|  |  |  |
| --- | --- | --- |
| Variable | Operation | Rationale |
| AppReceiveDate | Converted dates to the number of days since New Year, 2015 | Helps with parsimony when interpreting the coefficient values in the model |
|  | Created squared and cubed transformations of the converted dates, AppReceiveDate\_sq and AppReceiveDate\_cu | These improve the model |
| ModifiedCreditScore | Created dummy variables for different grades of credit. isCredit\_F is 1 if ModifiedCreditScore = 0, otherwise it is 0. isCredit\_D works similarly for scores of 400 to 629, as does isCredit\_C for scores of 630 to 669, and isCredit\_B for scores of 670 to 739. | These are necessary to match the behavior of ModifiedCreditScore with respect to loan approval. (see graph) |
|  | Created squared and cubed transformations, ModifiedCreditScore\_sq and ModifiedCreditScore\_cu | As above |
| ModifiedBankruptcyScore | Created squared transformation, ModifiedBankruptcyScore\_sq | A cubic transformation of this variable was found to be unnecessary |
|  | Created BankruptcyCredit, which is ModifiedCreditScore \* ModifiedBankruptcyScore | An interaction term, accounts for correlation between credit and bankruptcy scores |
| EmployedMonths | Removed from the training set all loans with values greater than 1200 (4 in total) | Improves the fit by removing obviously erroneous data |
|  | Created CurrentEmployedMonths\_ln, which is the same as log(1 + EmployedMonths) except it is set to 0 if the EmploymentStatus is anything other than “Employed” | This variable only seems to matter when applicant is currently employed, in which case the relationship between months of employment and approval rate looks logarithmic |
| TotalMonthlyIncome | Removed from the training set all loans where TotalMonthlyIncome was less than $1000 (629 in total) | Almost all of these were declined, and they interfere with the fitting process. Any loans like this in the test set will have their predictions overridden |
|  | Created TotalMonthlyIncome\_ln, which is log(1 + TotalMonthlyIncome) | Past the $1000 mark, the relationship between income and approval rate looks logarithmic |
| TotalMonthlyDebtBeforeLoan | Created isLowDebt, which is 1 if TotalMonthlyDebtBeforeLoan is below $300, 0 otherwise | Approval rates suddenly go up right after the $300 mark. (see graph) |
| VehicleMake | Created isUndecided, which is 1 if VehicleMake is “Undecided”, 0 otherwise | VehicleMake was too troublesome, but Undecided vehicles were too much of an outlier to ignore |
| VehicleMileage | Created isHighMileage, which is 1 if Mileage >= 100000, 0 otherwise | Approval rates suddenly drop to nearly nothing at the 100,000 mile mark |
| TotalVehicleValue | Withheld vehicles worth more than $1M from the training set (1 in total) | This one observation was clearly an error, unless he was trying to buy a Nissan helicopter or something like that. It was causing a lot of problems, so out it goes. |
|  | Created TotalVehicleValue\_ln, which is log(1 + TotalVehicleValue) | Vehicle value has a roughly logarithmic relationship with approval rate |
| AmountRequested | Withheld loans of more than $200,000 from the training data | Extreme outliers |
| OccupancyStatus | Created isHomeowner, which is 1 if OccupancyStatus is either Buying or Own, 0 otherwise | The Buying and Own categories seemed to behave one way, and the rest of the categories behaved another way |
| EstimatedMonthlyPayment | Created EstimatedProfit, which is AmountRequested – (Loanterm \* EstimatedMonthlyPayment). Loans with an EstimatedProfit between $0 and $1 are removed from the training data, and overridden after being predicted (24,942 in total) | 3 of those were approved. Throwing these out and training on what remains substantially improves the model when combined with an arbitrary override |
| LTV | Set all null values to 1 | These are all cases where the make is undecided and the vehicle value is set to 0 (therefore the null values are caused by division by 0). I think it’s reasonable to assume that these people would want a car roughly the same amount as their loan |
|  | Set all values higher than 2.5 to 2.5 | Reduces the influence of outliers |
|  | Created LTV\_sq, which is LTV \* LTV | Roughly polynomial relationship w approval rate |
| DTI | Set all values higher than 0.5 to 0.5 | Virtually all of these were rejected, so this helps prevent some overfitting |
|  | Set all null values to 0.5 | Assume the worst in this case |
|  | Created DTI\_sq and DTI\_cu | Explicit polynomial relationship with approval rate |

**Model equation:** Approved ~ AppReceiveDate\_cu + AppReceiveDate\_sq + AppReceiveDate + isCredit\_F + isCredit\_D + isCredit\_C + isCredit\_B + ModifiedCreditScore\_cu + ModifiedCreditScore\_sq + ModifiedCreditScore + BankruptcyCredit + ModifiedBankruptcyScore\_sq + ModifiedBankruptcyScore + CurrentEmployedMonths\_ln + TotalMonthlyIncome\_ln + isLowDebt + isUndecided + isHighMileage + isNewVehicle + TotalVehicleValue\_ln + AmountRequested + isHomeowner +LTV\_sq + LTV + DTI\_cu + DTI\_sq + DTI + MemberIndicator + CoApplicantIndicator