Lending Club Data Prediction Model Documentation.

Prepared by: Patrick Hunter

Statement of Purpose:

The purpose of this model is to make a baseline prediction as to the future loan status of a given loan from Lending Club, at the time of the origination of the loan. This prediction should be independent of any factors which arise from the underwriting of the loan itself. The goal is to predict if a loan will be fully paid or charged off, in order to reduce the quantity of charged off loans for Lending Club.

Stakeholders:

I am the model developer and primary stakeholder for this project.

Data Description & Preparation:

- Data quality assessment / treatment:

The original full dataset contains 151 columns and roughly 1.8 million rows. For ease of use, this model was trained on a sample of the full dataset containing 20,000 rows.

- Filtering Features and Observations:
    - All ID columns were removed from the dataset. These would not be useful in the model.
    - The choice was made to focus the model on non-joint applications. All columns related to join applications were removed from the dataset, and the dataset was filtered to remove observations for joint applications.
    - Hardship and settlement related columns were removed from the dataset, so as to prevent target leakage, since these columns were only relevant when a hardship was present and the loan was more likely to be charged off in that case.
    - Date related columns were first inspected to provide better context for the data, before either transforming them or removing them from our dataset.

        - The ‘last_payment_d_date’ field revealed that the Jan-2019 was the timing of the last data in our dataset.
        - In analyzing the relationship between issuance date and the percentage of loans for each date which were charged off, it seemed that the relatively steady relationship between these features began to change in January of 2017. It was hypothesized that this was likely due to the fact that with data ending in 2019, we were not seeing if some of the loans that were issued after January of 2017 were fully paid or charged off because they had not come close enough to maturity by January 2019. For this reason, the dataset was filtered to include records for loans that were issued prior to 2017.
    - Features related to the loan amount ("loan_amnt", "funded_amnt", "funded_amnt_inv", "int_rate", "installment", "issue_d") were removed from the dataset. These features were removed so as to remove any underwriting related factors from the prediction. The prediction is meant to be made independently of any decisions made in underwriting.
    - Features such as the ‘Grade’, Subgrade’ and ‘Term’ were removed for the same reason.
        - At first, the ‘Grade and Subgrade categorical features were converted to a numeric scale and used in the model. Using the permutation method of feature importance, that feature was by far the most important feature in the model. This raised a red flag, and it was determined that the grade was reflective of the underwriters assessment of risk, which we are trying to maintain independent from our prediction. Removing the Grade from the model had minimal impact on the AUC score, which indicated that is was likely heavily correlated with our other features of importance.
        - Because the term of the loan was determined in the underwriting process in relation to the overall risk of the loan, term was also highly correlated with the other features of importance which were related to risk of the loan being charged off.
    - Features related to post loan information which may present target leakage ("loan_status", "out_prncp", "out_prncp_inv", "total_pymnt", "total_pymnt_inv", "total_rec_prncp", "total_rec_int", "total_rec_late_fee", "recoveries", "collection_recovery_fee", "last_pymnt_d", "last_pymnt_amnt", "next_pymnt_d", "last_credit_pull_d","last_fico_range_high","last_fico_range_low"), were also removed from the dataset.
    - The ‘emp_title’ feature was removed from the dataset because of very high cardinality, making it’s use difficult for the purpose of prediction.
    - The ‘pymnent_plan’ feature was dropped due to it basically having a cardinality of 1. Only 3 of roughly 13000 observations had the value ‘y’.
    - Location based features like ‘zip_code’ and ‘addr_state’, representing the zip code and State respectively, were removed so as to not introduce any ethical issues related to prediction charge off based on location to the model.

    - The ‘desc’ feature was dropped from the model. If was sparsely populated, but it was analyzed using text analysis methods such as TF-IDF analysis and word cloud analysis to attempt to extract predictive features. It was determined that any predictive nature of this field was related to the ‘purpose’ feature, and thus not necessary to include in the mode.
        - The ‘Title feature was also found to be related to the ‘purpose’ feature, but this feature had much higher cardinality. For that reason, it was dropped from the dataset.
    - The ‘fico_range_high’ feature was dropped from the dataset, as this is correlated heavily to fico_range_low, and provides the same information to our models.
- Null Treatment:
    - There were a number of fields: ('tot_coll_amt', 'tot_cur_bal','mort_acc', 'total_rev_hi_lim', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m','num_tl_op_past_12m', 'percent_bc_gt_75', 'pub_rec_bankruptcies','tot_hi_cred_lim', 'total_bal_ex_mort','total_bc_limit','total_il_high_credit_limit', 'bal_to_inc', 'emp_length_num', ‘inq_fi’, ‘inq_last_12m’, ‘total_cu_tl’, "open_acc_6m", "open_act_il", "open_il_12m", "open_il_24m", "mths_since_rcnt_il", "total_bal_il", "il_util", "open_rv_12m", "open_rv_24m", "max_bal_bc", "all_util", "acc_open_past_24mths", "avg_cur_bal", "bc_util", 'mort_acc', 'num_accts_ever_120_pd') which had a high number of missing values, and those missing values more than likely were were due to the field not being filled out because it was not relevant to the application. It was hypothesized that those were not relevant to the individual observation because they would have a value of 0. For that reason those values were set to 0, and a missing value flag was created for each of these features.
- Feature Engineering:
    - A feature for the length of the observations credit history (earliest_cr_line_years) was created using the features for issue due date (issue_d_date) and first credit line date (earliest_cr_line_dt).
    - A total account balance to income ratio feature was created by dividing ‘total_cur_balance’ by ‘annual_inc’
    - Similarly , an installment to income ratio feature was created.
    - A ‘never_delinquent’ flag was created to indicate when the ‘mths_since_last_record’ and mths_since_last_delinq’ fields were both null.
    - The’grade’ and ‘sub_grade variables were converted to a numeric scale. This feature was shown to be highly predictive, but was later removed from the model to maintain model independence from underwriting decisions.
    - The ‘emp_length’ variable was converted from string to numeric, to make it useful for prediction.

- Target Engineering:
    - The target variable was set to 1, when ‘loan_status’ equaled ‘charged off’.
- Data splitting:
    - In the XGBoost model, the data was split into a test, train split, using a test size of 0.2.

Methodology:

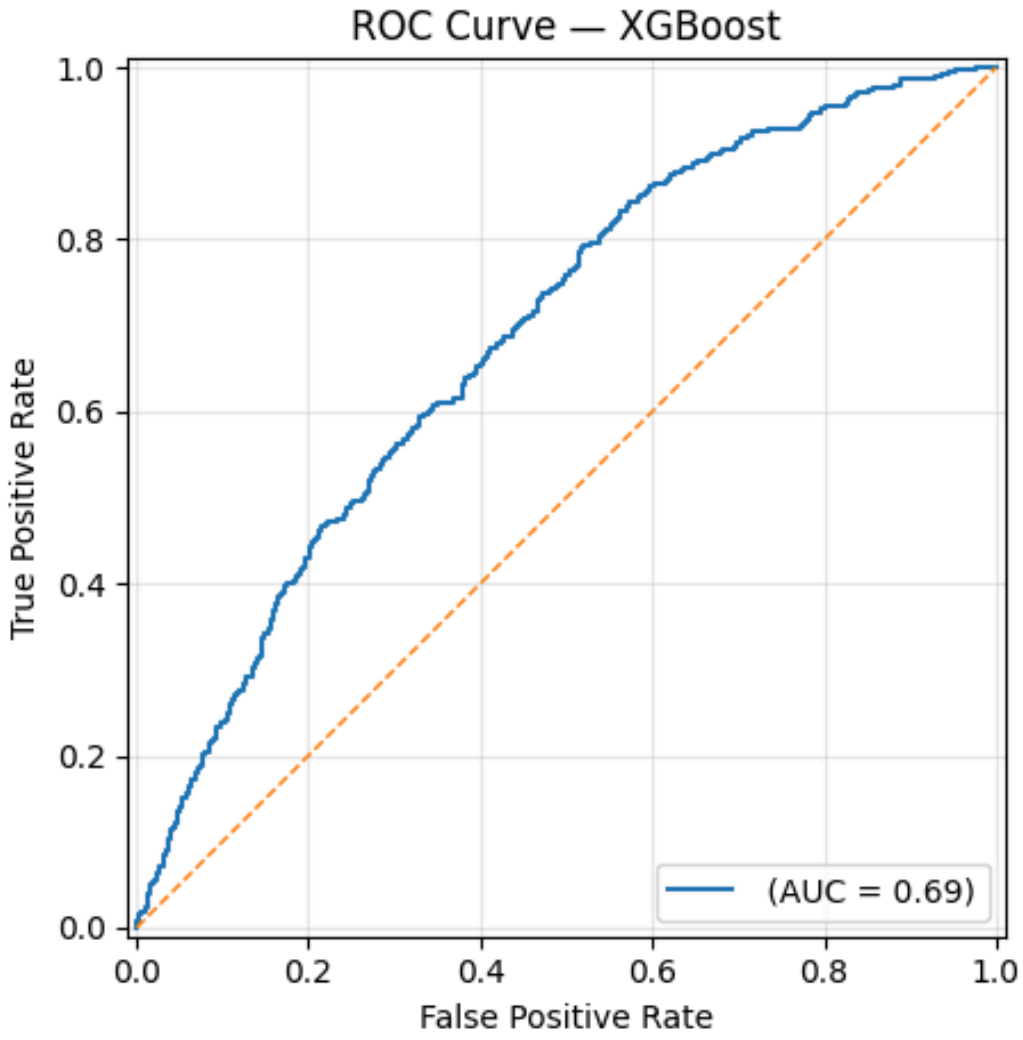
- Rational for chosen models:
    - The XGboost model was chosen because it is consistently one of the most predictive models for classification purposes. It handles categorical data well, even with high cardinality present, which was important with our ‘purpose’ feature. It also handles null values effectively, which was important as we had a number of fields that, after cleaning, still contained null values.
    - To counterbalance the impact of potential overfitting that could be present due to the XGboost model, a Random Forest model was chosen as the comparison model. This was done because the random forest increases the overall bias in the model with its regularization effects, which should decrease model variance and overfitting. Comparing the feature importances of the random forest model with the XGboost model will give us an idea of the overall robustness of the primary features which are responsible for making predictions in our XGboost model.
- Assumptions:
    - In both models, we are making the assumption that features involving the underwriting decisions such as ‘grade’, ‘subgrade’, and ‘term’ should be left out of the model. While those features increase prediction accuracy, they are reflective of the underwriters risk assessment of each observation, and we are assuming that the model is used to make a prediction on the likelihood of the loan being charged off, which would be used prior to underwriting.
    - We are assuming that the explainability of the XGboost and Random forest models are acceptable in the eyes of regulators, for the purpose of this model.
- hyperparameter selection & tunning
    - XG Boost Model:
        - Randomized search was performed along with 5 fold cross validation to choose the best model values for the following hyperparameters:
          "Model__n_estimators", "model__max_depth",
          "model__min_child_weight", "model__subsample",
          "model__colsample_bytree", "model__gamma",
          "model__reg_alpha", "model__reg_lambda",
          "model__learning_rate".

            - The model used 60 iterations, and was scored using ROC AUC.
        - After the search, the following values were chosen for each hyperparameter:
            - ‘model__subsample': 0.7, 'model__reg_lambda': 3.0, 'model__reg_alpha': 0.1, 'model__n_estimators': 250, 'model__min_child_weight': 3, 'model__max_depth': 6, 'model__learning_rate': 0.02, 'model__gamma': 0.5, 'model__colsample_bytree': 1.0
        - These parameters were chosen because these are the most impactful parameters in the XGboost Model.
    - Random Forest Model:
        - I manually set the following hyperparameters for this model. I decided to not perform hyper parameter tuning on this model, due to the high compute required to do so, and the resources at my disposal. My goal is to use this model is meant to serve as a benchmark for the XGBoost model, rather than to extract maximum predictive ability from the model.
            - n_estimators=600, max_depth=12, min_samples_leaf=2, min_samples_split=10, max_features="sqrt", class_weight="balanced_subsample", bootstrap=True.

Performance Assessment:

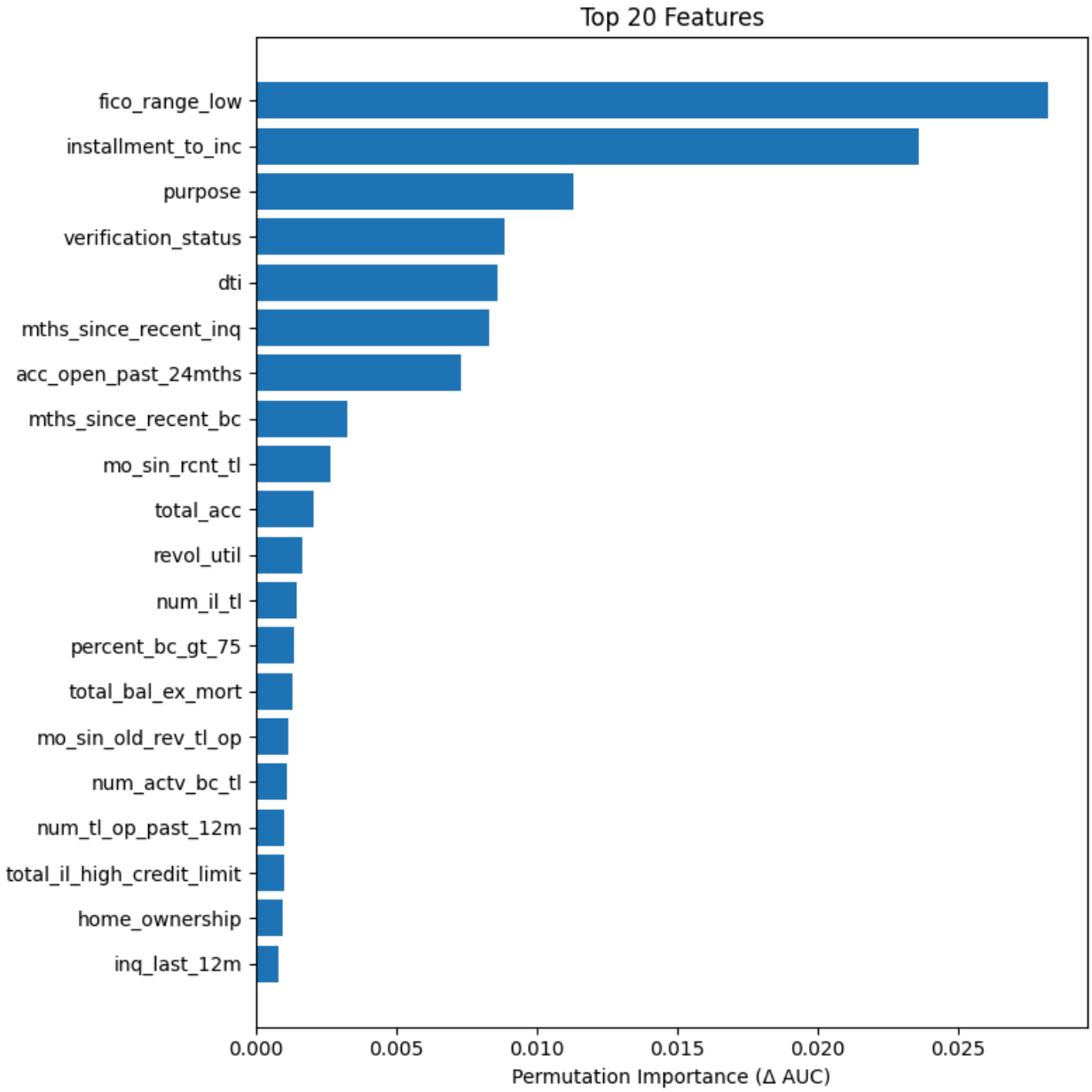
XGBoost Model:
The XGBoost model showed a ROC AUC value of 0.687 for the test or ‘hold out’ data sample, and a mean ROC AUC value of 0.676 for the cross validation sets. This shows that the model is not overfitting the training set, because the test set AUC outperformed

the training set. The ROC plot for the hold out sample is shown below:

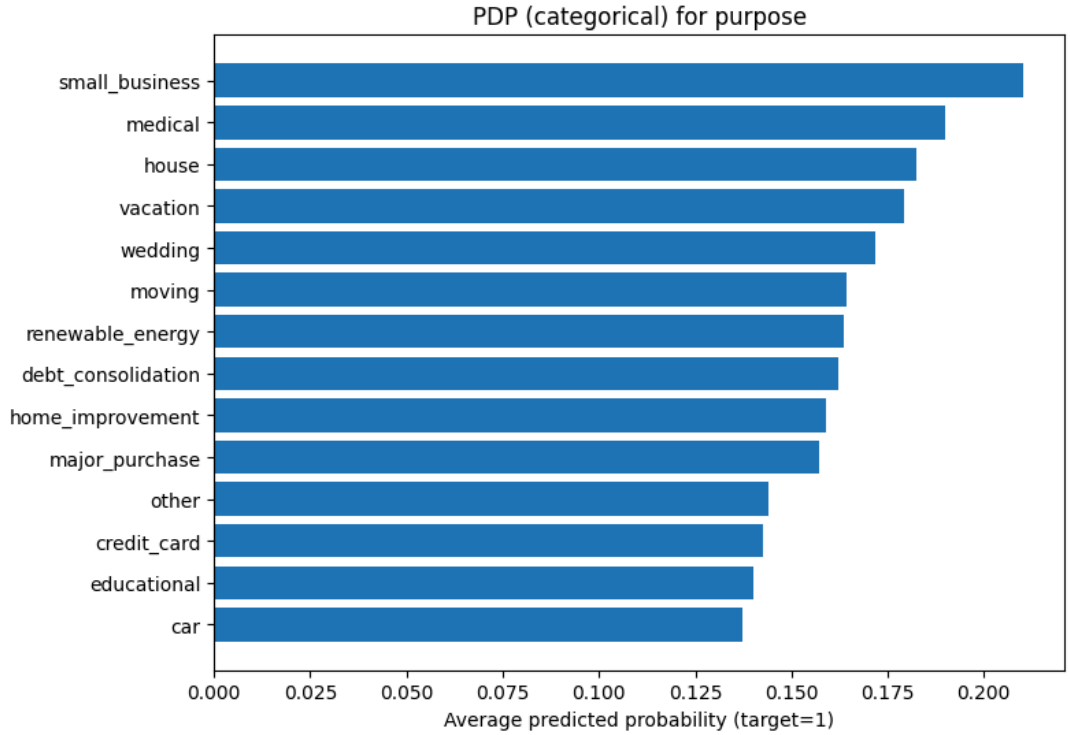


Interpretability:

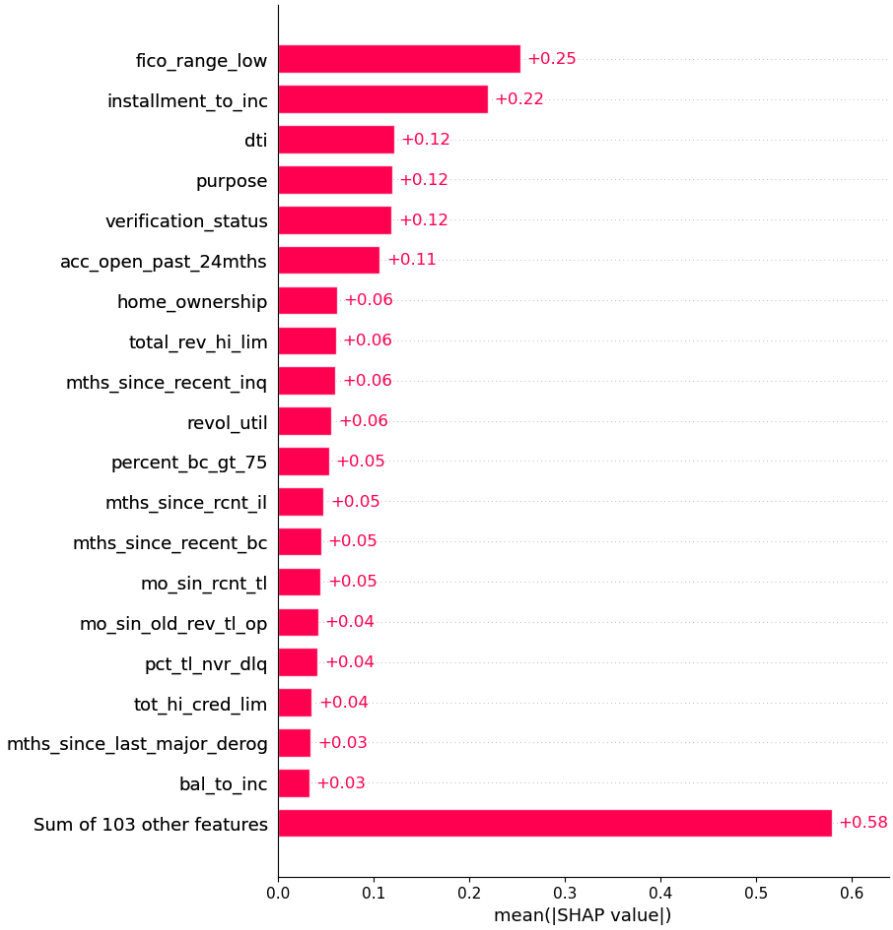
Feature Importances were evaluated using the Permutation Method. The most predictive features in this model are shown in the chart below:



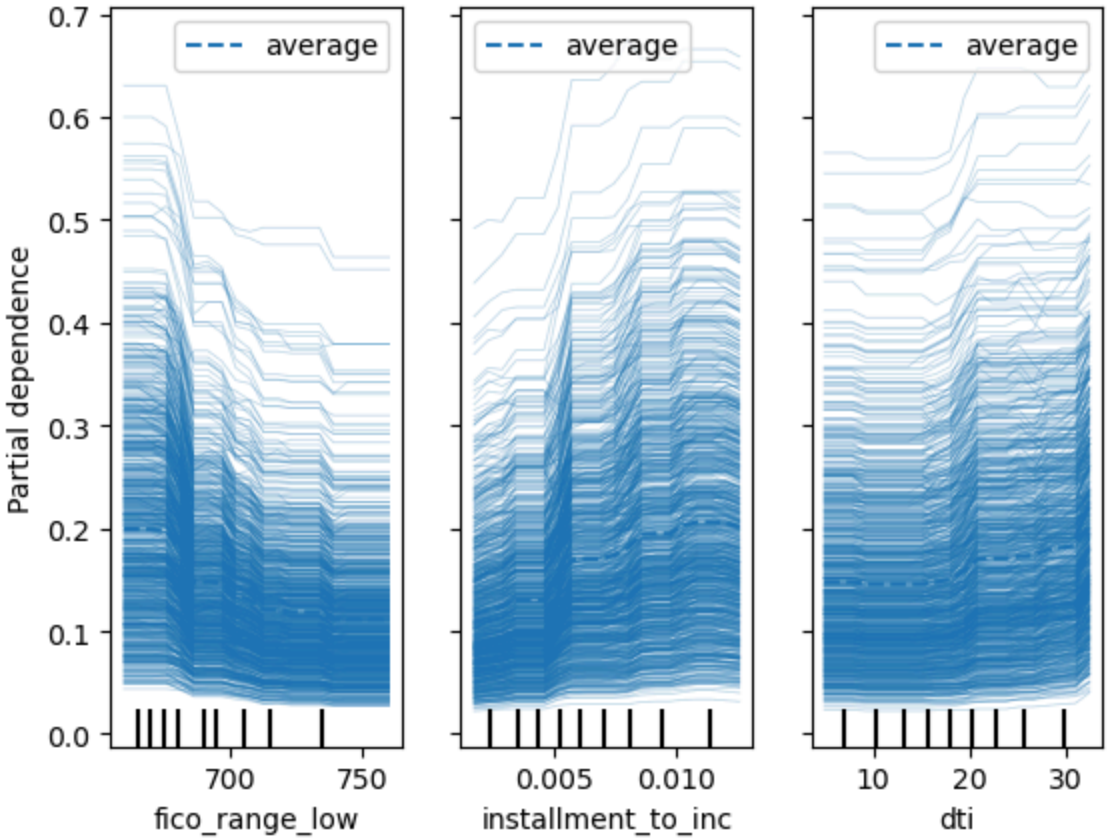
The ‘Purpose’ feature is a categorical feature with a relatively high cardinality. To get a better idea of how each value of ‘Purpose’ impacted the prediction of the model, a categorical PDP plot was created. The values of ‘Purpose’ with the highest average predictive probability are the ones that increase the probability of the loan being ‘Charged Off’.



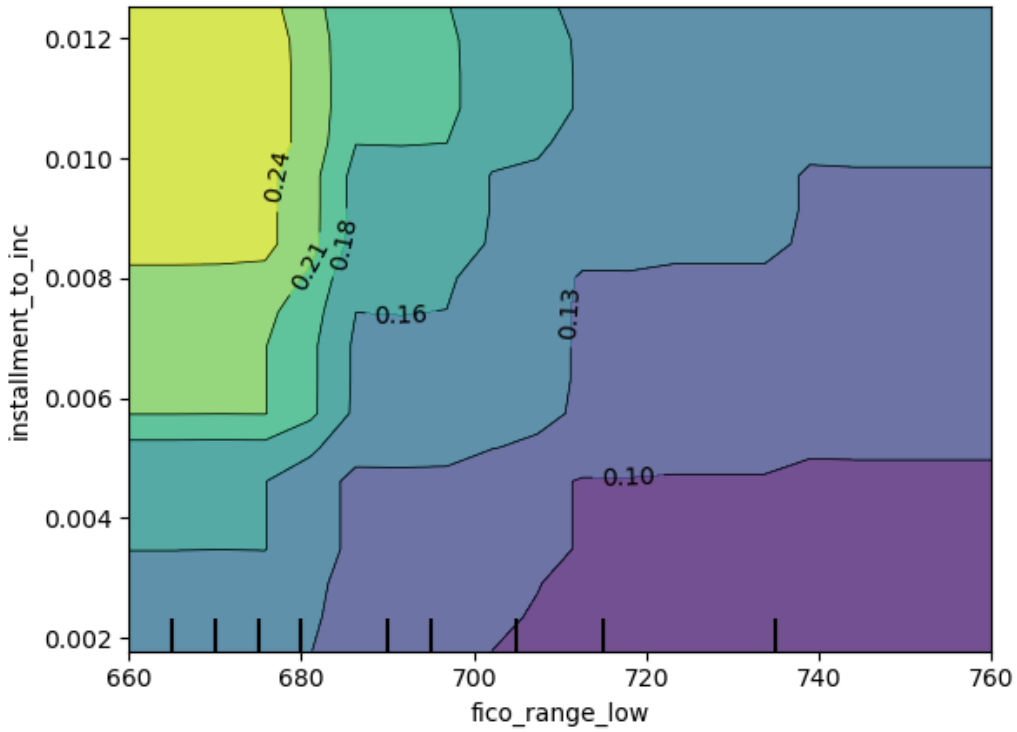
Global SHAP feature importances were also calculated as an alternative to the permutation method. These are shown in the plot below:



Here we can see that ‘Fico Range Low’, ‘Installment to Inc’ and ‘dti’ are the top 3 features in terms of mean SHAP value. To further investigate the individual impacts of the features with the highest mean SHAP value, Partial Dependence Plots for those features can be seen in the plots below:

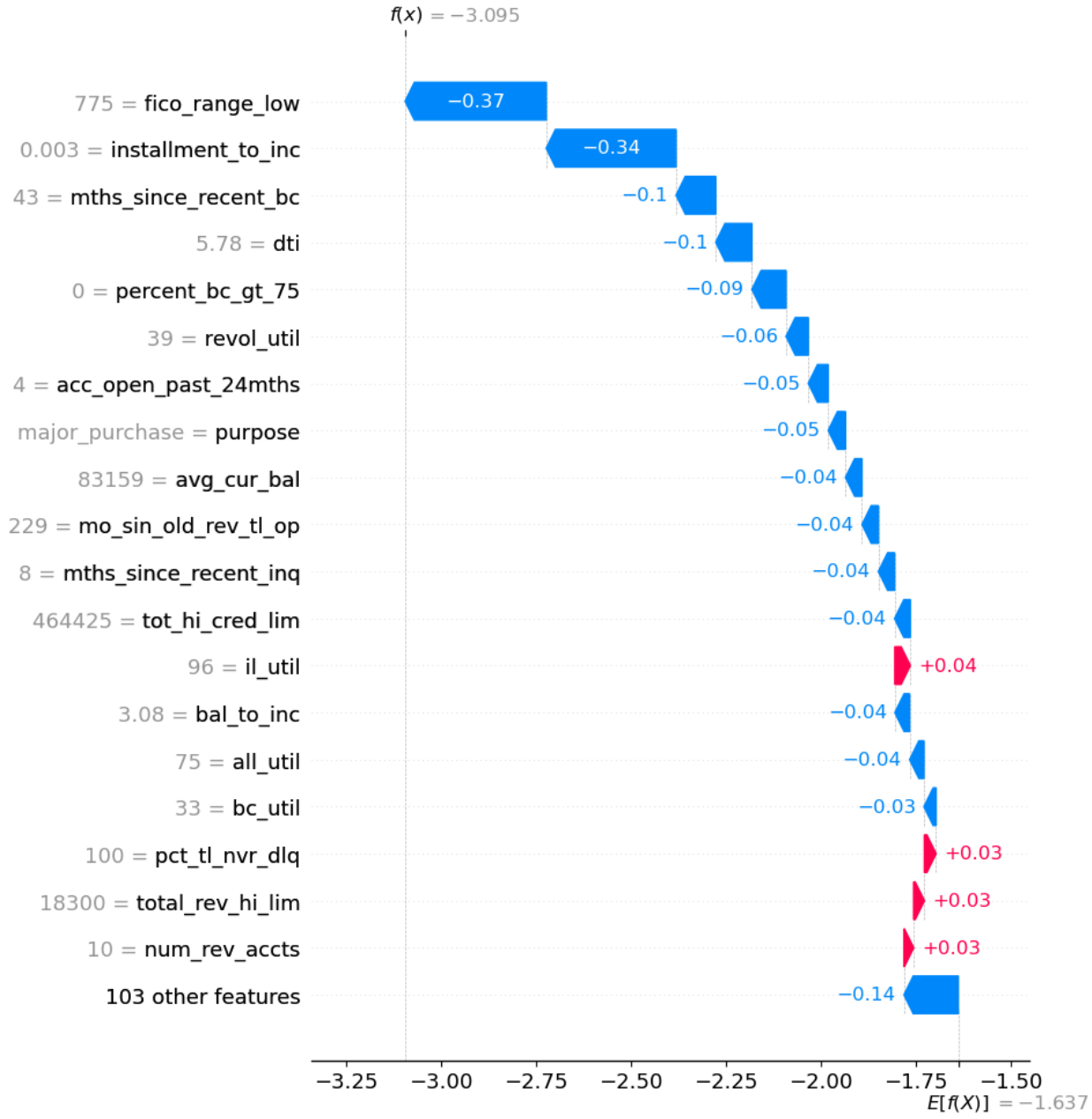


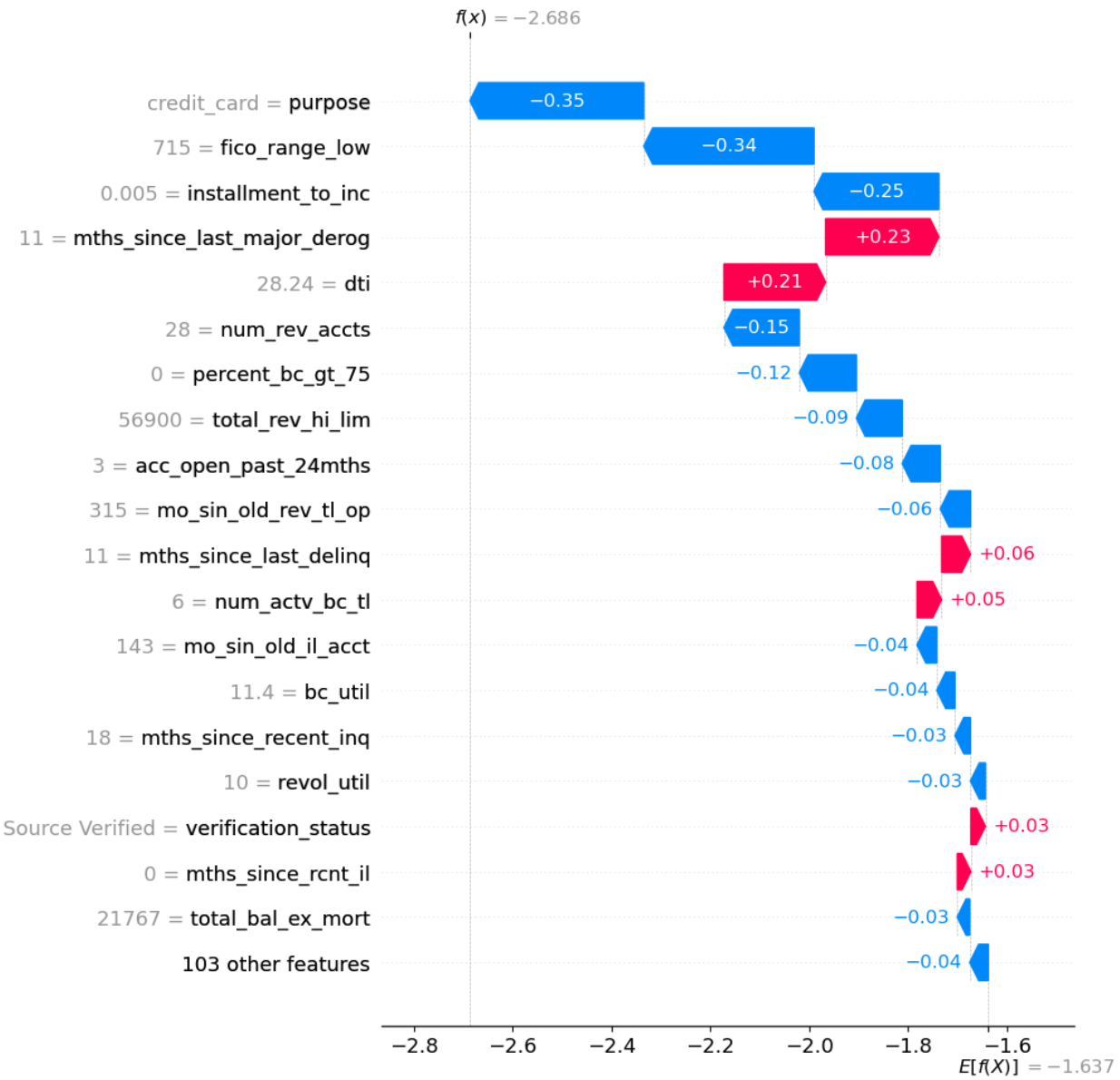
These plots show how an increasing lower bound range for fico scores decreases probability of a loan being charged off, whereas increasing the installment to income and debt to income ratios has an upward trend on the probability of default. This makes logical sense. To further investigate the relationship between variables, and how their interactions impacted the model's classification, a 2 way partial dependence plot for 'fico_range_low' and 'installment_to_inc' is shown below.

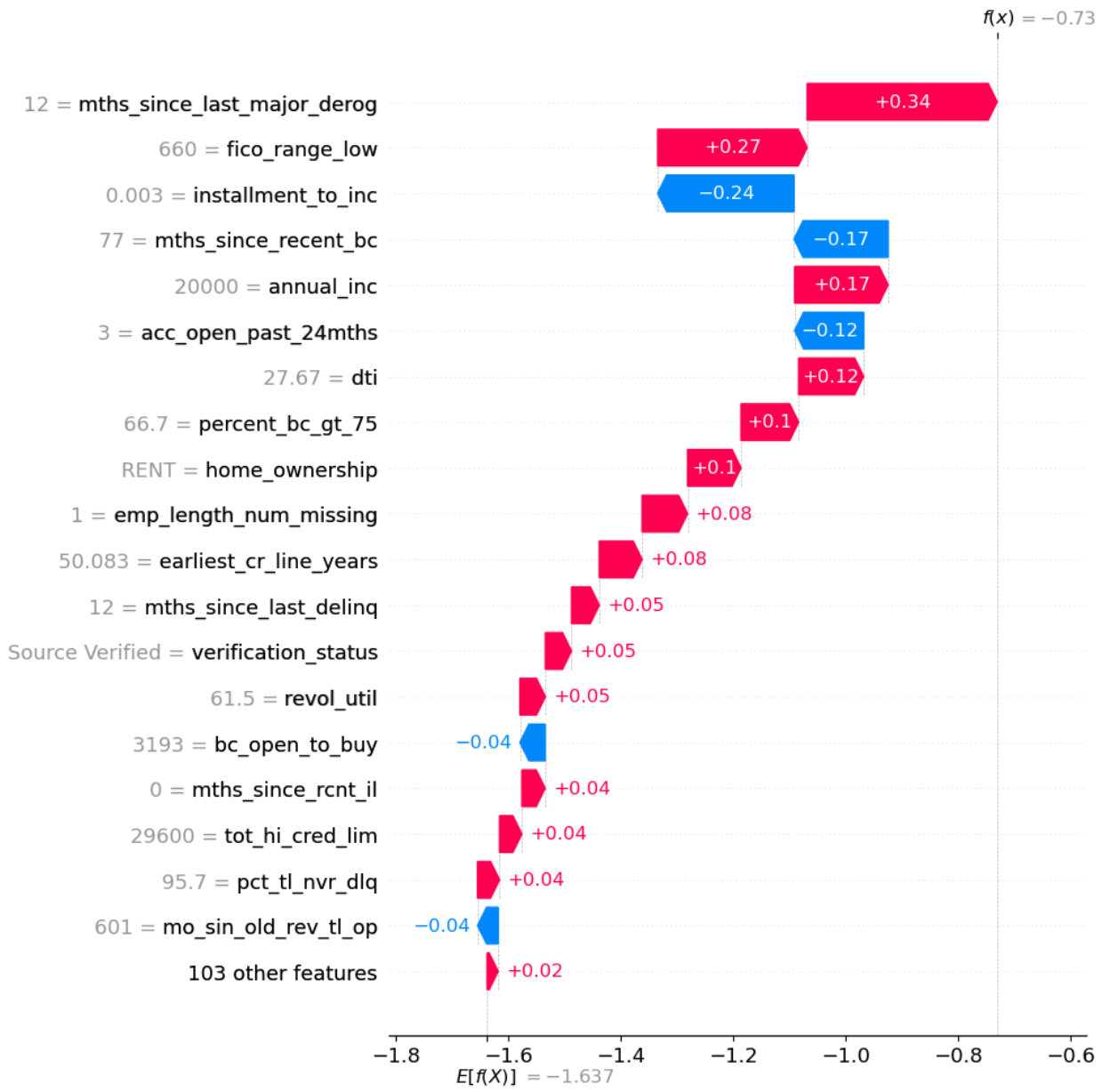


This 2 Way PDP plot illustrates the interaction between these two features. The relative effects of these two features have a compounding effect on one another. Observations with low fico ranges have an even greater chance of default when paired with high installment to income.

To explore the local explainability of observations for this model, three observations were picked at random, and their SHAP waterfall plots are shown below. These plots illustrate the local effect of each feature on the overall probability of default for the observation.


f(x) = −3.095
775 = fico_range_low
0.003 = installment_to_inc
43 = mths_since_recent_bc
5.78 = dti
0 = percent_bc_gt_75
39 = revol_util
4 = acc_open_past_24mths
major_purchase = purpose
83159 = avg_cur_bal
229 = mo_sin_old_rev_tl_op
8 = mths_since_recent_inq
464425 = tot_hi_cred_lim
96 = il_util
3.08 = bal_to_inc
75 = all_util
33 = bc_util
100 = pct_tl_nvr_dlq
18300 = total_rev_hi_lim
10 = num_rev_accts
103 other features
−0.37
−0.34
−0.1
−0.1
−0.09
−0.06
−0.05
−0.05
−0.04
−0.04
−0.04
−0.04
+0.04
−0.04
−0.04
−0.03
+0.03
+0.03
+0.03
−0.14
−3.25
−3.00
−2.75
−2.50
−2.25
−2.00
−1.75
−1.50
E[f(X)] = −1.637

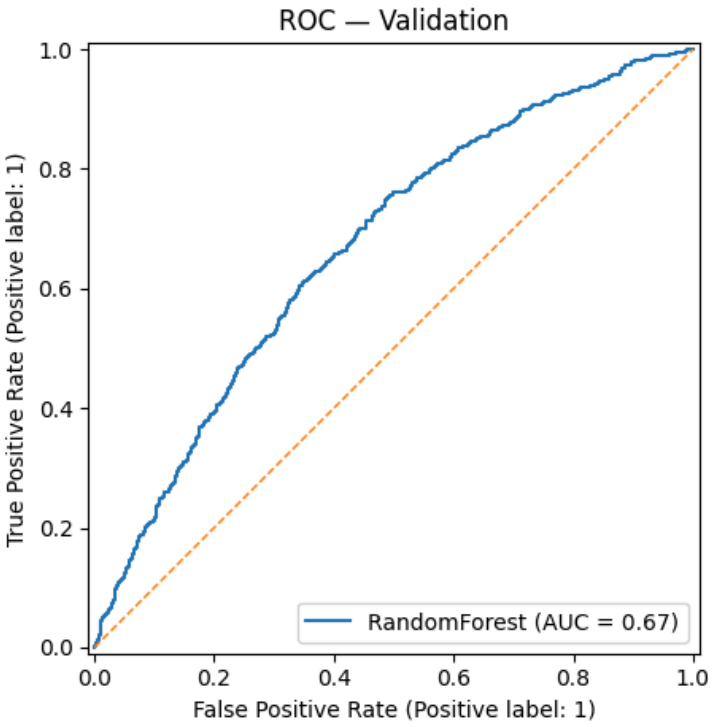

f(x) = −2.686
credit_card = purpose
715 = fico_range_low
0.005 = installment_to_inc
11 = mths_since_last_major_derog
28.24 = dti
28 = num_rev_accts
0 = percent_bc_gt_75
56900 = total_rev_hi_lim
3 = acc_open_past_24mths
315 = mo_sin_old_rev_tl_op
11 = mths_since_last_delinq
6 = num_actv_bc_tl
143 = mo_sin_old_il_acct
11.4 = bc_util
18 = mths_since_recent_inq
10 = revol_util
Source Verified = verification_status
0 = mths_since_rcnt_il
21767 = total_bal_ex_mort
103 other features
−0.35
−0.34
−0.25
+0.23
+0.21
−0.15
−0.12
−0.09
−0.08
−0.06
+0.06
+0.05
−0.04
−0.04
−0.03
−0.03
+0.03
+0.03
−0.03
−0.04
−2.8
−2.6
−2.4
−2.2
−2.0
−1.8
−1.6
E[f(X)] = −1.637



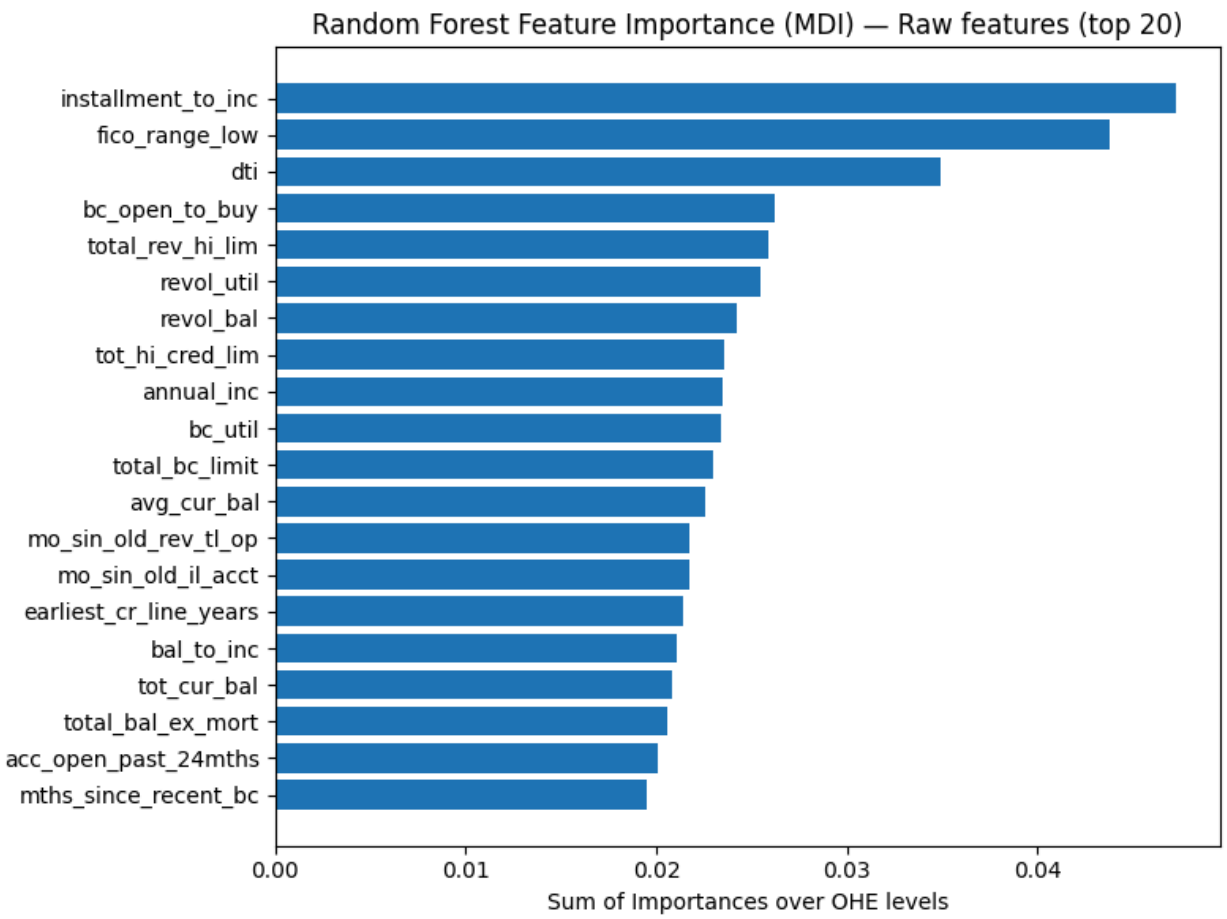
Here we can see that not every observation is impacted in the same way by the same features. For example, in all 3 plots, the lower bound of the fico range has one of the highest importances of any feature, but in the third plot, the ‘mths_since_last_major_derog’ variable has the greatest impact of this observation's predicted probability of default.

Random Forest Benchmark Model

The Random Forest model used for benchmarking showed a mean ROC AUC value of 0.662 for the cross validation sets. The ROC curve is shown below:



The feature importances for the raw features used in the Random Forest model are shown below. This chart reveals that the random forest model relied heavily on the same top 3 features as the XGBoost model, but the features beyond the top 3 deviate. The random forest model relies less on the categorical features which showed high predictive power in the XGBoost model, such a ‘Purpose’ and ‘Verification Status’.



Risk Controls and Limitations:

- Limitations and use cases: This model should be used to help guide risk based decision making in the funding of loans through Lending Club. Because the data is platform specific, it is not certain that this model would have any predictive value for loans issued outside of Lending Club. This model also relied on a limited subset of years in which various macro-economic factors were present. It is not certain that the model would have the same predictive power for loans issued outside of the date range of observations in this model. This model does not take into account any of the underwriting decisions used to structure the loans. This should be used alongside the individual risk assessment of the underwriting body in deciding various aspects of the loan issuance.
- Ongoing Monitoring: If this model is to be used outside of the date range of the original training data set, the model should be re-trained and re-evaluated on a regular basis. Quarterly evaluation should seem sufficient for re-training, so as to identify potential drift in the data or outcomes of the model.