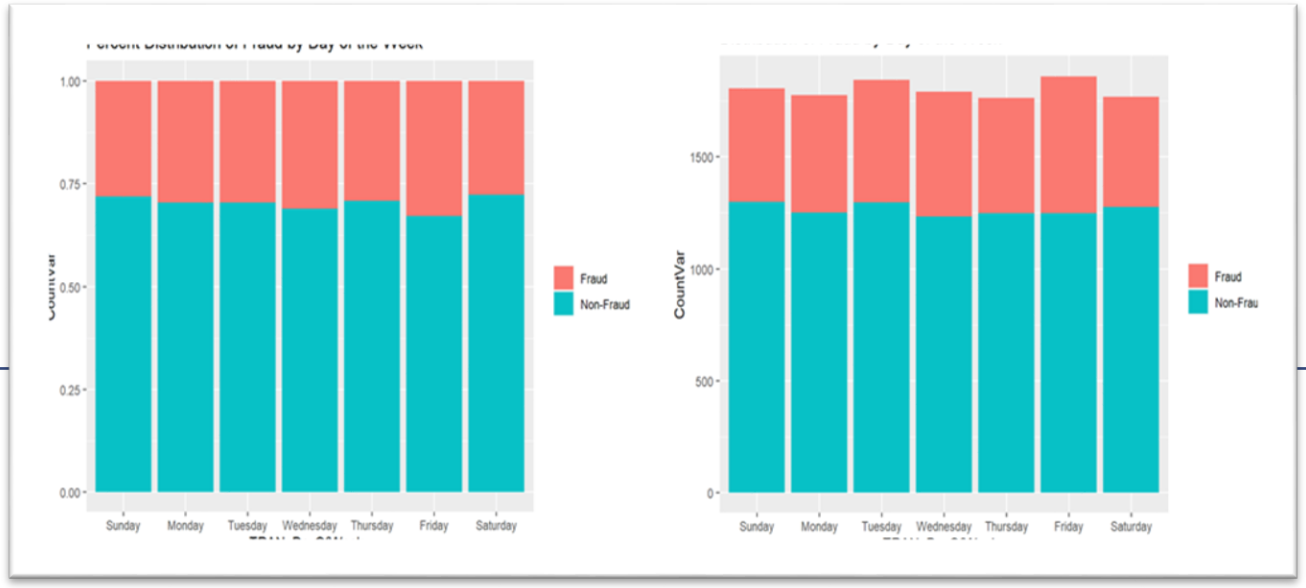
- For date variables, we replaced them with the latest transaction date. The rationale was that perhaps the database was not completely updated when the query was run. We plan to work with the New Alliance Bank to determine if there are other (better) dates to replace them with. For example, if the blanks are due to a change in database (Oracle -> SQL), then we can replace them with the date the change occurred.
- For categorical variables, we replaced the missing values with "Other."

Some variables in the original data were removed for several reasons:

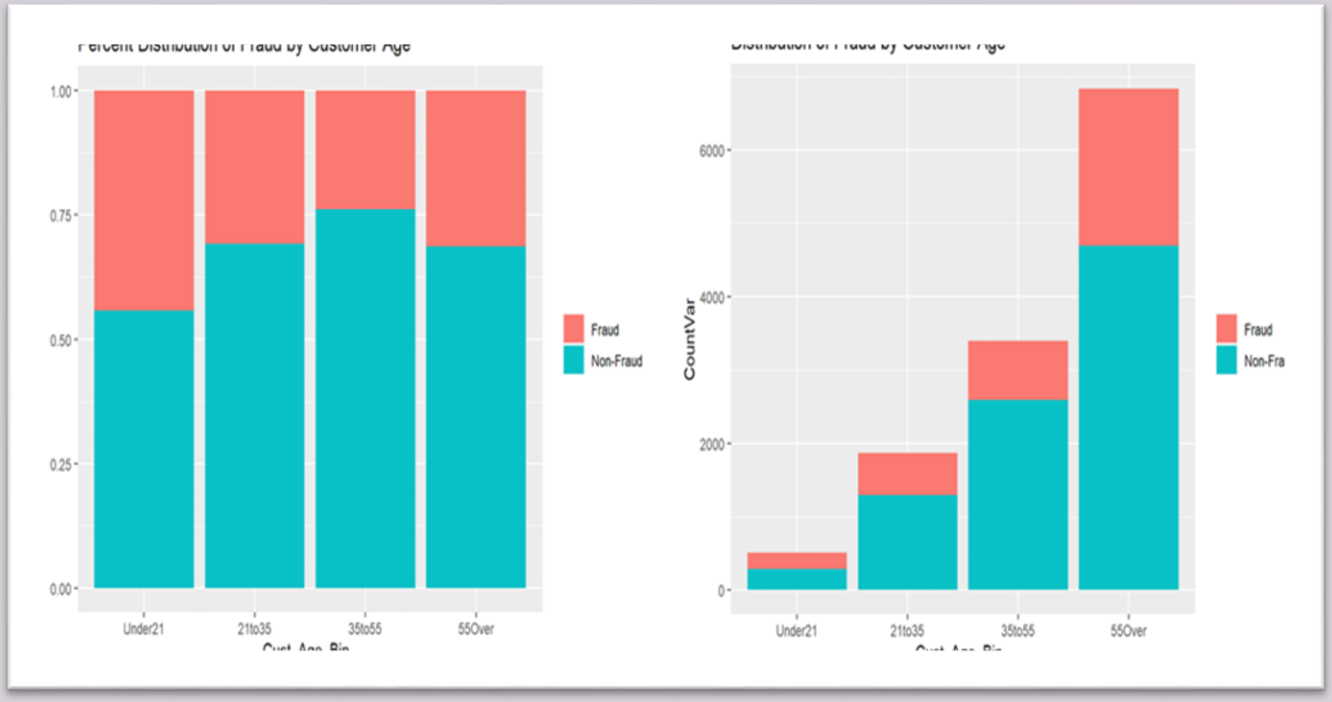
- Predictor variables that contain dates and times can be difficult to work with when training a predictive model. So, many of them were used to calculate another variable and then the original predictor was deleted. For example, CUST_SINCE_DT was used to calculate the number of years someone was a customer of New Alliance Bank. That new data was stored in Cust_Since_Years_ and then CUST_SINCE_DT was deleted.
- The categorical variables 'ACTN_CD', 'ACTN_INTNL_TXT', 'TRAN_TYPE_CD will be removed as they have only one unique value and do not have an impact on prediction.
- TRAN_DT and ACTVY_DT can be dropped as both provide the same data which is available in TRAN_TS.

Below are some examples of the new variables we created, broken down by Fraud vs. Non-Fraud:

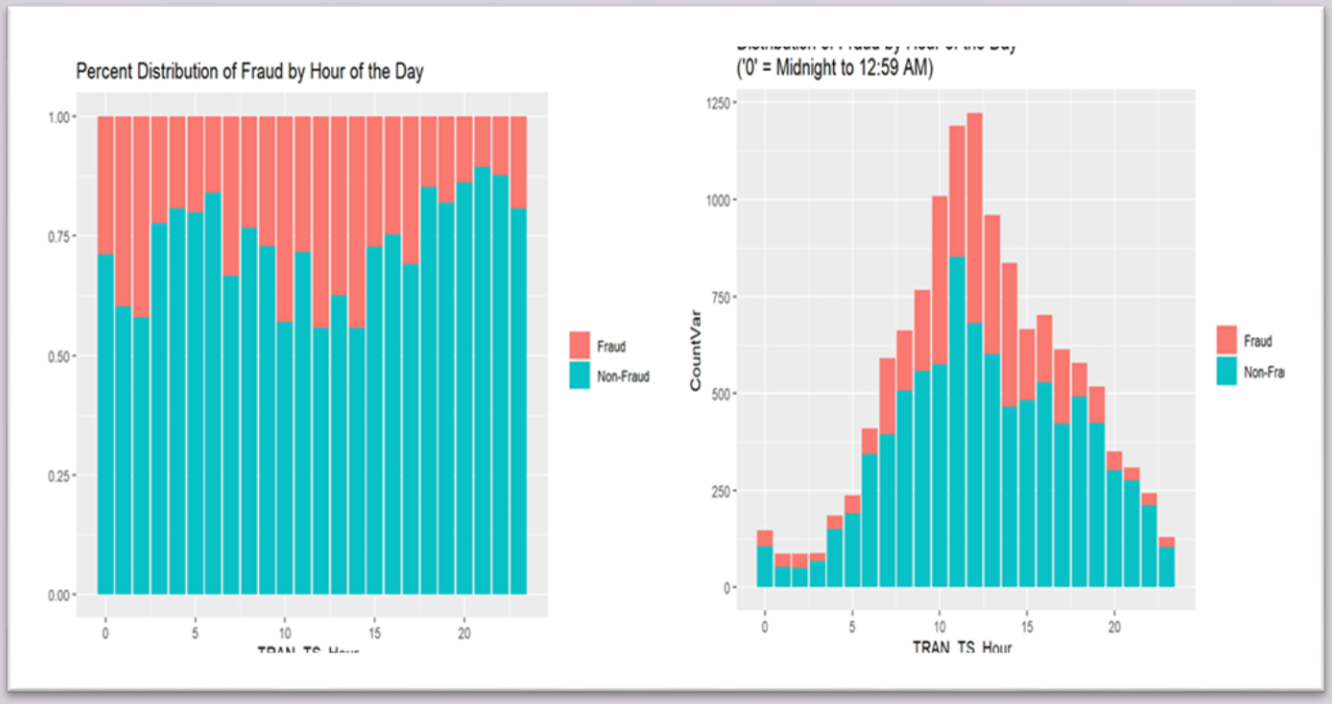
**Figures 6 & 7: Day of the Week**




**Figures 8 & 9: Customer Age**




## Figures 10 & 11: Hour of the Day



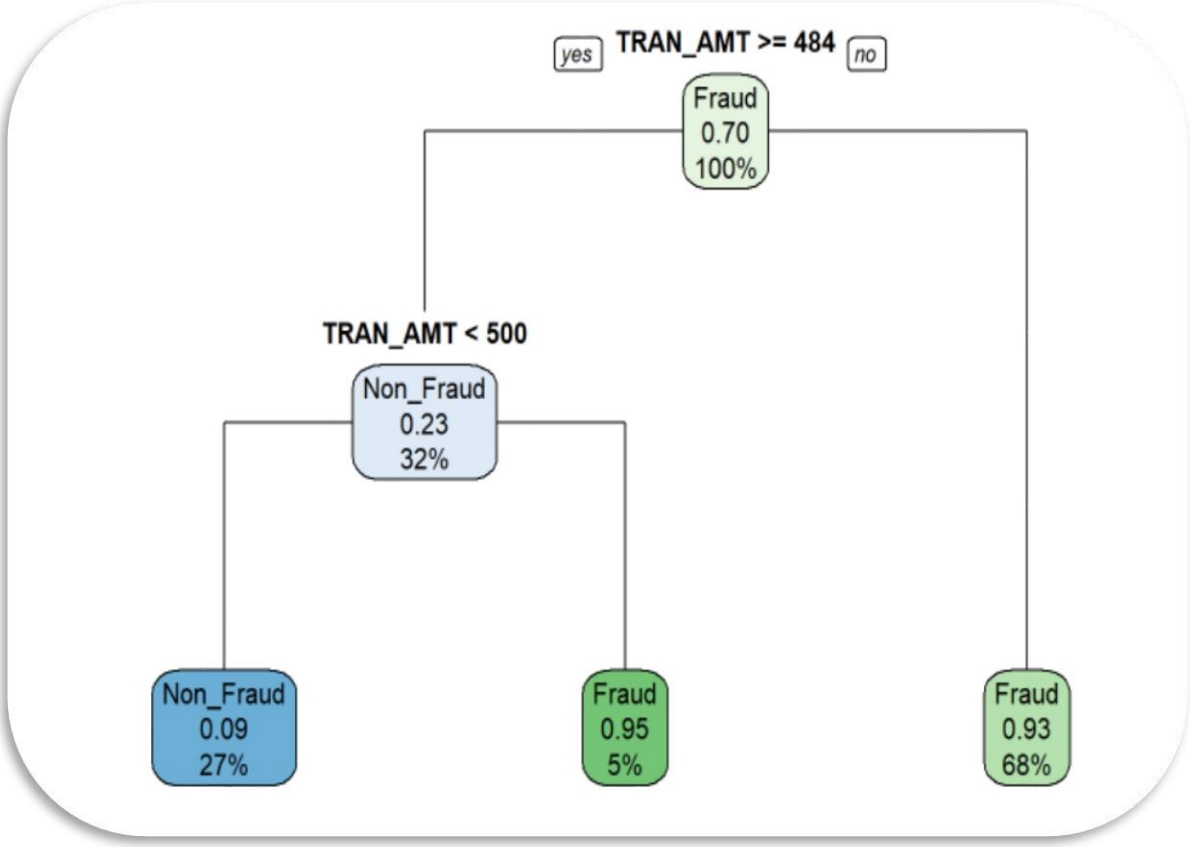
# Model Exploration

As we began to explore potential models, we divided the training data into three pieces:

80% data for Training & Validation and 20% for Testing.

**Classification Model:** Because this is a classification problem, we first started with a simple Classification Model.

**Fig 12: Decision tree for Classification Model**



**Figures 13: Variables by Importance for Classification Model**

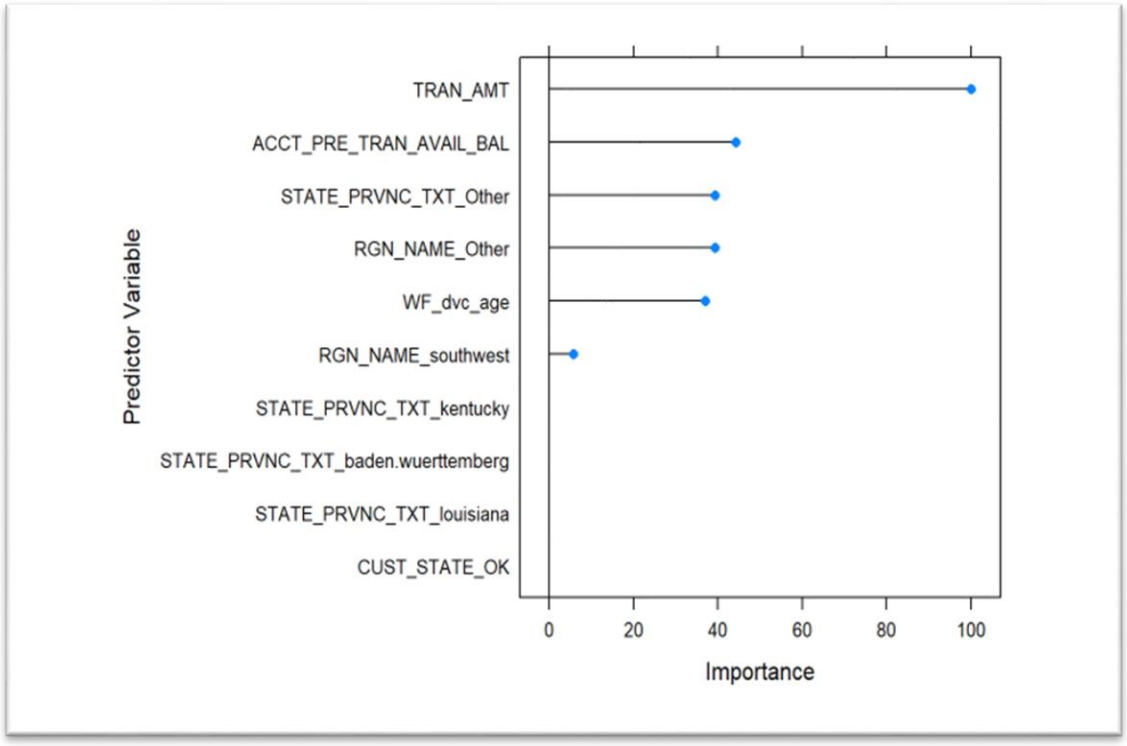
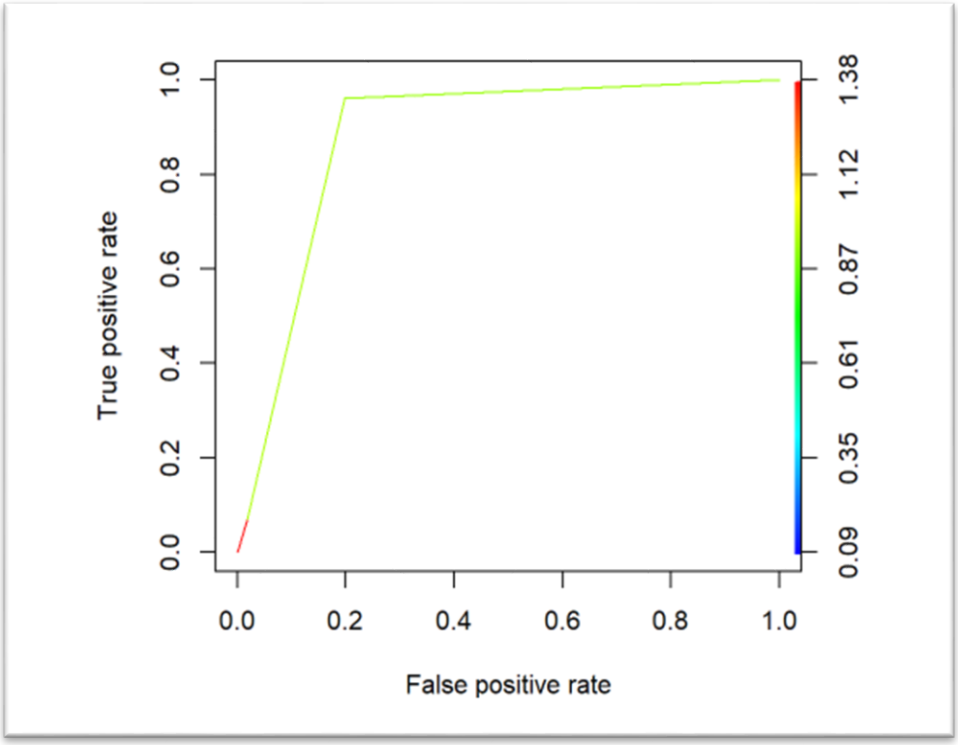


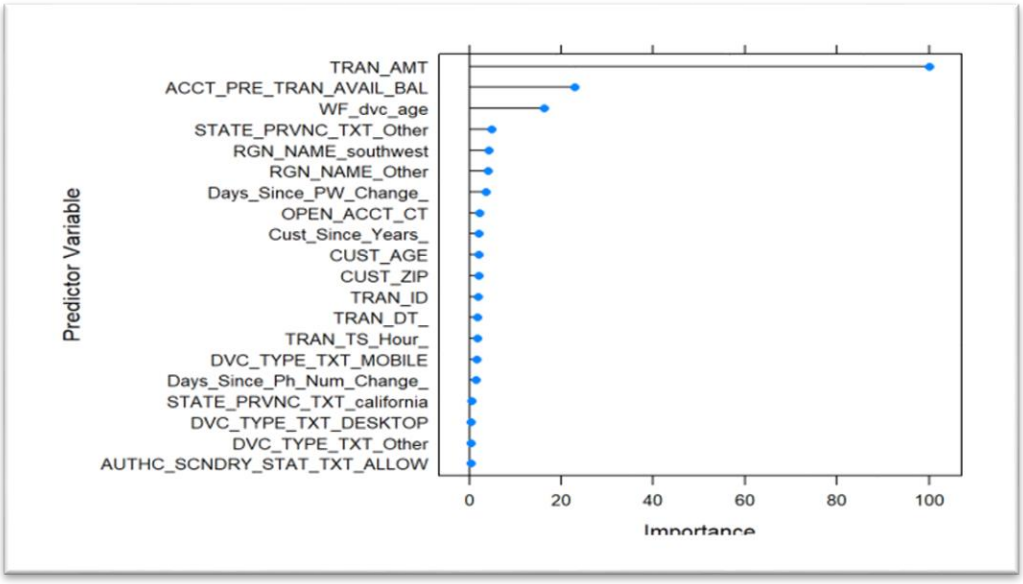

**Fig 14: Variables by Importance to the Classification model**

The Area Under the Curve (AUC) for the Classification model was **0.88015.**



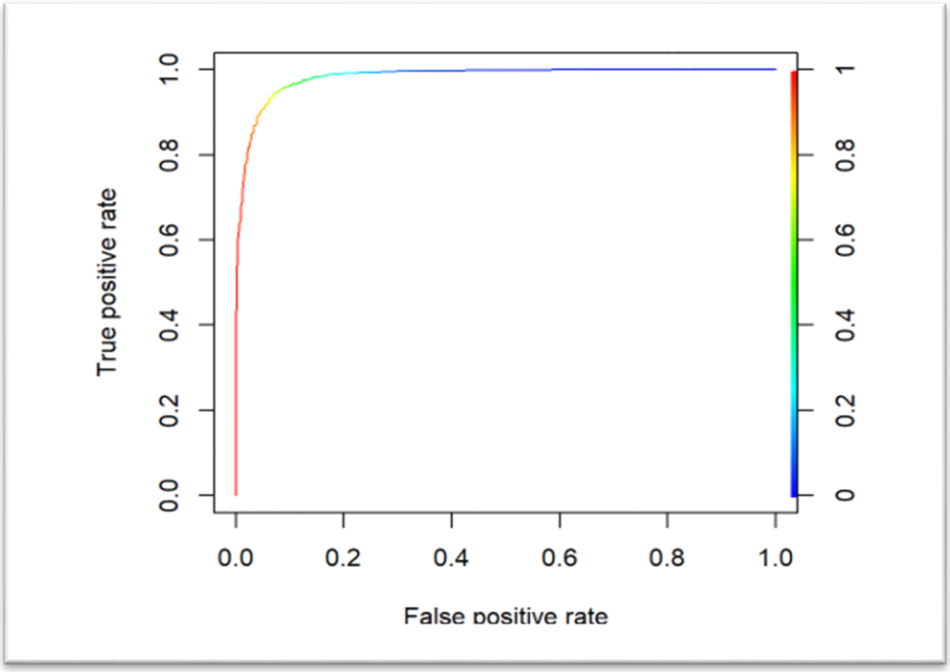
**Random Forest:** Next, we ran a Random Forest model on our data, Random Forest is an ensemble decision tree model that uses bootstrap aggregation (referred to as bagging) to essentially simulate a larger dataset.

### Fig 16: Variables by Importance to the Classification model



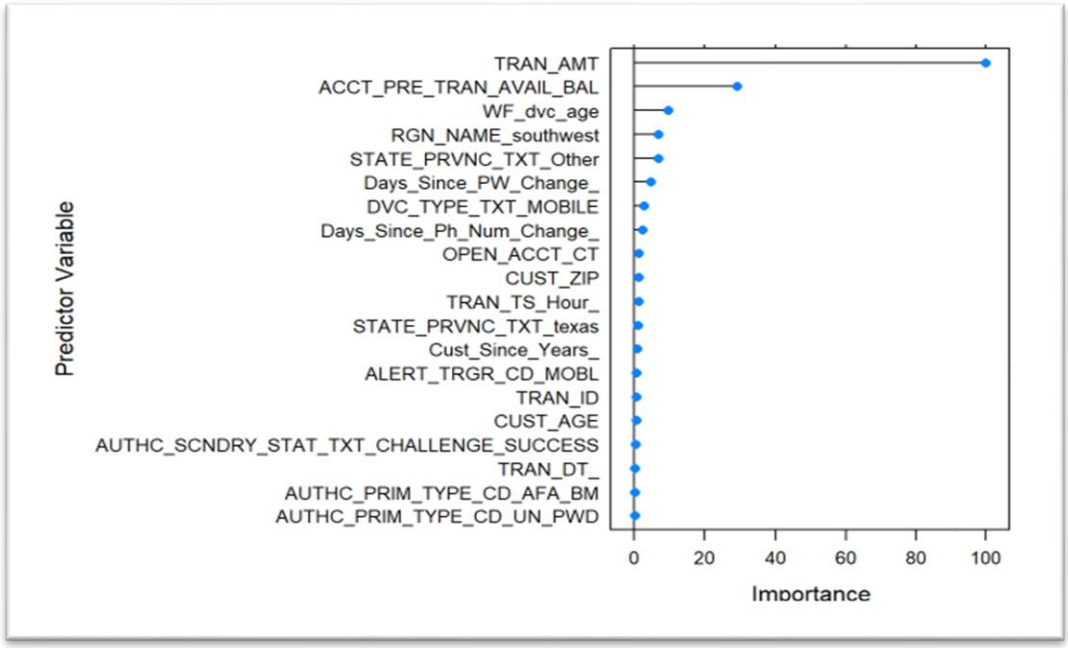

### Fig 17: ROC for Random Forest model

The Area Under the Curve (AUC) for the Random Forest model was **0.9832**



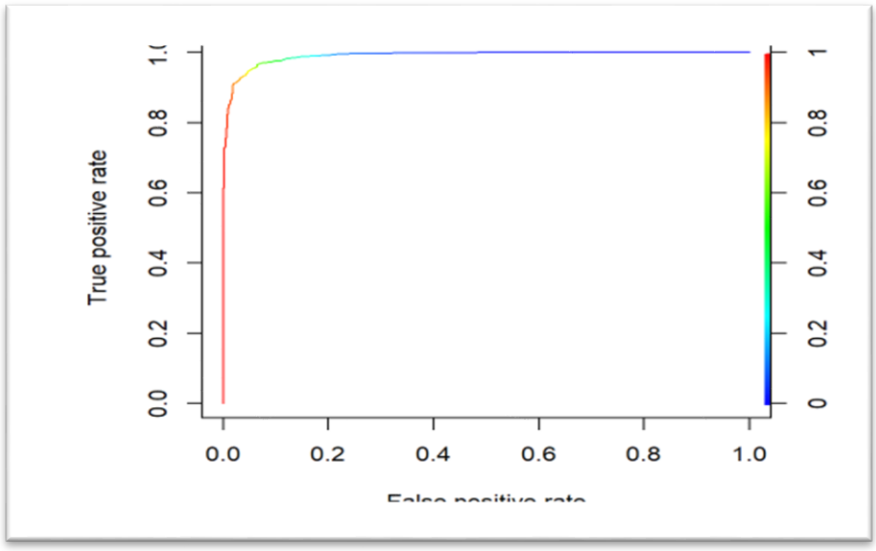
**Gradient Boosting Tree:** Next, we tried another variation of decision trees, a gradient boosting tree. This model sequentially adds predictors to a decision tree ensemble, each one correcting its predecessor. This method attempts to fit the new predictor (or successor) to the residual errors made by the previous predictor/predecessor. In general, gradient boosting trees have shorter training time than random forests as they train much fewer trees.

### Fig 18: Variables by Importance to the Gradient Boost model




### Fig 19: ROC for XGBoost model

The Area Under the Curve (AUC) for the XGBoost model was 0.99028



# Model Selection:

**Table 3: Model Evaluation Comparison**

| Model Type | AUC |
|---|---|
| **Classification** | 0.88015 |
| **Random Forest** | 0.97297 |
| ***XGBoost*** | ***0.9907*** |

Given these three models, all with adequate ROC AUC scores, we decided to use **XGBoost** as our Final model for the following reasons:

- **XGBoost** has the highest AUC of the 3 models we built
- **XGBoost** reduces variances and increases prediction accuracy by learning slowly from errors/residuals of observations
- Though it's difficult to interpret **XGBoost**, it has better prediction accuracy over classification tree.

End Of Report