\

LendingClub Case Memo

To: Emily Figel

8/10/2025
Prashant Chopra

Prediction Tool to Invest in LendingClub Optimally

Executive Summary

In 2018, rapid fintech growth and declining confidence in traditional banks made credit assessment increasingly complex. Consequently, Emily Figel had several factors to evaluate when determining the most effective predictive strategy for managing investment decisions on the LendingClub platform.

LendingClub is a leading peer-to-peer lending marketplace offering unsecured personal loans across seven primary credit grades (A-G) and thirty-five distinct subgrades. Each subgrade represented a unique balance between yield potential and credit risk, with default rates running from approximately five percent in the highest tiers to over forty percent in the lowest. As such, designing an effective prediction tool requires balancing quantitative precision with qualitative interpretability to maintain investor confidence and long-term portfolio sustainability.

Analysis

Variable Selection and Justification

We first selected variables that best capture borrower credit quality and repayment behaviour to include in the logistic regression and classification models. The loan_status variable was converted into a binary outcome; default (Default or Charged Off) and repaid (*exhibit 1*), to reflect whether borrowers had fulfilled their obligations. The FICO score was averaged from its high and low range to yield a single indicator of creditworthiness. Annual income was adjusted for joint applications to better reflect total repayment capacity. These transformations ensured the feature set was consistent, interpretable, and met the model's predictive objective.

Both subgrade and and interest rate were included as model testing showed lower log loss when both were retained, indicating added predictive value despite correlation. Loan amount captures repayment burden and risk exposure, while debt-to-income (dti) and debt-to-income joint measures financial leverage. Finally, loan purpose was included since, according to the case, debt consolidation borrowers often continued accumulating credit card debt, making it a strong indicator of default risk.

Approach and Analysis

With the main variables selected, our next step was to build and test models that could best predict whether a borrower would default. Instead of relying solely on the hold-out set, we went ahead with a 4-fold cross-validation approach in order to get a more accurate and balanced read on model performance. This let us test our models across different portions of the data and confirm the results were not dependent on one specific split, all in a time-efficient manner. This setup forms the basis for the models we tested next: Logistic Regression and Classification (CART).

Logistic Regression

Among the models tested, logistic regression was the stronger candidate for predicting loan default on LendingClub, allowing Figel to make more informed and risk-adjusted investment decisions. The model's coefficients reveal how borrower characteristics influence default risk. Consistent with financial metrics, higher FICO scores reduce the odds of default, while higher debt-to-income ratios increase them (*exhibit 4*). Quantitatively, logistic regression demonstrated slightly superior predictive performance compared to other approaches, achieving a LogLoss of approximately **0.339** and an accuracy of **0.874**. These results suggest well-calibrated probability estimates and reliable overall performance.

While logistic regression offers clear advantages, it has some limitations that Emily Figel should consider. The model assumes linear relationships between predictors and defaults risk in log-odd space, which may oversimplify complex behaviours. An example would be the effect of dti on default not being perfectly linear. Moreover, interpreting coefficients and log-odds can be less intuitive for non-technical users. Thus, Figel's investment team may face some challenges in translating model outputs into actionable decisions. Finally, although logistic regression slightly outperforms alternative models numerically (LogLoss of **0.339** versus **0.37**), the gain may be marginal. As a result, the improvement may not justify the added complexity for teams that value straightforward, visually interpretable reasoning.

Classification Tree (CART)

The classification tree model closely mirrors the intuitive reasoning that investors like Emily Figel already use when evaluating borrowers. By sequentially splitting data based on key risk indicators, the model effectively simulates human decision-making. This makes it more intuitive and trustworthy for non-technical users. Additionally, the model enhances transparency and accountability since every decision can be traced back to clear, rule based conditions within the tree. The traceability supports compliance and risk management by allowing stakeholders to easily audit loan risk classifications. The classification tree also achieved slightly higher overall accuracy (**85%** compared to **84%**), performing comparably to logistic regression while remaining easy to interpret and communicate.

While the classification tree is intuitive and easy to explain, it comes with some trade-offs. Its predictions are slightly less precise than those of logistic regression, with a higher LogLoss of about **0.376**. Because decision trees classify borrowers into broad risk categories, they offer less nuance of Figel's investment decisions. The model is also prone to overfitting. Without techniques of pruning or restricting tree depth, it can end up memorizing training data rather than learning patterns that generalize to new borrowers. Additionally, in our case, the model predicts that all individuals will repay their loans (*exhibit 3*), which prevents us from visualizing the rule-based conditions. This is problematic because the main benefit of a classification tree is the ability to see the different paths a given data point can follow.

Recommendation

Both LogLoss and accuracy are useful for evaluating model performance, but they offer different benefits for Figel's investment strategy. LogLoss measures how well predicted probabilities match actual outcomes. Although lower values are better, extremely low ones may indicate overfitting. Given the uncertainty in borrower behaviour, this risk is worth noting.

Accuracy, by contrast, is easier to interpret and communicate. It shows how often the model correctly classifies borrowers. When paired with a confusion matrix, the model helps visualize trade-offs between false positives and false negatives. This makes it more practical for Figel's team, as they can choose a model that aligns with their risk tolerance. For reference, a risk-averse individual would prioritize minimizing false negatives, since they would catch more potential defaulters (*exhibit 2*).

Ultimately, the logistic regression achieved a slightly lower LogLoss and nearly identical accuracy to the classification tree, thus offering the most balanced and reliable prediction tool. We recommend adopting the **logistic regression model alongside cross-validation** to support data-driven, interpretable, and consistent lending decisions on LendingClub.

Appendix

Exhibit 1:

```
# Ensuring that loan_status only includes values that Figel highlighted
lc_data <- lc_data %>%
  filter(loan_status %in% c(
    "Fully Paid", "Charged Off", "Default",
    "Late (31-120 days)", "Late (16-30 days)", "Current"
  ))


lc_data <- lc_data %>%
  mutate(
    loan_status = if_else(loan_status %in% c("Charged Off", "Default"),
                          "Default", "Repaid"),
    loan_status = factor(loan_status, levels = c("Repaid", "Default")),
    fico = (fico_range_low + fico_range_high) / 2,
    income = ifelse(application_type == "Individual",
                    annual_inc, annual_inc_joint),
    combined_dti = ifelse(application_type == "Individual",
                          dti, dti_joint)
  )
```

Exhibit 2

```
> print(log_conf_matrix)
         y_pred
y_true    Repaid Default
  Repaid   86686     379
  Default  12135     293
```

Exhibit 3

```
> print(cart_conf_matrix)
         y_pred
y_true    Repaid Default
  Repaid   87065       0
  Default  12428       0
```

Exhibit 4

```
Coefficients:
                                     Estimate    Std. Error z value             Pr(>|z|)    
(Intercept)                       -0.2576029659  0.1709138529  -1.507             0.131757    
purposecredit_card                 0.1648226865  0.0570600918   2.889             0.003870 ** 
purposedebt_consolidation          0.1922040056  0.0563027229   3.414             0.000641 ***
purposeeducational                -0.3016160405  0.4983273319  -0.605             0.545008    
purposehome_improvement            0.1785692715  0.0594981689   3.001             0.002689 ** 
purposehouse                       0.2438071102  0.0810325837   3.009             0.002623 ** 
purposemajor_purchase              0.2212941824  0.0657523334   3.366             0.000764 ***
purposemedical                     0.2186122292  0.0718562346   3.042             0.002347 ** 
purposemoving                      0.3079250365  0.0781792903   3.939  0.00008192312730821 ***
purposeother                       0.1135828585  0.0592857698   1.916             0.055384 .  
purposerenewable_energy            0.0512692020  0.2101523816   0.244             0.807260    
purposesmall_business              0.4248001762  0.0706501729   6.013  0.00000000182428529 ***
purposevacation                    0.2334258004  0.0820312013   2.846             0.004433 ** 
purposewedding                    -0.4844508540  0.2227357151  -2.175             0.029630 *  
int_rate                         -21.8317396281  0.3735563997 -58.443 < 0.0000000000000002 ***
loan_amnt                          0.0000088431  0.0000006615  13.368 < 0.0000000000000002 ***
verification_statusSource Verified  0.1685008861  0.0127231465  13.244 < 0.0000000000000002 ***
verification_statusVerified        0.2794169516  0.0137545525  20.315 < 0.0000000000000002 ***
fico                              -0.0035139254  0.0002051478 -17.129 < 0.0000000000000002 ***
combined_dti                       0.0044397521  0.0006124078   7.250  0.00000000000041780 ***
sub_gradeA2                        0.5924921931  0.0759571789   7.800  0.00000000000000617 ***
sub_gradeA3                        0.8469340439  0.0724739436  11.686 < 0.0000000000000002 ***
sub_gradeA4                        0.9893849352  0.0671126380  14.742 < 0.0000000000000002 ***
sub_gradeA5                        1.4586646284  0.0645233630  22.607 < 0.0000000000000002 ***
sub_gradeB1                        1.9341352290  0.0632945075  30.558 < 0.0000000000000002 ***
sub_gradeB2                        2.1817110482  0.0636927996  34.254 < 0.0000000000000002 ***
sub_gradeB3                        2.4888119984  0.0638233184  38.995 < 0.0000000000000002 ***
sub_gradeB4                        2.7071981593  0.0639706536  42.319 < 0.0000000000000002 ***
sub_gradeB5                        2.9602497159  0.0646206782  45.810 < 0.0000000000000002 ***
sub_gradeC1                        3.3062345921  0.0654109231  50.546 < 0.0000000000000002 ***
sub_gradeC2                        3.5775875909  0.0668345364  53.529 < 0.0000000000000002 ***
sub_gradeC3                        3.7732616757  0.0677912219  55.660 < 0.0000000000000002 ***
sub_gradeC4                        4.0440234582  0.0691823028  58.455 < 0.0000000000000002 ***
sub_gradeC5                        4.2313169838  0.0711729562  59.451 < 0.0000000000000002 ***
sub_gradeD1                        4.5707354771  0.0737816568  61.949 < 0.0000000000000002 ***
sub_gradeD2                        4.7609455816  0.0766869838  62.083 < 0.0000000000000002 ***
sub_gradeD3                        5.0597981766  0.0787479750  64.253 < 0.0000000000000002 ***
sub_gradeD4                        5.3650243164  0.0805349103  66.617 < 0.0000000000000002 ***
sub_gradeD5                        5.6570383767  0.0832658062  67.940 < 0.0000000000000002 ***
sub_gradeE1                        5.8107844810  0.0849491708  68.403 < 0.0000000000000002 ***
sub_gradeE2                        6.0579865759  0.0873861608  69.324 < 0.0000000000000002 ***
sub_gradeE3                        6.3082389728  0.0901887732  69.945 < 0.0000000000000002 ***
sub_gradeE4                        6.5500580485  0.0933487457  70.168 < 0.0000000000000002 ***
sub_gradeE5                        6.7331134474  0.0968807687  69.499 < 0.0000000000000002 ***
sub_gradeF1                        6.9134475417  0.1028884704  67.194 < 0.0000000000000002 ***
sub_gradeF2                        7.2352748652  0.1102423644  65.631 < 0.0000000000000002 ***
sub_gradeF3                        7.4639403072  0.1154339190  64.660 < 0.0000000000000002 ***
sub_gradeF4                        7.5456449696  0.1206354602  62.549 < 0.0000000000000002 ***
sub_gradeF5                        7.8956252850  0.1257409616  62.793 < 0.0000000000000002 ***
sub_gradeG1                        7.9590480982  0.1363923230  58.354 < 0.0000000000000002 ***
sub_gradeG2                        8.0034538134  0.1469125801  54.478 < 0.0000000000000002 ***
sub_gradeG3                        8.2960115707  0.1626452341  51.007 < 0.0000000000000002 ***
sub_gradeG4                        8.2068495653  0.1747333501  46.968 < 0.0000000000000002 ***
sub_gradeG5                        8.2230059255  0.1838698520  44.722 < 0.0000000000000002 ***
income                            -0.0000032424  0.0000001333 -24.330 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```