Defaulting on Loans

A Risk Assessment

Daniel Alvarado

What is the problem?

-Loan default is the problem!

-Loan default occurs when a borrower fails to pay back a debt according to the initial arrangement.

-Key is to identify the variables that are indicators of potential default.

-Predictive model was built to determine the likelihood of payment issues based on the features most correlated with defaulting.

Who cares?

-Anyone looking for a loan.

-Financial Institutions.

-Focusing on personal loans in this project.

The Data

-Downloaded from https://www.kaggle.com/mishra5001/credit-card

-Case Study in risk assessment

-Built to practice using EDA in a real world scenario

-Target variable in this case is whether the borrower defaulted or not

-0 = no payment issues (no default)

-1 = payment issues (default)

The Data

Total Borrowers:

307,511

% Default:

8% (24,825 borrowers)

A few questions that guided this project:

-Do all borrowers share common features associated with default?

-Do borrowers with payment issues have red flags exclusive to them?

-What does the socioeconomic situation of borrowers look like?

Exploratory Data Analysis: Hypothesizing

-Focused on borrowers that listed their income type as 'working'.

-Hypothesize to compare the top correlated feature (the rating of the region and city the client lives in) with a borrower's employment situation.

Null: There is no significance between employment, the region a borrower lives in, and payment issues.

Alt: The relationship between employment, region rating, and payment issues is significant.

```
Ttest_indResult(statistic=42.243840533593996, pvalue=0.0)
```

Our low p-value provides statistical evidence to reject the null hypothesis and accept the alt hypothesis! There is a significance between the region rating of working borrowers and the likelihood of payment issues.

EDA

-I will summarize the relationships that were visualized and briefly include a few of the illustrations

-Distribution of float variables for all borrowers

-Income ranges of all borrowers and by gender, loan type

-Age groups

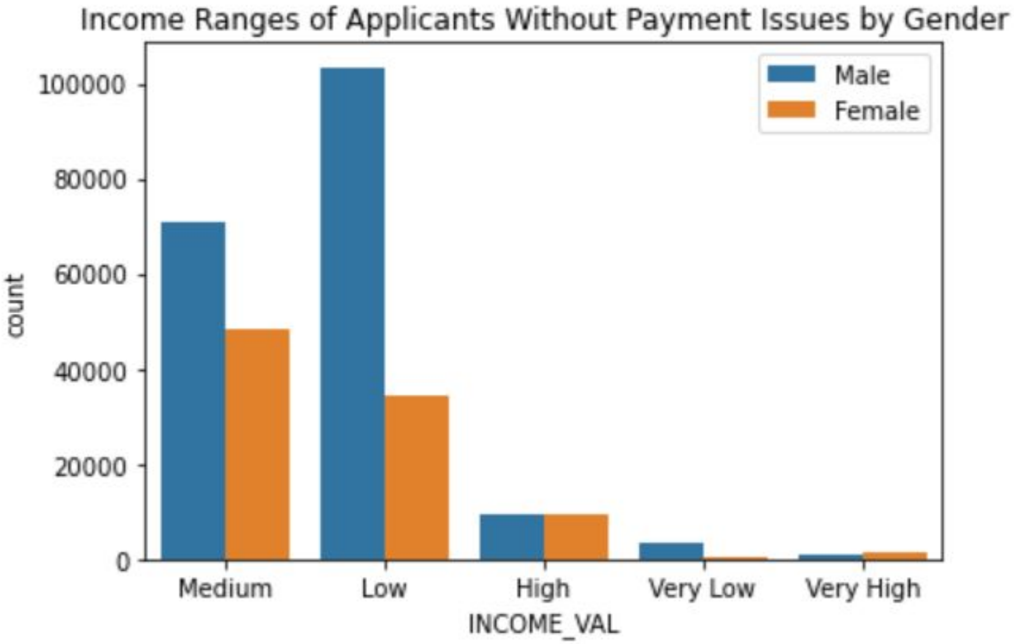
-Education types

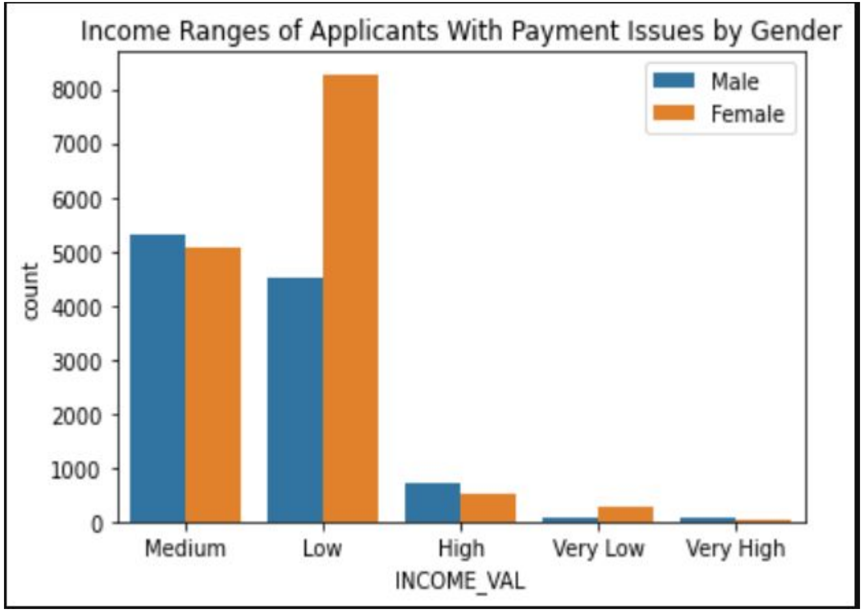
-Family status

-Region rating

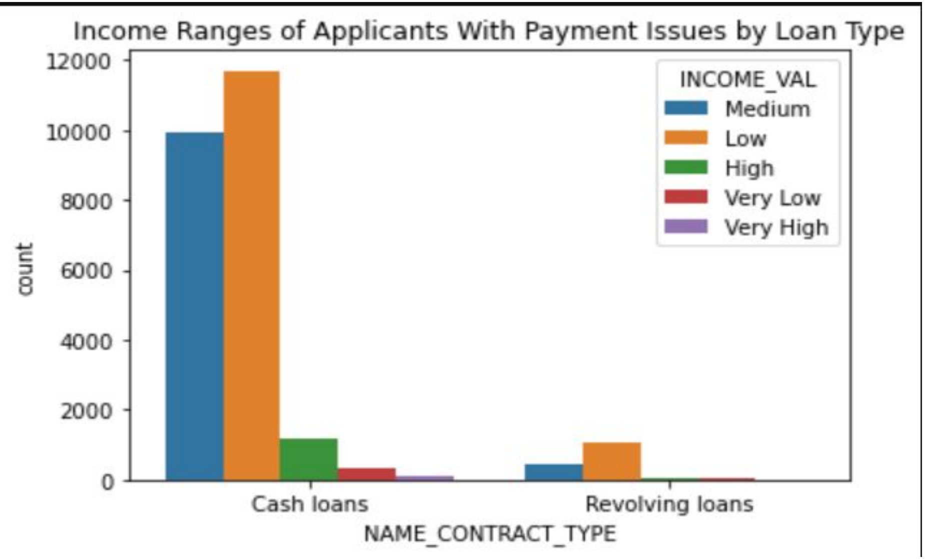
-Income totals and credit amount

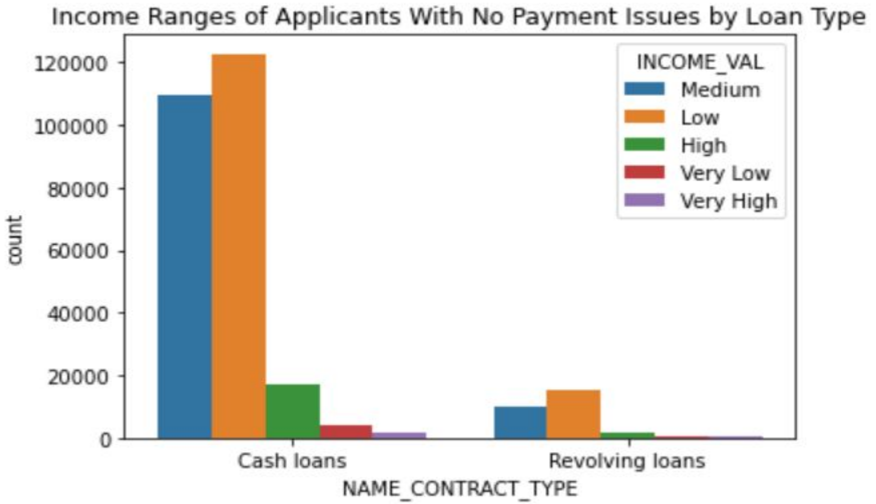
EDA: Income Ranges



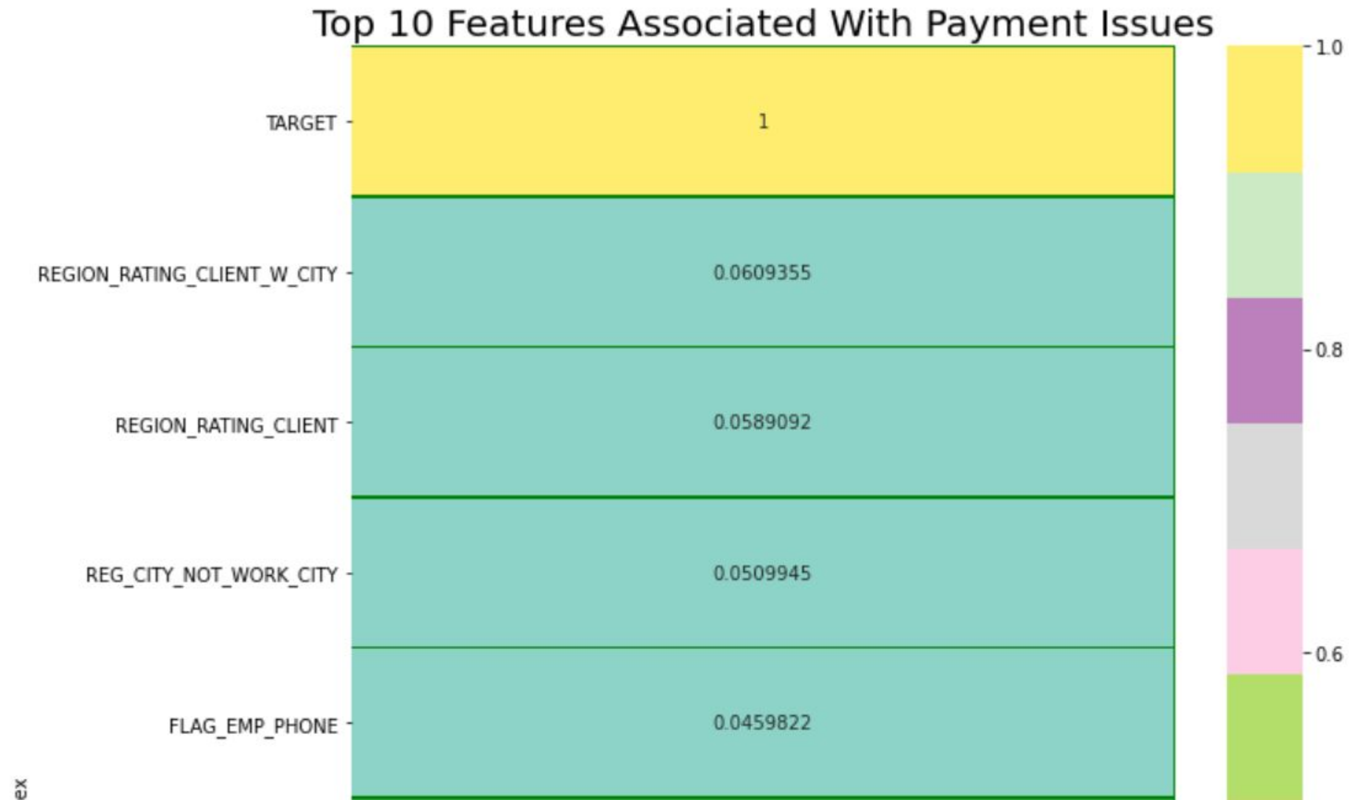


EDA: Income Ranges by Loan Type





EDA: Features Correlated With Payment Issues



Modeling and Methods

-Supervised learning problem

-Binary classification: 0 = no payment issues (no default), 1 = payment issues (default)

-Data is highly imbalanced. Oversampling and SMOTE used to remedy.

Models

The models built are as follows:

-Support Vector Machine

-Random Forest

-Logistic Regression

-K-Neighbors Classifier

-Decision Tree Classifier

Accuracy was a misleading measure of success due to the imbalanced data. In this situation, precision, recall, and f1 score are a better metric.

Modeling Steps: SVM

Pre-processing:

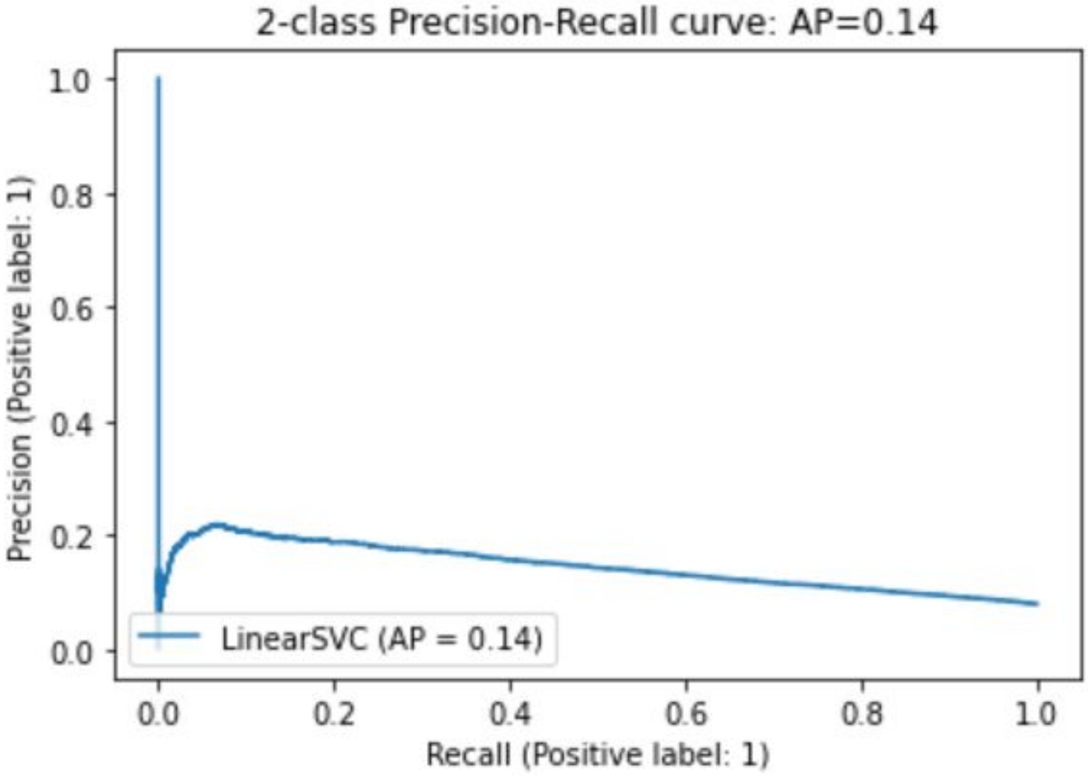
-Dummy feature generation

-Train/Test split (70/30)

-Addressed imbalance using synthetic minority oversampling technique (SMOTE) method

-Scaling numerical data

Precision-Recall Curve for Linear Support Vector Classifier



-Average precision of 0.14

-High recall, low precision means the model has a low false negative rate

Model Comparisons

	Model Name	Precision	Recall	F1 score
2	DecisionTreeClassifier	0.09	0.18	0.12
1	KNeighborsClassifier	0.14	0.01	0.03
0	LinearSVC	0.11	0.00	0.00
3	LogisticRegression	0.00	0.00	0.00

-Vanilla models
-Due to imbalanced data, the LinearSVC and Linear regression models aren't useful

	Model Name	Precision	Recall	F1 score
0	LinearSVC	0.13	0.64	0.21
3	LogisticRegression	0.13	0.63	0.21
1	KNeighborsClassifier	0.10	0.28	0.14
2	DecisionTreeClassifier	0.11	0.11	0.11

-Models with oversampling technique
-Precision, recall, and f1 scores went up dramatically for all models.

Model Comparisons

	Model Name	Precision	Recall	F1 score
0	LinearSVC	0.13	0.62	0.21
3	LogisticRegression	0.13	0.61	0.21
1	KNeighborsClassifier	0.10	0.43	0.16
2	DecisionTreeClassifier	0.08	0.94	0.15

-SMOTE models
-The KNeighbor and Decision Tree recall scores went up

-The scores of the final model
-SVM model with SMOTE technique applied for imbalanced data

```
svm_pre = sklearn.metrics.precision_score(y_test, svm_predict, pos_label = 1, average='weighted')

svm_recall = sklearn.metrics.recall_score(y_test, svm_predict, pos_label = 1, average='weighted')

svm_f1 = f1_score(y_test, svm_predict, pos_label = 1, average='weighted')

print('Support Vector Machine (SVM): precision-score=%.3f' % (svm_pre))
print('Support Vector Machine (SVM): recall-score=%.3f' % (svm_recall))
print('Support Vector Machine (SVM): f1-score=%.3f' % (svm_f1))
```

```
Support Vector Machine (SVM): precision-score=0.885
Support Vector Machine (SVM): recall-score=0.615
Support Vector Machine (SVM): f1-score=0.703
```

Further Research

-A borrower’s career would be an interesting variable to gauge default potential

-Instead of a binary classification, a probability of default would hopefully lead to an agreement that would benefit both parties.

-Rejecting a potential borrower because a machine learning model predicted that they would default is playing it safe on the lender’s part.

-A fluid scale of default potential would open up a more nuanced conversation between both parties that would ideally reduce predatory loan practices and empower the borrower to confidently take out a loan.

Recommendations

-Incentivize borrowers to opt into revolving loans instead of cash loans. This will provide a small safety net incase borrowers can't make a payment.

-Cap the amount of credit offered to higher risk individuals to 500K. Higher risk individuals in this case refers to working borrowers with low region-rating scores.

-Provide leniency/assistance to borrowers with lower socio-economic standing. Low income women make up a lot of the default cases. Getting rid of predatory loans and capping the credit limit would reduce default rates, particularly among low income women.