Optimizing Loan Approval:

A Data-Driven Model For Assessing Installment-to-Income Risk

Authors: Cameron Ela, Jackson Crawford, Jinwen “Eddie” Zhao, and Ravish Kamath

Problem Definition

Loaning monetary resources has risks that scale with rewards. As such, it is important for all parties loaning money to accurately assess clients. Successfully assessing the probability that a client will default on a loan saves money for the loaner and keeps good clients coming for business. Thus, using machine learning to predict loan default probability is a valuable application that can increase efficiency in loan application assessment. Most banks, however, already have models for predicting loan default probabilities. Therefore, we have decided to focus on building a prediction model for peer-to-peer loans. With peer-to-peer lending, individuals lend monetary funds rather than banks. A useful prediction model for peer-to-peer lending could popularize it as a viable investment strategy and expand the market for peer-to-peer lending mediators such as LendingClub. Businesses like LendingClub give individuals a common platform to negotiate peer-to-peer loans. With a successful loan default prediction model, we hope to develop a new opportunity for passive income.

Background

This project aims to significantly enhance the accuracy of loan default predictions for peer-to-peer loans facilitated by LendingClub. The predictive models employed are suboptimal, leading to ineffective loan decision-making processes. There are significant limitations to existing models and their inability to ensure reliable future results. In response, our task involves the development of an advanced model that not only incorporates traditional quantitative financial data but also integrates behavioral factors that might affect default probabilities. This comprehensive approach aims to produce a more robust model that better predicts defaults. The ultimate goal of this initiative is to refine and improve the criteria for loan approvals, thus managing risk more effectively. By doing so, we intend to enhance the stability of financial outcomes for lenders and borrowers on the LendingClub platform, thereby strengthening the reliability and efficiency of the peer-to-peer lending model.

Description of Dataset

We obtained the 5,000 loan record dataset from Lending Club. Each data point has 21 features and a “loan default" target variable. The target variable is a binary variable indicating a person's default status. Some features are traditional financial metrics used in loan assessment such as annual income, loan amount, and loan grade. There are also behavioral features such as borrower’s job, employment length, and home ownership. Through several cleaning procedures and variable creation, we divided these variables into categorical, ordinal, discrete, and continuous variables.

Methods & Experiments

1. Data Exploration

Our first method is data exploration, where we use data visualization and statistical techniques to describe the dataset's characteristics. We first separate our target and independent variables, perform univariate and bivariate analysis on key features, and conduct feature engineering as assessed through statistical tests.

<u>Target Variable:</u>

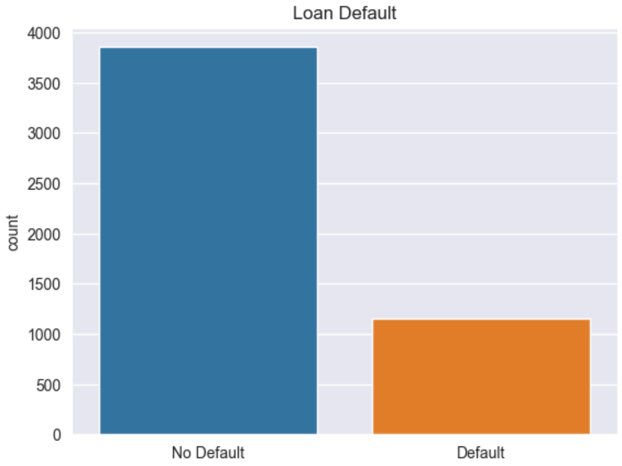

Figure 1: Loan Default Count Plot

In Figure 1, we see out of the 5,000 loan records, approximately 23% identify as defaults. The imbalance can create a bias toward the no-default class. In Section 2 (Resampling Technique), we overcome this issue using resampling.

<u>Predictor Variables:</u>

Annual Income & Installment

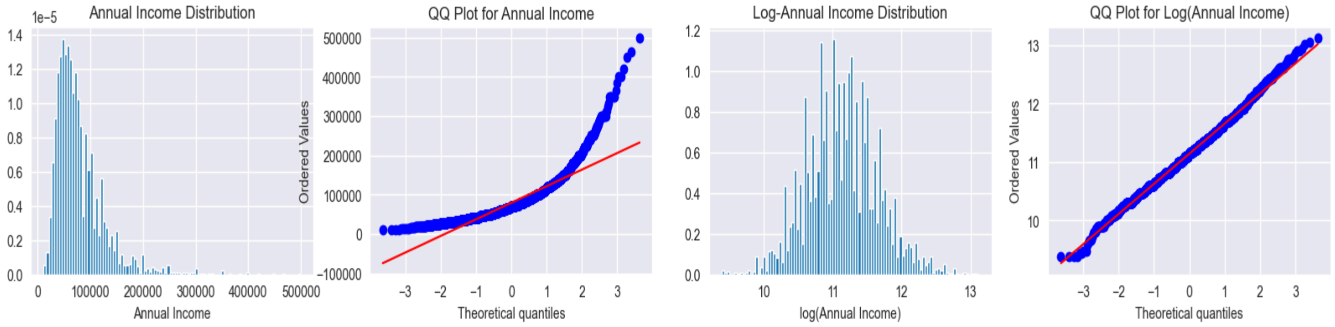

Figure 2: Annual Income Distribution and QQ plot

We first look at annual income since this is a major factor in assessing customers' loan applications. In Figure 2, the distribution of annual income is skewed heavily towards the right; hence, we apply a logarithmic transformation to create a more symmetrical distribution. This procedure allows us to standardize our features, producing better model accuracy and mitigating the influence of varying scales.

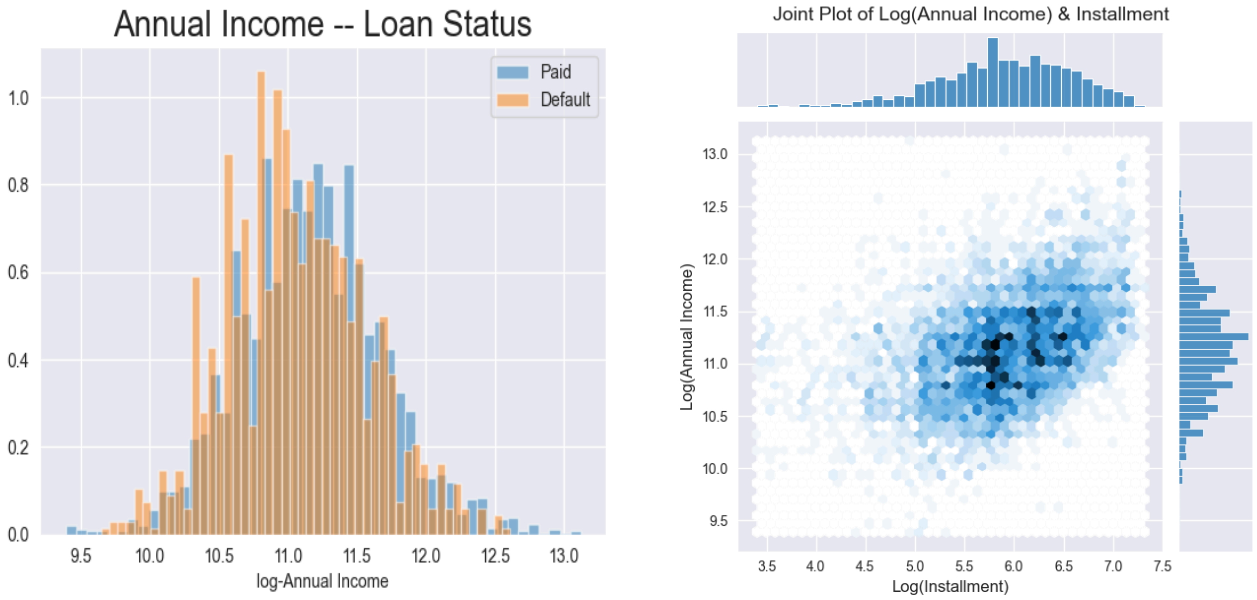

Figure 3: (left) Annual Income & Loan Status (right) Annual Income & Installment

Is annual income the key numerical metric for assessing customers? Not necessarily, since other factors like monthly installment payments could better classify customers' chances of defaulting. In the left plot of Figure 3, we look at the marginal distribution of loan status based on annual income. Both distributions are nearly symmetrical, indicating that annual income is quite similar in both classes. However, when looking at a joint distribution between annual income and installment (with a logarithmic transformation for scaling purposes), there is a positive association between the variables. This indicates that a variable transformation should be applied. Debt-to-income (DTI), or installment-to-annual income ratio, is a key measurement lenders use to assess whether a customer can pay off their loan. As shown in Figure 4, the chances of default increase as the ratio increases.

To test for the association between the install-to-income ratio and loan default, we can use the t-test. The t-test is a common statistical test used to compare the means of two groups. One important assumption of the t-test is that our records are independent and approximately normally distributed. We apply a Box-Cox transformation rather than a logarithmic transformation of the install-to-income ratio because the support is between 0 and 1. In Figure 5, we see a more symmetrical distribution. Let μ_1 be the mean of the install-to-income ratio given that the person has not defaulted, and μ_2 be the mean of the install-to-income ratio given that the person has defaulted. Here is our following hypothesis, and its alternative:

$$H_0 : \mu_1 = \mu_2 \quad H_A : \mu_1 \neq \mu_2$$

By applying the t-test, we have a p-value of 3.105367751892538e-35 ≈ 0, indicating the installment-to-income ratio is a statistically significant contributor to loan default. As such, we can firmly reject the null hypothesis and conclude that the ratio of installment to income is associated with loan default.

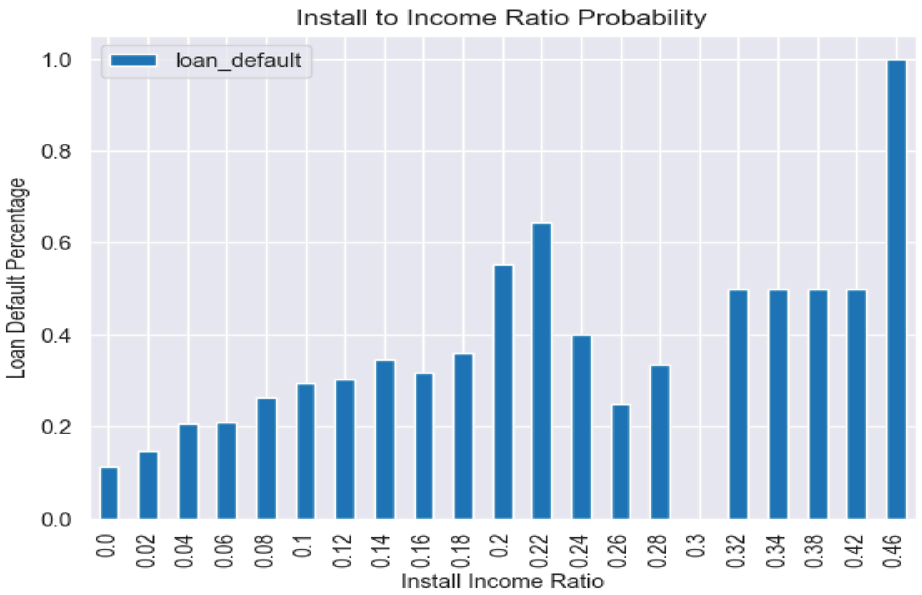

Figure 4: Install to Income Ratio Probability

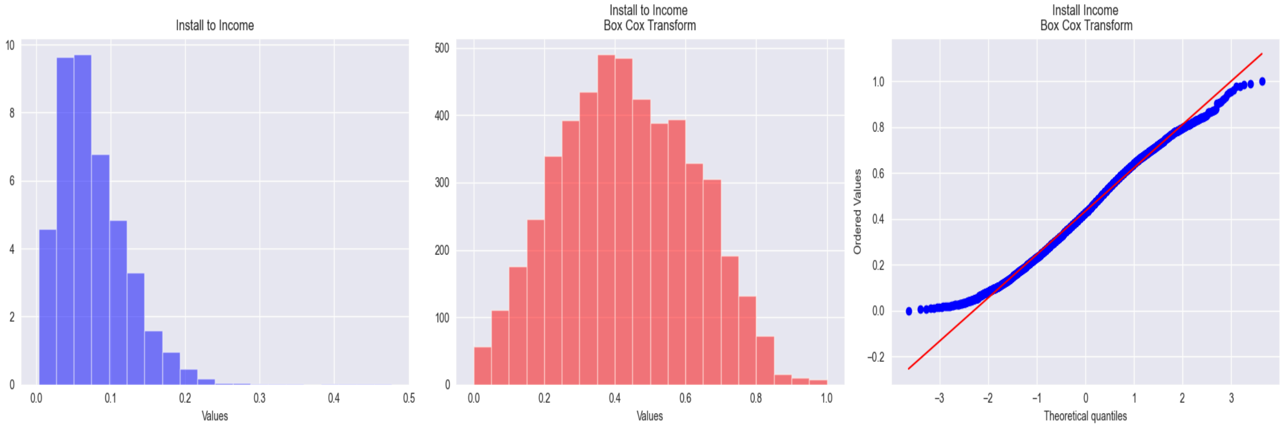

Figure 5: Installment-to-Income Ratio Transformation

Behavioral Factors

The installment-to-income ratio is a key metric to assess loan default. However, certain behavioral factors may also persuade a lender to issue loans to customers. From our dataset, we identified behavioral factors as follows:

- Home Ownership
- Employment Length
- Job

In Figure 6, depending on the individual's home ownership status, their probability of defaulting can increase on average. An interesting takeaway from this figure is people who take on a mortgage tend to have a higher annual income and a lower chance, on average, of defaulting compared to a homeowner. When looking at employment length as a behavioral contributor to loan default, we do not see a discernible pattern other than an increase in a person’s annual income.

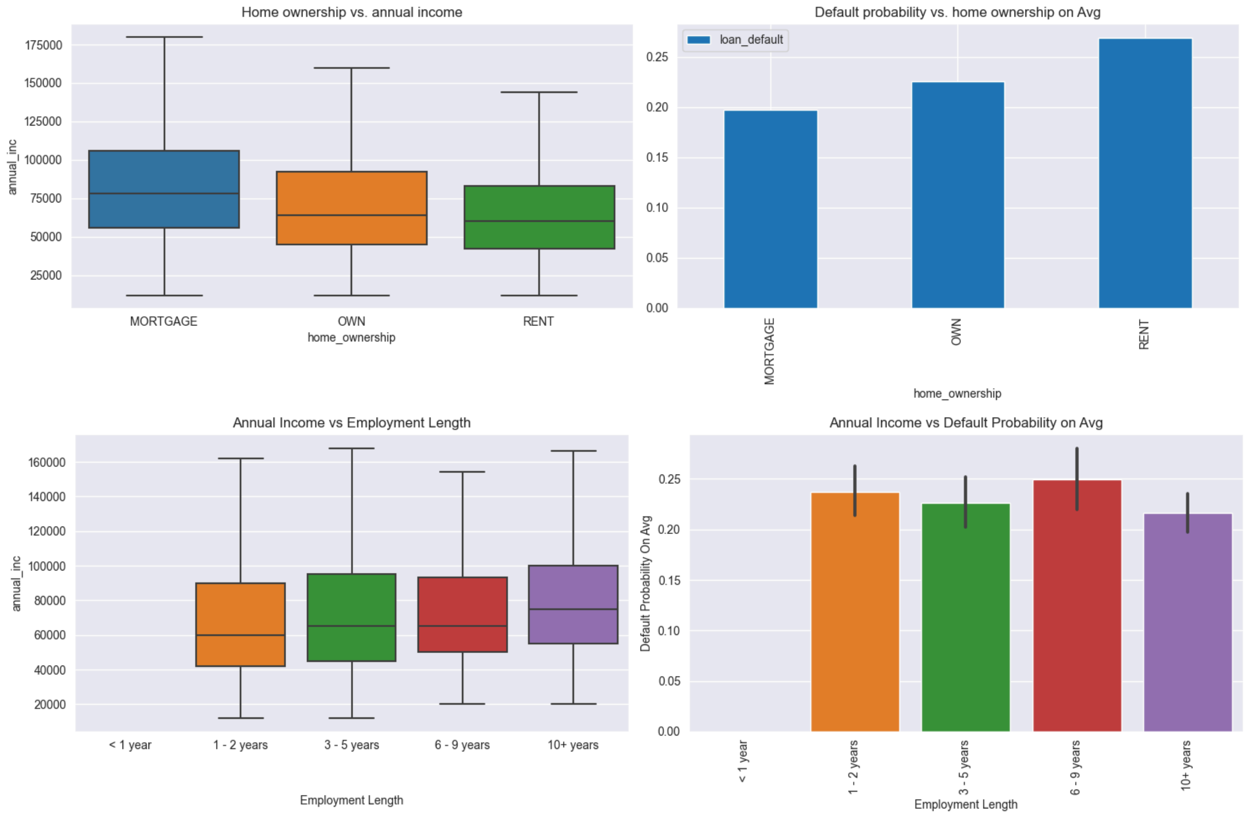

Figure 6: (Above) Home Ownership; (Below) Employment Length

The records in the dataset have 119 distinct jobs, which is dimensionally too big. Hence, we reduced the dimensions based on general job sectors such as science and technology, law, and health care. We narrowed jobs to 15 job sectors. In Figure 7, we see the percentage of records for these sectors. A key takeaway is almost half the records identify 'other' as a profession, which may not give us much information about
loan defaults.

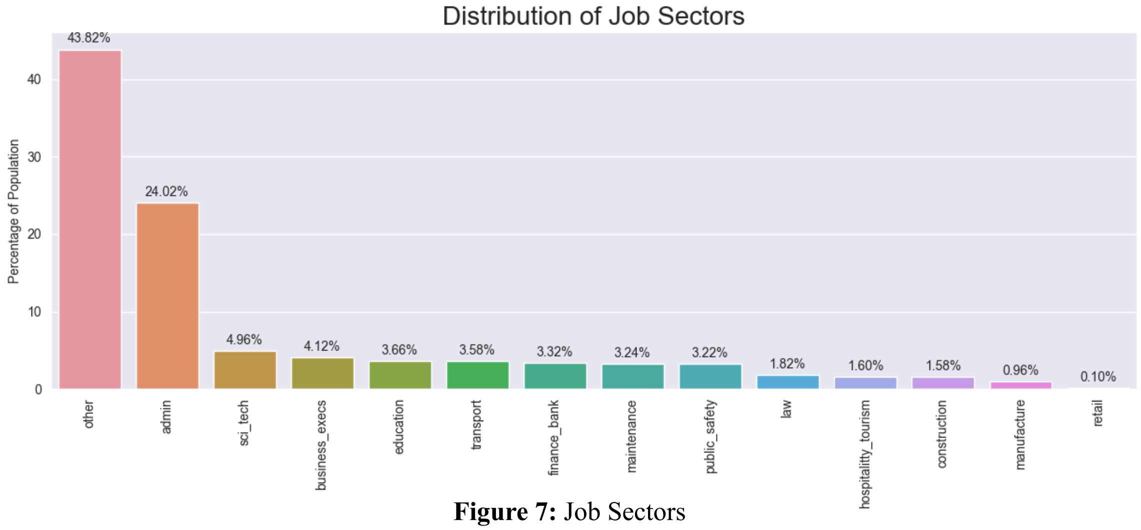

Figure 7: Job Sectors

2. Imbalanced Data Handling

In our data exploration section, we first looked at our target variable to see how balanced it is. We concluded that the data was imbalanced and needed to be reshaped. Imbalanced data is problematic because classification models can be biased towards the majority class. A classification algorithm could always predict the majority class and still be considered highly accurate. For our dataset, we handled imbalanced data in three ways.

SMOTE with Random Under Sampling:

Synthetic Minority Over-Sampling is a pre-processing technique addressing target class imbalances. Nitesh Chawla, author of the paper "SMOTE: Synthetic Minority Over-sampling Technique", suggests also using random undersampling for the majority class.

ADASYN:

This method is similar to SMOTE but generates a different number of samples depending on an estimate of the local distribution of the class to be oversampled.

SMOTE & ENN:

This combines SMOTE with Edited Nearest Neighbor (ENN) which stems from the K-Nearest Neighbors algorithm. This method combines SMOTE's ability to generate synthetic examples for the minority class and ENN's ability to delete some observations having a different class from its K-nearest neighbor majority class.

We first split the data into training and testing data, applying the sampling technique only to the training data. We then applied a logistic regression model and used F1 score and recall to assess which sampling method works best for this data. The F1 score is a good indicator for handling imbalanced data, and recall is important because we are focused on identifying loans that may default. Based on the results in Table 1, we conclude that the SMOTE ENN provides the highest score for both metrics.

Sampling Technique	F1 Score	Recall Score
No Sampling	9.77 %	16.26%
SMOTE & Down Sampling	55.37%	47.42%
ADASYN	38.11%	38.11%
SMOTE & ENN	38.11%	38.11%

Table 1: Sampling Technique Results

3. Model Building & Selection

Based on the data exploration and imbalanced data handling sections, we can proceed to test several models using all of the features. Additionally, we can select features to maximize simplicity and improve classification metrics. We will continue to use F1 score and recall to measure each model with a 10-fold cross-validation.

Model Predictors:
Model 0: Using all features available in the dataset
Model 1: Install/Income
Model 2: Install/Income + Home Ownership
Model 3: Install/Income + Employment Length
Model 4: Install/Income + Job Sector
Model 5: Install/Income + Home Ownership + Employment Length
Model 6: Install/Income + Home Ownership + Job Sector
Model 7: Install/Income + Job Sector + Employment Length
Model 8: Install/Income + Home Ownership + Employment Length + Job Sector

The classification models and their parameters are as follows:

Logistic Regression:

To fine-tune the logistic regression, we decided to use an optimization algorithm called limited memory broyden fletcher goldfarb shanno (LBFGS), which is commonly used for large scale datasets. Furthermore, we added a maximum iteration of 1000 to allow for convergence.

Decision Tree:

For our splitting criterion, we used entropy since we learned more about it in class. The other possible criterion is Gini impurity, which provides very similar results to entropy.

Support Vector Machine (SVM):

We set regularization strength to 1, giving stronger misclassification penalties. We aim to reduce the misclassifications since these could have severe negative impacts on lenders.

K-Nearest Neighbor (KNN):

We will be using the Euclidean distance as our distance metric.

Based on the results in Table 2, we see that logistic regression performs best, although all of the classification methods have similar scores. However, by using SMOTE upsampling and random undersampling, we see vast differences in all modeling scores. Comparing all the subset models to the full model (Model 0), we see that Model 2 does best in F1 score, while Model 5 does best in recall. Most importantly, these two models contain installment-to-income ratio and home ownership as predictors, suggesting these features are strongly associated with loan default.

Model	Logisitic Regression				Decision Tree				SVM				KNN			
	Resample		Original		Resample		Original		Resample		Original		Resample		Original	
	F1 %	Recall %	F1 %	Recall %	F1 %	Recall %	F1 %	Recall %	F1 %	Recall %	F1 %	Recall %	F1 %	Recall %	F1 %	Recall %
0	*42.8*	*51.99*	*20.98*	*13.17*	*35.73*	*50.28*	*30.92*	*32.12*	*39.63*	*43.48*	*0*	*0*	*39.41*	*57.54*	*23.89*	*17.90*
1	*37.55*	*48.88*	*0*	*0*	*32.09*	*46.33*	*25.08*	*25.11*	*37.16*	*47.03*	*0*	*0*	*33.54*	*46.38*	*18.84*	*13.48*
2	*38.27*	*49.75*	*0.52*	*0.26*	*33.36*	*48.06*	*25.89*	*25.98*	*37.60*	*49.61*	*0*	*0*	*33.58*	*47.84*	*19.47*	*14.06*
3	*37.42*	*48.50*	*0.11*	*0.06*	*32.40*	*47.72*	*26.36*	*26.77*	*37.57*	*47.31*	*0*	*0*	*33.11*	*47.08*	*18.08*	*12.96*
4	*37.01*	*47.69*	*0.74*	*0.38*	*32.79*	*46.64*	*25.79*	*26.13*	*34.32*	*39.99*	*0*	*0*	*33.70*	*47.34*	*19.41*	*14.18*
5	*38.09*	*50.36*	*0.52*	*0.26*	*33.02*	*48.39*	*26.80*	*27.12*	*37.43*	*47.57*	*0*	*0*	*34.35*	*47.07*	*18.10*	*13.02*
6	*38.09*	*49.96*	*1.20*	*0.61*	*32.47*	*49.12*	*27.16*	*27.70*	*36.12*	*45.08*	*0.06*	*0.03*	*32.49*	*45.33*	*18.41*	*13.02*
7	*37.06*	*48.15*	*0.80*	*0.41*	*32.53*	*47.78*	*26.69*	*27.29*	*36.88*	*46.99*	*0*	*0*	*33.85*	*46.99*	*15.64*	*11.01*
8	*38.12*	*49.66*	*1.20*	*0.61*	*33.90*	*47.71*	*27.57*	*27.70*	*37.42*	*47.25*	*0*	*0*	*32.85*	*46.73*	*17.25*	*12.35*

Table 2: Model Selection Results

Conclusion

Our goal is to improve the accuracy of loan default predictions for peer-to-peer loans facilitated by the LendingClub site. By using machine learning to predict loan default probability, we can increase efficiency in loan application assessment while increasing accuracy. Our task involves developing an advanced model to predict loan defaults based on traditional quantitative financial data and behavioral factors. Our first identifiable quantitative feature was a transformation of installments and annual income to create a debt-to-income ratio. This ratio had a significant association with our target variable. Furthermore, we identified home ownership, employment length, and job sector as possible behavioral factors that can help build a better prediction of defaulting. Before model selection, we resampled our training data to overcome the imbalance of the loan default classes. By measuring multiple classification models (F1score and recall), we concluded that having home ownership as a behavioral factor plus an installment-to-income ratio serves as the simplest and best model to predict loan defaults.

References

1. Brownlee, Jason. “Smote for Imbalanced Classification with Python.” MachineLearningMastery.com, March 16, 2021. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.
2. Chaudhary, Kartik. “How to Deal with Imbalanced Data in Classification?” Medium, September 23, 2023. https://medium.com/game-of-bits/how-to-deal-with-imbalanced-data-in-classification-bd03cfc66066.
3. Murphy, Chris B. “Debt-to-Income (DTI) Ratio: What’s Good and How to Calculate It.” Investopedia. Accessed May 1, 2024. https://www.investopedia.com/terms/d/dti.asp.