Identifying At-Risk Customers

Before Approving Credit Applications.

Milestone 2, Milestone 3, and Milestone 4

Michelle Helfman

Bellevue University

DSC630-T302 Predictive Analytics (2241-1)

Professor Andrew Hua

October 28, 2023

Term Project Milestone 2

Data Selection and Project Proposal

Introduction:

As the US economy is emerging from the COVID-19 pandemic, more people are falling behind on their credit card and car loan payments. Rents have increased dramatically to offset the loss of rental income during the pandemic. Student loan payments paused for over three years and will resume in October. Overpriced homes purchased during the bidding wars of 2021 and 2022 have lost considerable value, and with home prices dropping and interest rates at record-high levels, selling these homes will be challenging, and mortgages will begin to default.

The Federal Reserve's (Fed) raising interest rates to control inflation makes purchasing more challenging. Also, banks and other lenders had already begun tightening credit for months, which went into overdrive after the bank failures in the spring of 2023. These factors place additional importance on identifying customers at risk of defaulting before approving credit applications.

Project Data:

This project uses a credit risk dataset from Kaggle to simulate the credit bureau information used on loan applications and other credit approval forms.

Attribute Information:

- person_age – Age
- person_income – Annual Income
- person_home_ownership – Homeownership (Rent, Mortgage, Own, Other)
- person_emp_length – Employment length (in years)

- loan_intent – Loan intent
- loan_grade – Loan grade (A through G)
- loan_amnt – Loan amount
- loan_int_rate – Interest rate for the loan
- loan_status – Loan status (0 is current, 1 is in default)
- loan_percent_income – Percent of income
- cb_person_default_on_file – Historical default (Y, N)
- cb_person_cred_hist_length – Credit history length

Which Models and Why Use Them?:

Decision Tree Classifiers are supervised machine learning models using pre-labeled data to train an algorithm to be used to make a prediction. Decision Tree Classifiers are easy to understand and explain because they follow natural thought processes but are challenging to display since they produce an extremely large graph. In this case, the Decision Tree uses the variable that will best split the dataset, resulting in whether to approve or reject a credit application.

Logistic regression estimates the probability of an event occurring based on a given dataset of independent variables. In this case, the probability of a default on a consumer loan or extension of credit for a given period. Logistic regression predicts the model's accuracy by identifying the relationship between a continuous dependent variable and one or more independent variables.

Evaluating the Results:

The information will be evaluated before and after using the Decision Tree Classifier and Logistic Regression models. Exploratory Data Analysis (EDA) will visually investigate how the main characteristics fit together and go beyond what the models produce. EDA can give insights into the contributing factors in defaulting on credit and loans. After performing the models, the accuracy will be evaluated, and using the Classification Report, the precision, recall, and f1-score will also be reviewed. A ROC Curve will also be employed to confirm the accuracy of the Logistic Regression model. Finally, additional testing will be performed on the Decision Tree Classifier and Logistic regression models to confirm each model's accuracy.

What Do You Hope to Learn?:

This project will give me more insight into how these models determine their results. This model type is used in decisions that significantly affect people's lives. To grant or deny credit can change someone's future. A better understanding will give me a better comfort level when I use these models in the real world.

Assessing Ethical Concerns and Risks:

The models must be carefully trained since obtaining credit is very important to someone's future. If the training data is skewed toward either approval or denial, that would disproportionately affect the outcome to one side or the other. When training the model, an immense amount of data must be used. Every permutation and nuance of this information must be represented repeatedly within the dataset to avoid bias. Gender, race, and zip code do not play a part in any of these models, so that form of bias does not need to be addressed.

The most significant risk is relying solely on the model's results. Examining some of the individual components of the credit report, such as employment length, may affect the final determination. Finally, overall, this data contains personal information beyond what is represented in the sample dataset and must be securely stored in such a way as to avoid theft or misuse.

The Contingency Plan:

If the dataset for identifying credit risk is unsuitable for this project, I must look into an alternate scenario. For this, I will create a dataset combining natural gas indexes from the Henry Hub terminal with weather data from two other locations to explore the effect of high and low temperatures on pricing.

Include Anything Else of Importance:

When using the Decision Tree Classifier with such a large amount of information, it is impossible to depict the actual Decision Tree. The graph is so big that it cannot be rendered on a single sheet of paper. The alternative is to use a confusion matrix. Even though a Decision Tree Classifier and Logistic Regression models produce binary results, the confusion matrix gives four results: True Positive, True Negative, False Positive, and False Negative. The False Positive and False Negative results signify that additional research is required.

Term Project Milestone 3

Preliminary Analysis

Will I be able to answer the questions I want to answer with the data I have?

The Question: Can customer data on age, income, employment, home ownership, credit history, and current and historical loan information be used to predict whether credit should be extended based on prior loan history? Which dataset features are most predictive of defaulting on loans?

Yes, the dataset has 32581 rows, which will be sufficient to answer the question. The credit risk data comprises 12 features (4 categorical and 8 numeric), with Loan_Status being the target already set to 0 and 1.

Dataset Description and NULL Percentages

```
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   32581 non-null  int64
 1   Income                32581 non-null  int64
 2   Home_Ownership        32581 non-null  object
 3   Employment_Length     31686 non-null  float64
 4   Intent                32581 non-null  object
 5   Loan_Grade            32581 non-null  object
 6   Loan_Amount           32581 non-null  int64
 7   Interest_Rate         29465 non-null  float64
 8   Current_Loan_Status   32581 non-null  int64
 9   Loan_Percent_Income   32581 non-null  float64
 10  Prior_Defaults        32581 non-null  object
 11  Credit_History_Length 32581 non-null  int64
dtypes: float64(3), int64(5), object(4)
```

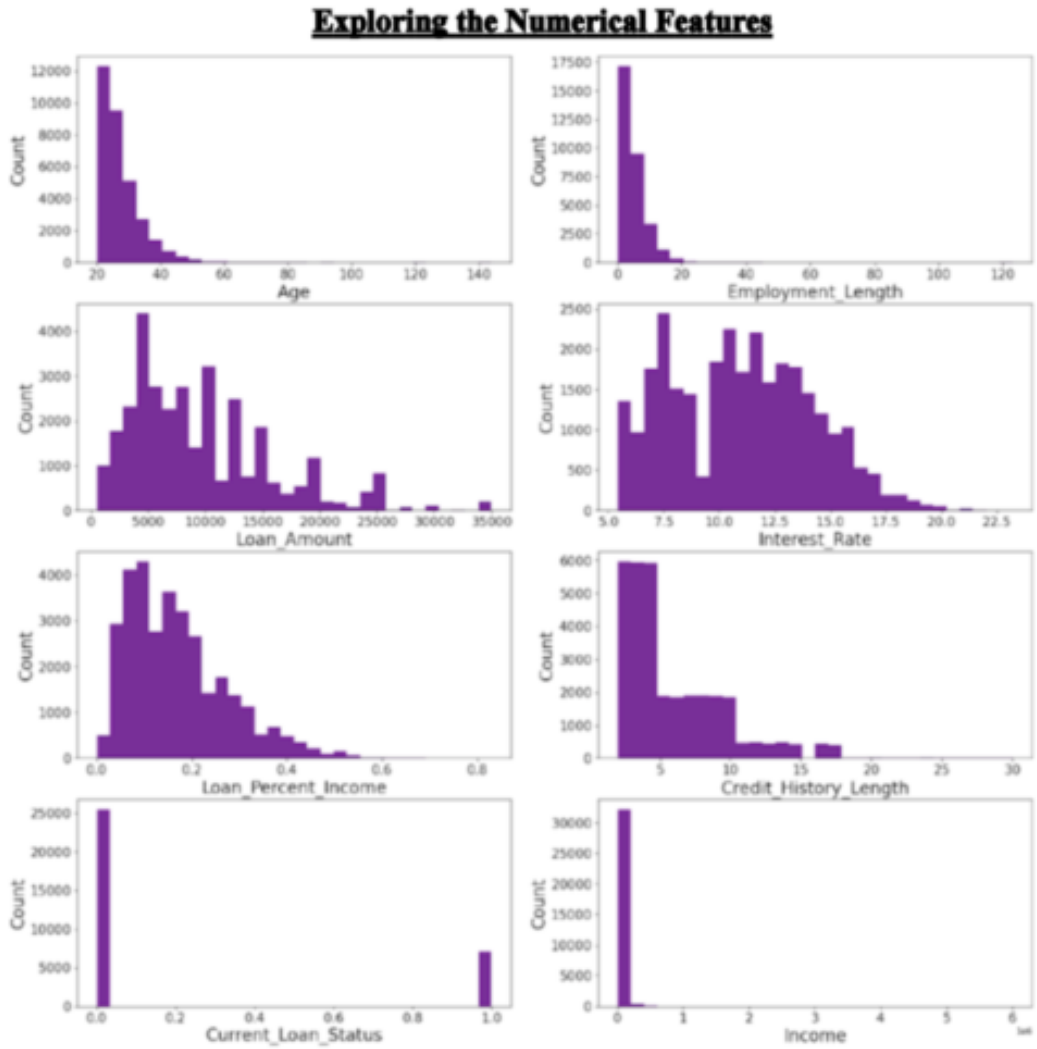
```
Percentage Columns with NULLS
Age                     0.0
Income                  0.0
Home_Ownership          0.0
Employment_Length       2.7
Intent                  0.0
Loan_Grade              0.0
Loan_Amount             0.0
Interest_Rate           9.6
Current_Loan_Status     0.0
Loan_Percent_Income     0.0
Prior_Defaults          0.0
Credit_History_Length   0.0
```

- There are 2 columns with missing information: Employment_Length and Interest_Rate.
- The overall average employment length will replace the NULLS in the Employment_Length column.

- The NULLS in the Interest_Rate column will be filled in with the average interest rate per Loan_Grade.

What visualizations are especially useful for explaining my data?

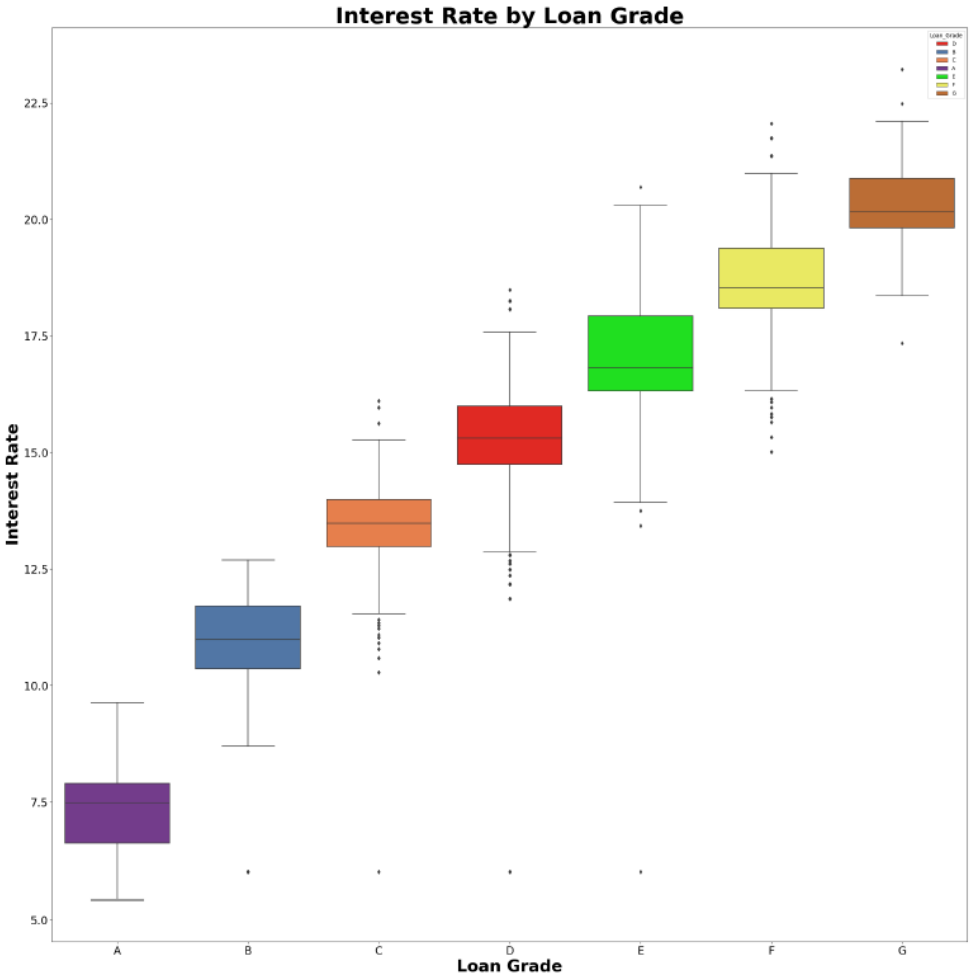
Histograms are used to describe the initial state of the data. This includes the minimum, maximum, averages, and identifying outliers that could be candidates for removal. A boxplot visualizes the correlation between Loan_Grade and Interest_Rate prior to filling in the NULLS.



- **Age -** Most of the loans are between 21 and 35, with an average age of 27.73 years and a maximum of 144. The outliers are over 84 years old, and these rows will be removed.
- **Employment_Length -** The average employment length is 4.8 years, with a minimum of 0 and a maximum of 123 years. The row with 123 will be deleted, and the NULLS will be replaced with the average employment length.
- **Loan_Amount** – Over 75% of the loans are below $12,000; the maximum loan amount is $35,000 with no outliers.
- **Interest_Rate** – With a minimum of 5.4% and a maximum of 23.2%, the average is 11%, but 9.6% of the interest rates are missing. These will be filled in using the average interest rate for each Loan_Grade.
- **Loan_Percent_Income** – No loan is more than 8% of someone's income, and 75% is less than 2.4%, with an average of 1.7%.
- **Credit_History_Length** – The range of credit history is 2 to 30 years, with an average of 5.8 years. Most people have a credit history of 8 years or less because the majority of ages are less than 30 years old.
- **Income -** While 95% of income is less than $300,00, the maximum is $6 million. The top 5% (168 records) will be considered outliers, and the rows will be deleted.
- **Loan_Status -** 0 is current, and 1 is defaulted on this loan. This is the target variable. While loan status is numerical in nature, this is actually a categorical feature.

Exploring the Categorical Features

- **Home_Ownership** – Home ownership consists of 4 categories: Mortgage, Rent, Own, and Other. Less than 10% of people own their homes, most rent or have mortgages.
- **Loan_Grade** – Loan grades are A through G, and most loans are graded A or B. The average interest rate associated with each loan grade will fill in the NULL interest rate records.
- **Prior_Default** – Has this person defaulted on a prior loan: Y for Yes, N for No.
- **Intent** – What is the purpose for this loan: Education, Medical, Venture, Personal, Debt Consolidation, and Home Improvement in the number of occurrences reverse order.



This picture depicts that the higher the loan grade, the lower the interest rate. Using the average interest rate per loan grade would be an excellent way to fill in the missing Interest_Rate information.

After preparing the data, A Pearson Correlation Matrix will be used to display the relationship between the target feature, Current_Loan_Status, and the other numerical features: Age, Employment_Length, Loan_Amount, Interest_Rate, Loan_Percent_Income, Credit_History_Length, and Income. A Spearman Correlation Matrix will depict the relationships between all the variables to each other, with an added focus on the Current_Loan_Status variable. Additional bar charts will compare the most correlated features

from the Spearman Correlation Matrix to the Current_Loan_Status. Finally, Confusion Matrices will display each model's performance After training and testing the Decision Tree Classifier and Logistic regression models. A ROC Curve will also show if the Logistic regression model will be a good predictor of loan default.

Do I need to adjust the data and/or driving questions?

The data is straightforward in explaining the driving questions about predicting future loan defaults during the credit risk assessment process. Fill in the missing data (NULLS) and removing outliers are the only changes to the original dataset needed. Creating additional dummy variables to transform the categorical features for use in the correlation matrices, the Decision Tree Classifier, and Logistic regression models.

Do I need to adjust my model/evaluation choices?

No adjustments will be necessary since the Decision Tree Classifier and Logistic regression models work well with binary target variables in predicting positive and negative outcomes. Using the Classification Report to evaluate the precision, recall, and f1-score, we can confirm the accuracy of the Decision Tree Classifier and Logistic regression models.

Are my original expectations still reasonable?

Yes, my expectations are still reasonable that, provided with a predefined set of information, a future loan default can be predicted, and which features contribute the most to this prediction when evaluating overall credit risk.

Term Project Milestone 4

Finalizing Your Results

Explain your process for prepping the data:

In Milestone 3, Exploratory Data Analysis (EDA) was performed to get a better understanding of the Credit Risk dataset. For each numeric column, the minimum, maximum, mean, and percentile statistics were given. Details of the categorical, non-numeric columns, consisting of the various elements, were described. Columns with outliers and missing information were identified, and plans were created to correct this information.

```
# Import Functions
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.figure_factory as ff
import plotly.express as px

from sklearn import tree
from sklearn.dummy import DummyClassifier
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.metrics import plot_confusion_matrix
from sklearn.linear_model import LogisticRegression
from yellowbrick.classifier import ROCAUC

import warnings
warnings.filterwarnings('ignore')
```

```
# Read dataset and create data frames for all records

risk_df = pd.read_csv('credit_risk_dataset.csv')
```

```
risk_df.describe().apply(lambda s: s.apply('{0:.5f}'.format))
```

	Age	Income	Employment Length	Loan Amount	Interest Rate	Current Loan Status	Loan Percent Income	Credit History Length
count	32581.00000	32581.00000	31686.00000	32581.00000	29465.00000	32581.00000	32581.00000	32581.00000
mean	27.73460	66074.84847	4.78969	9589.37111	11.01169	0.21816	0.17020	5.80421
std	6.34808	61983.11917	4.14263	6322.08665	3.24046	0.41301	0.10678	4.05500
min	20.00000	4000.00000	0.00000	500.00000	5.42000	0.00000	0.00000	2.00000
25%	23.00000	38500.00000	2.00000	5000.00000	7.90000	0.00000	0.09000	3.00000
50%	26.00000	55000.00000	4.00000	8000.00000	10.99000	0.00000	0.15000	4.00000
75%	30.00000	79200.00000	7.00000	12200.00000	13.47000	0.00000	0.23000	8.00000
max	144.00000	6000000.00000	123.00000	35000.00000	23.22000	1.00000	0.83000	30.00000

Data Preparation Process

- Fill in missing information and remove rows with outliers.

- Create dummy variables to replace the non-numeric categorical columns
- Identify correlated columns and remove them, if necessary
- Split the data into training and testing datasets

There are two columns with information that needs to be added: Employment Length and Interest Rate. The NULLs/NaNs in the Interest Rate column will be filled in with the average interest rate per Loan Grade. The overall average employment length will replace the NULLS in the Employment Length column. The outliers are records with Ages over 84 years old, Employment Length of 123 years, and records with income over $300,000.

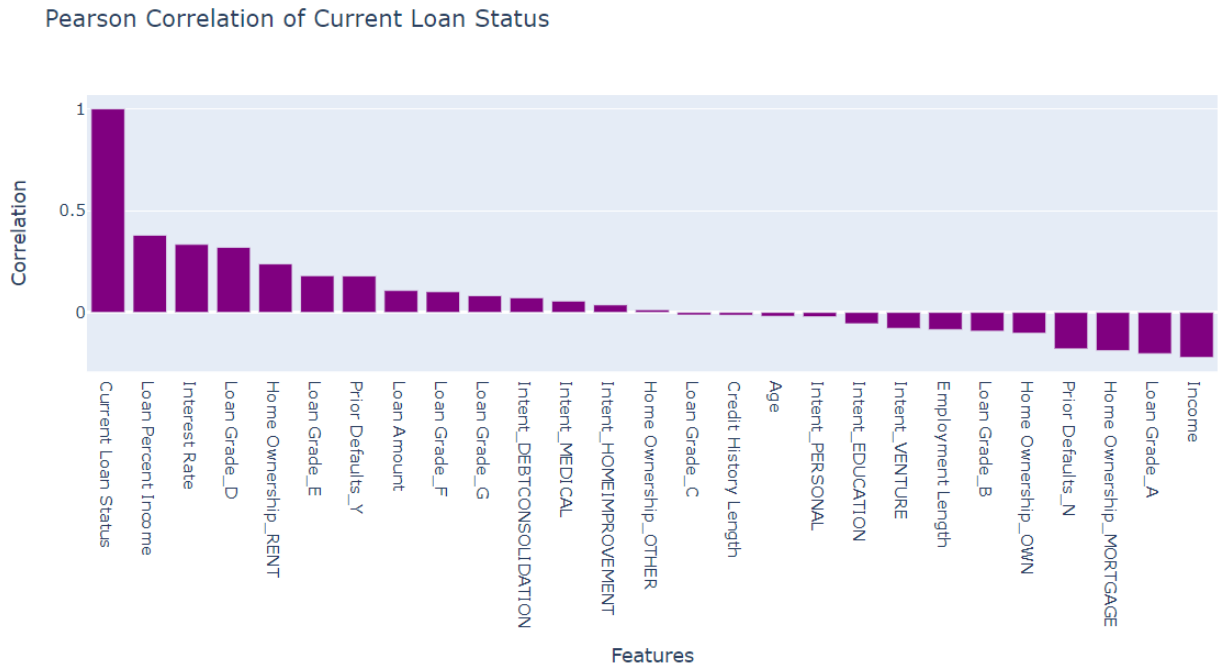
Data Cleanup (Fill In Missing Data and Remove Rows With Outliers)

```
# Fill in the interest rate with average by loan grade

risk_df['Interest Rate'] = risk_df.groupby('Loan Grade')['Interest Rate'].transform(lambda x: x.fillna(x.mean()))

risk_df['Employment Length'] = risk_df['Employment Length'].fillna(risk_df['Employment Length'].mean())

# Remove Outliers 1 Outlier at a time
# Remove rows with age > 84
risk_df = risk_df[(risk_df.Age <= 84)]

# Remove row with Employment Length = 123 years
risk_df = risk_df[(risk_df['Employment Length'] != 123)]

# Remove rows with income >= 300000
risk_df = risk_df[(risk_df.Income < 300000)]
```

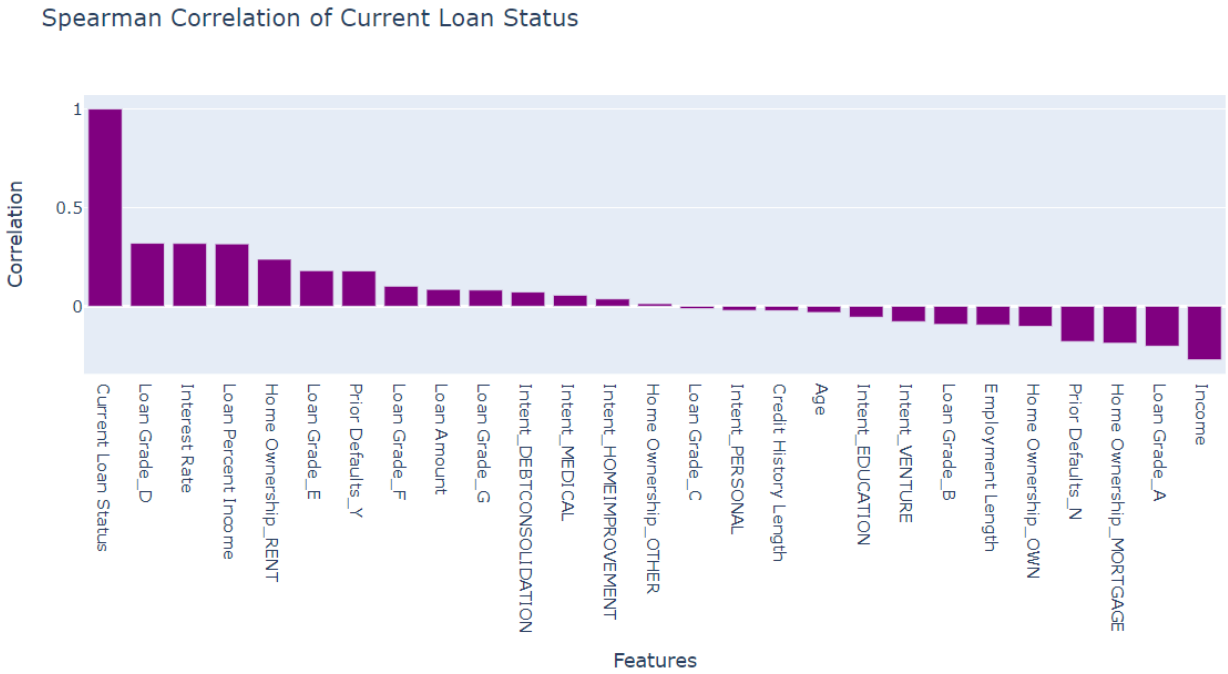
Using Pearson and Spearman Correlation methods, the only heavily correlated columns are Age to Credit History Length (89%). This makes sense since older people have longer credit histories. The Credit History Length column will remain the same.

```
# Get Categorical Columns and Create Dummy Columns

# Get the categorical columns
object_cols = risk_df1.select_dtypes("object").columns
object_cols = list(set(object_cols))

# Create dummy records
risk_dummy_df = pd.get_dummies(risk_df1, columns = object_cols)
```

```
# Using the Pearson method, Correlate all columns including
# dummy variables to the target variable (Current Loan Status)

p_corr_cls = risk_dummy_df.corr()['Current Loan Status'].sort_values(ascending=False)

p_corr_bp = px.bar(x = p_corr_cls.keys(), y = p_corr_cls.values,
                   color_discrete_sequence = ["purple"],
                   title = 'Pearson Correlation of Current Loan Status',
                   labels = {'x':'Features', 'y': 'Correlation'})

p_corr_bp.show()
```




```
# Using the Spearman method, Correlate all columns including
# dummy variables to the target variable (Current Loan Status)

sp_corr_cls = risk_dummy_df.corr(method = 'spearman')['Current Loan Status'].sort_values(ascending=False)

sp_corr_bp = px.bar(x = sp_corr_cls.keys(), y = sp_corr_cls.values,
                color_discrete_sequence = ["purple"],
                title = 'Spearman Correlation of Current Loan Status',
                labels = {'x':'Features', 'y': 'Correlation'})

sp_corr_bp.show()
```



Continuing with the Pearson and Spearmen Correlation methods in comparing all the columns, including the dummy variables, to the Current Loan Status (Target Variable), none are highly correlated. Again, no columns need removal.

Finally, the data, including the dummy variables, is split into training and testing datasets.

```
# Split credit risk data into training and testing data sets

# Delete the Status columns
no_status_df = risk_dummy_df.drop(['Current Loan Status'], axis = 1)

# Create Status Target column
status_df = risk_dummy_df['Current Loan Status']

# Split data into 80/20 sets
x_train, x_test, y_train, y_test = train_test_split(no_status_df, status_df,
                                                    test_size = 0.2, random_state = 1)

print('Number of x_train Rows and Columns = ', x_train.shape)
print('Number of y_train Rows and Columns = ', y_train.shape)
print('Number of x_test Rows and Columns = ', x_test.shape)
print('Number of y_test Rows and Columns = ', y_test.shape)
```

```
Number of x_train Rows and Columns =  (25924, 26)
Number of y_train Rows and Columns =  (25924,)
Number of x_test Rows and Columns =  (6482, 26)
Number of y_test Rows and Columns =  (6482,)
```

Build and evaluate at least one model:

Decision Tree Classifier

```
# Decision Tree Classifier

# Create the Decision Tree Classifier,
# train the model, and test the results

# Create a Decision Tree object
dtc = DecisionTreeClassifier()

# Train the Decision Tree model
dtc.fit(x_train, y_train)

# Train model to make predictions
y_pred = dtc.predict(x_test)

# Calculate accuracy
ac_score = accuracy_score(y_test, y_pred)
print('The Accuracy Score For Decision Tree Classifier = ', round(100 * ac_score, 2), '%', sep = '')
```

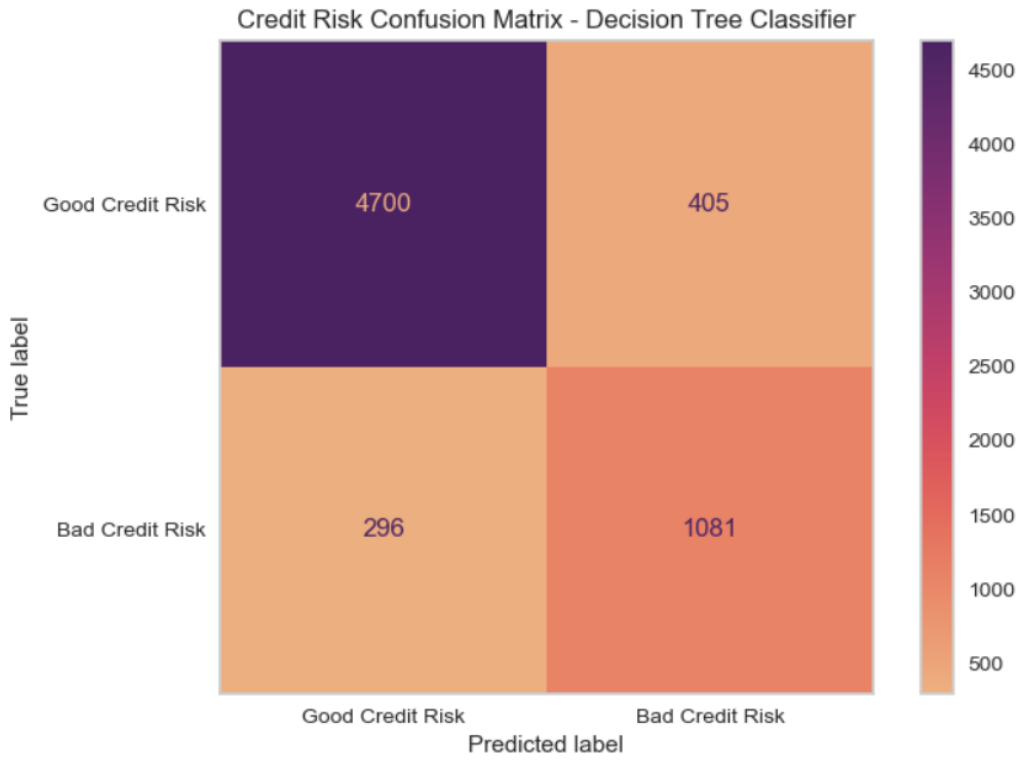
```
The Accuracy Score For Decision Tree Classifier = 89.19%
```

```
# Double check the accuracy

target_names = ['Good Credit Risk', 'Bad Credit Risk']
class_rpt = classification_report(y_test, y_pred, target_names=target_names)

print('Classification Report For Decision Tree Classifier')
print(class_rpt)
```

```
Classification Report For Decision Tree Classifier
                  precision    recall  f1-score   support

Good Credit Risk       0.94      0.92      0.93      5105
 Bad Credit Risk       0.73      0.79      0.76      1377

        accuracy                           0.89      6482
       macro avg       0.83      0.85      0.84      6482
    weighted avg       0.90      0.89      0.89      6482
```

```
# create and plot a confusion matrix

# Set up labels
labels = ['Good Credit Risk', 'Bad Credit Risk']

# Plot confusion matrix
fig = plt.figure()
plot_confusion_matrix(dtc, x_test, y_test, display_labels = labels, cmap = "flare")
plt.title('Credit Risk Confusion Matrix - Decision Tree Classifier')
plt.grid(False)

plt.show()
```

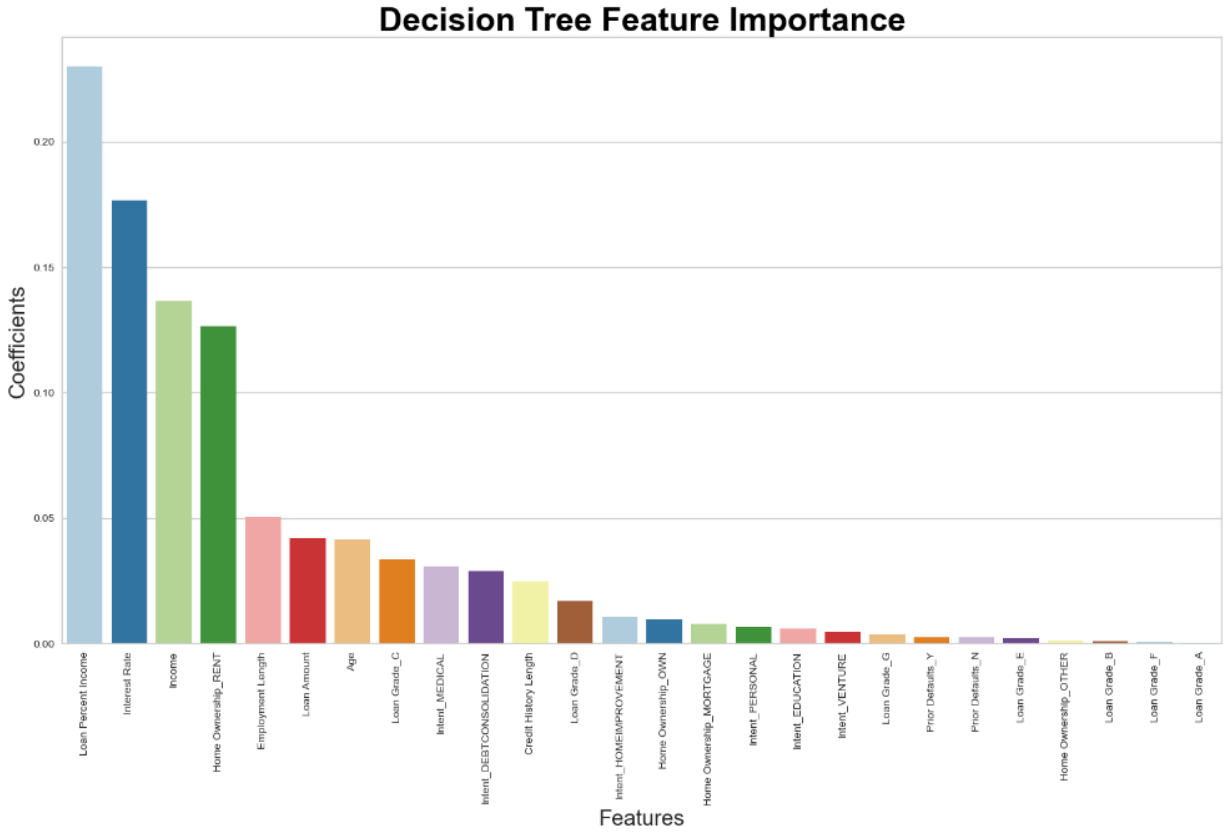
```
<Figure size 800x550 with 0 Axes>
```



The Decision Tree Classifier is a better predictor of Good Credit Risk than Bad Credit Risk. Of all the records with a Current Loan Status in good standing, 94% are considered a Good Credit Risk, while the Current Loan Status is in default, and only 73% are considered a Bad Credit Risk. The Recall scores of all the records with Good and Bad Credit Risk predicted correctly are 92% and 79%. The F1-Score combines precision and recall at 93% and 76%, respectively. Overall, the accuracy score is very good at 89%, but 11% of the records are classified as a false Good or Bad Credit Risk.

Features of the Decision Tree Classifier

```
# Display the features for the Decision Tree in order of importance

# Identify and rank the important features
importances_df = pd.DataFrame(data={
    'Attribute': x_train.columns,
    'Importance': dtc.feature_importances_
})
importances_df = importances_df.sort_values(by = 'Importance', ascending = False)

# Display the features in order of importance
fig, axes = plt.subplots(figsize = (19, 10))

import_bp = sns.barplot(x = 'Attribute', y = 'Importance', data = importances_df,
                        ci = None, palette = 'Paired')

import_bp.set_title('Decision Tree Feature Importance',
                        fontdict={'size': 30, 'weight': 'bold', 'color': 'black'})
import_bp.set_xlabel('Features', fontdict={'size': 20})
import_bp.set_ylabel('Coefficients', fontdict={'size': 20})

# rotate x-axis labels by 40 degrees
plt.xticks(rotation = 90)

# Show the plot
plt.show()
```



The top 5 features that contribute the most to predicting the accuracy of the Decision Tree Classification model are Loan Percent Income, Interest Rate, Income, HomeOwnership_Rent, and Employment Length. The five features that contribute the least are Lone Grade_E, Home Ownership_Other, Lone Grade_B, Lone Grade_F, and Lone Grade_A.

Logistic Regression Model

```
# Create the Logistic Regression Model,
# train the model, and test the results

# Create logistic regression object
logit = LogisticRegression(solver="liblinear", random_state=0)

# Train the logistic regression model
logit.fit(x_train, y_train)

# Train model to make predictions
y_pred = logit.predict(x_test)

# Calculate accuracy
ac_score = accuracy_score(y_test, y_pred)
print('The Accuracy Score For Logistic Regression = ', round(100 * ac_score, 2), '%', sep = '')
```

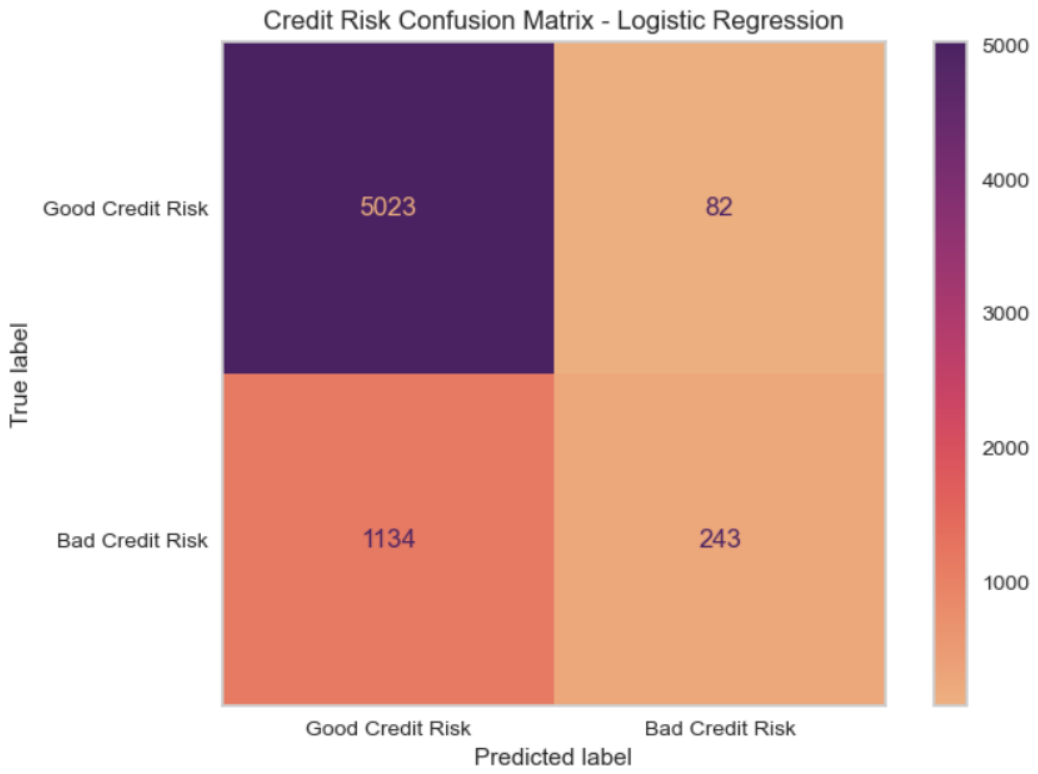
```
The Accuracy Score For Logistic Regression = 81.24%
```

```
# Double check the accuracy

class_rpt = classification_report(y_test, y_pred, target_names=target_names)

print('Classification Report For Logistic Regression')
print(class_rpt)
```

```
Classification Report For Logistic Regression
                  precision    recall  f1-score   support

Good Credit Risk       0.82      0.98      0.89      5105
 Bad Credit Risk       0.75      0.18      0.29      1377

        accuracy                           0.81      6482
       macro avg       0.78      0.58      0.59      6482
    weighted avg       0.80      0.81      0.76      6482
```

```
# create and plot a confusion matrix

# Plot confusion matrix using the labels in the Decision Tree
fig = plt.figure()
plot_confusion_matrix(logit, x_test, y_test, display_labels = labels, cmap = "flare")
plt.title('Credit Risk Confusion Matrix - Logistic Regression')
plt.grid(False)

plt.show()
```

```
<Figure size 800x550 with 0 Axes>
```

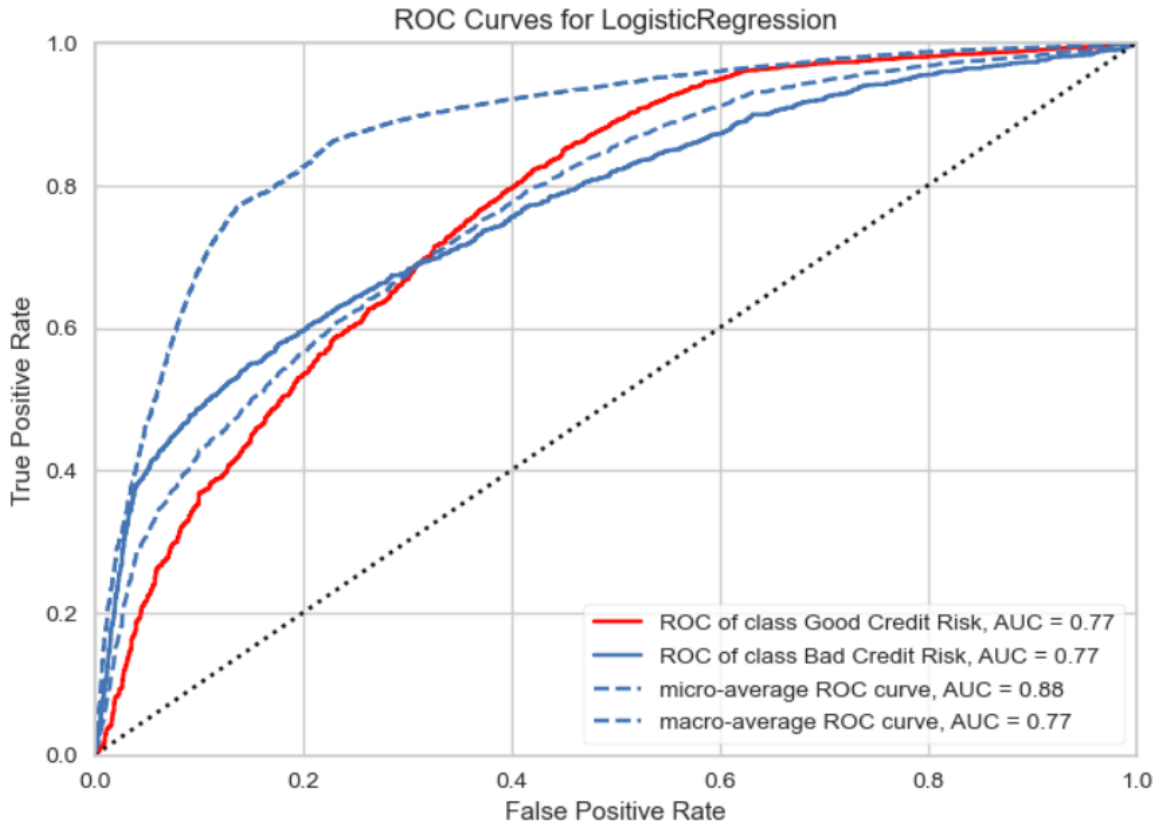


The Logistic Regression Model also predicts Good Credit Risk better than Bad Credit Risk. Of all the records with a Current Loan Status in good standing, 82% are considered a Good Credit Risk, while the Current Loan Status is in default, and only 75% are considered a Bad Credit Risk. The Recall scores of all the records with Good and Bad Credit Risk predicted correctly are 98% and 18%. The F1-Score combines precision and recall at 89% and 28%, respectively. Overall, the accuracy score is good at 81%, but 19% of the records are classified as a false Good or Bad Credit Risk.

Receiver Operating Characteristic (ROC) with Area Under the Curve (AUC) for the Logistic Regression Model

```
# Create Receiver Operating Characteristic (ROC) with Area Under the Curve (AUC)
# For the Logistic Regression Model

# Set up the ROC visualizer
ROC_labels = {0: 'Good Credit Risk', 1: 'Bad Credit Risk'}
roc_curve = ROCAUC(logit, encoder = ROC_labels, solver = 'liblinear')

# Fit the test data
roc_curve.fit(x_test, y_test)

# Display the model on the test data
roc_score = roc_curve.score(x_test, y_test)
print(roc_score)
roc_curve.show()
```

```
0.7673002898800382
```



<AxesSubplot:title={'center':'ROC Curves for LogisticRegression'}, xlabel='False Positive Rate', ylabel='True Positive Rate'>

The ROC Curve shows that the Logistic Regression Model is a good predictor but not a great predictor of Credit Risk, with an overall accuracy score of 77%.

Interpret your results:

Both the Decision Tree Classifier and Logistic Regression models give results on whether an applicant is a good or bad credit risk based on the Current Loan Status. While both models return binary results, the accuracy of the models is very different.

- The Accuracy Score for the Decision Tree Classifier was 89%, while the accuracy of the Logistics Regression model was 81%. The results are the same as part of the Classification Report or when calculated separately.
- Both models have false Good and Bad Credit Risk rates above 10%. The Decision Tree Classifier was 11%, and the Logistics Regression model was 19%.

- The Precision score on the Classification Report reflects a correct percentage of Good Credit Risk predictions. The Decision Tree Classifier performed better than the Logistic Regression model at 94% to 82%, respectively.
- The Classification Report also gives a Recall score on the proportion of actual Good Credit Risk records that were identified correctly. The Logistics Regression model had a better Recall score of 98% over the Decision Tree Classifier, with 92%
- The Confusion Matrix gives a good representation of the test records that fall into each category and are used to create the Classification Report measurements.
- The ROC (Receiver Operating Characteristics) Curve graphically represents the effectiveness of the Logistic Regression model, and the AUC (Area Under the Curve) measures the overall performance of the model. The ROC Curve for the Logistic Regression shows the observations are correctly classified as a Good or Bad Credit Risk 77% of the time.
- For the Decision Tree Classifier, the 5 features that make the biggest contribution to predicting the accuracy of the model are Loan Percent Income, Interest Rate, Income, HomeOwnership_Rent, and Employment Length.

Begin to formulate a conclusion/recommendations:

Because of higher interest rates, property values declining, and inflation, lending institutions are looking more closely at loan applicants to identify which customers are more likely to default on loans to minimize risk before extending credit. Using readily available information, can customer data on age, income, employment, home ownership, credit history, and current and historical loan information be used to predict whether credit should be extended based on prior loan history?

Two models, Decision Tree Classifier and Logistics Regression, were explored to answer this question. Between the two models used to predict good or bad credit risk, the Decision Tree Classifier had better performance with an accuracy score of 89%. In contrast, the accuracy of the Logistics Regression model was only 81%. The Decision Tree Classifier was shown to be a better predictor of credit risk, but an accuracy score of over 93% (considered excellent) would make one more comfortable with the results.

The sample dataset needs more details: when a person applies for credit, a credit score from one of the reporting agencies, the length of time residing at the current address, and other credit-related information are included. More information could increase the accuracy of the model.

Credit is essential to someone's future; extending credit at a reasonable interest rate could decide whether someone can purchase a house, pay off medical bills, buy a vehicle, or go to college. Relying solely on the outcome of a statistical model would create a significant risk of a bad decision. I recommend that an actual person review the individual parts of someone's credit profile before making a final conclusion.

References

Tory Newmyer, Aaron Gregg, and Jaclyn Peiser (August 30, 2023)

Delinquencies rise for credit cards and auto loans, and it could get worse

https://www.washingtonpost.com/business/2023/08/30/delinquencies-credit-auto-loans/

Diana Olick (November 7, 2022)

Here's how much equity U.S. homeowners have lost since May

https://www.cnbc.com/2022/11/07/homeowners-lost-1point5-trillion-in-equity-since-may-as-home-prices-drop.html

Lao Tse (2020). Credit Risk dataset. (Retrieved, September 4, 2023)

https://www.kaggle.com/datasets/laotse/credit-risk-dataset

U.S. Energy Information Administration (Retrieved, September 4, 2023)

Henry Hub Natural Gas Spot Price

https://www.eia.gov/dnav/ng/hist/rngwhhdD.htm

Visual Crossing Weather (Retrieved, September 4, 2023)

Historical Weather Data

https://www.visualcrossing.com/weather-data