1. Appendix

1.1. Feature description and selection

Recall that we end up with 33,951 completed loans with 38 features and 1 response variable. The explanation of these 38 features and the response variable are shown in Table 1.

Table 1: Features and response variable description

Feature	Explanation	Type
BorrowerRate	The Borrower's interest rate for this loan.	Numerical
LenderYield	The Lender yield on the loan.	Numerical
OpenCreditLines	Number of open credit lines.	Numerical
EstimatedEffectiveYield	Effective yield that estimated by Prosper.	Numerical
EstimatedLoss	Estimated principal loss on charge-offs.	Numerical
ProsperRating	A custom rating score built by Prosper.	Numerical
ProsperScore	A custom risk score built by Prosper.	Numerical
ListingCategory	The category of the listing that the borrower.	Numerical
CurrentCreditLines	Number of current credit lines.	Numerical
TotalCreditLinespast7years	Number of credit lines in the past seven years.	Numerical
OpenRevolvingAccounts	Number of open revolving accounts.	Numerical
OpenRevolvingMonthlyPayment	Monthly payment on revolving accounts.	Numerical
TotalInquiries	Total number of inquiries.	Numerical
CurrentDelinquencies	Number of accounts delinquent.	Numerical
AmountDelinquent	Dollars delinquent.	Numerical
Occupation	The Occupation selected by the Borrower.	Categorical
PublicRecordsLast10Years	Number of public records in the past 10 years.	Numerical
RevolvingCreditBalance	Dollars of revolving credit.	Numerical
TradesNeverDelinquent	Trades that have never been delinquent.	Numerical
TotalTrades	Number of trade lines ever opened.	Numerical
StatedMonthlyIncome	The monthly income the borrower stated.	Numerical
AvailableBankcardCredit	The total available credit via bank card.	Numerical
MonthlyLoanPayment	The scheduled monthly loan payment.	Numerical
TradesOpenedLast6Months	Number of trades opened in the last 6 months.	Numerical
BankcardUtilization	The percentage of available credit that is utilized.	Numerical
IsBorrowerHomeowner	Specifies if the borrower is a homeowner or not.	Categorical
BorrowerAPR	The Borrower's Annual Percentage Rate.	Numerical
DebtToIncomeRatio	The debt to income ratio of the borrower.	Numerical
EstimatedReturn	Return that is estimated by Prosper.	Numerical
InquiriesLast6Months	Number of inquiries in the past six months.	Numerical
LoanOriginalAmount	The origination amount of the loan.	Numerical
CreditScoreRangeLower	The lower range of the borrower's credit score.	Numerical
EmploymentStatusDuration	The length in months of the employment status.	Numerical
DelinquenciesLast7Years	Number of delinquencies in the past 7 years.	Numerical
Term	The length of the loan expressed in months.	Numerical
BorrowerState	The state of the address of the borrower.	Categorical
EmploymentStatus	The employment status of the borrower.	Numerical
CurrentlyInGroup	Specifies if or not the Borrower was in a group.	Categorical
LoanStatus	The current status of the loan.	Categorical

The results of overall accuracy for three-label classification with forward selection, backward selection and recursive section are shown in Table 2. From Table 2, we can observe that under the same classification technique, the recursive selection has the highest overall accuracy. Hence we can state that the recursive selection outperform the forward and backward selection based on our sample dataset.

Table 2: Single Monte Carlo result for overall accuracy of three-label classification (in %) with different feature selection techniques on test datasets.

Technique \ Feature selection	Forward selection	Backward selection	Recursive selection
Multinomial Logistic Regression (MLOGIT)	61	60	62
Random Forest (RF)	65	65	67
Hierarchical Three Label Classification (H3LC)_RF	78	77	80
H3LC_Logistic Regression (LOGIT)	65	66	68

Recall that we use the recursive selection algorithm to filtrate out the irrelevant features for each classification method. The selected features are described as the following.

Selected features in two-label classification:

- *RF*: 'BorrowerRate', 'CurrentlyInGroup', 'EstimatedLoss', 'Term', 'EstimatedEffectiveYield', 'ProsperRating', 'BorrowerAPR', 'ListingCategory', 'StatedMonthlyIncome', 'BankcardUtilization', 'CreditScoreRangeLower', 'TotalTrades', 'ProsperScore', 'TotalInquiries', 'CurrentCreditLines', 'TradesNeverDelinquent', and 'IsBorrowerHomeowner'.

- *LOGIT*: 'EstimatedEffectiveYield', 'ProsperRating', 1InquiriesLast6Months', 'DebtToIncomeRatio', 'TradesNeverDelinquent', 'Term', 'EstimatedReturn', 'OpenRevolvingAccounts', 'TotalCreditLinespast7years', 'CurrentCreditLines', 'MonthlyLoanPayment', 'PublicRecordsLast10Years', 'CurrentDelinquencies', 'CurrentlyInGroup', 'BankcardUtilization', 'IsBorrowerHomeowner', 'BorrowerAPR', 'OpenRevolvingMonthlyPayment', and 'AmountDelinquent'.

Selected features in three-label classification:

- *RF*: 'BorrowerRate', 'EstimatedLoss', 'Term', 'BorrowerAPR', 'AmountDelinquent', 'InquiriesLast6Months', 'ProsperScore', 'TradesOpenedLast6Months', 'CurrentDelinquencies', 'EstimatedReturn', 'EstimatedEffectiveYield', 'CurrentlyInGroup', 'DelinquenciesLast7Years', 'PublicRecordsLast10Years', and 'ListingCategory'.

- *MLOGIT*: 'EstimatedEffectiveYield', 'ProsperRating', 'Term', 'ProsperScore', 'TradesNeverDelinquent', 'DebtToIncomeRatio', 'DelinquenciesLast7Years', 'MonthlyLoanPayment', 'OpenRevolvingAccounts', 'BorrowerRate', 'CreditScoreRangeLower', 'AmountDelinquent', 'ListingCategory', 'CurrentDelinquencies', and 'BorrowerAPR'.

- *H3LC_RF*: step-A: use the same features as two-label RF.

- *H3LC_RF*: step-B: 'BorrowerRate', 'EstimatedLoss', 'BorrowerAPR', 'Term', 'EstimatedEffectiveYield', 'DelinquenciesLast7Years', 'CreditScoreRangeLower', 'TradesNeverDelinquent (percentage)', 'CurrentCreditLines', 'OpenCreditLines', 'MonthlyLoanPayment', 'StatedMonthlyIncome', 'AmountDelinquent', 'ProsperRating', 'EstimatedReturn', 'BankcardUtilization', 'CurrentlyInGroup', and 'OpenRevolvingMonthlyPayment'.

- *H3LC_LOGIT*: step-A: use the same features as two-label LOGIT.

- *H3LC_LOGIT*: step-B: 'EstimatedReturn', 'CurrentDelinquencies', 'OpenRevolvingMonthlyPayment', 'CurrentCreditLines', 'IsBorrowerHomeowner', 'EstimatedLoss', 'EstimatedEffectiveYield', 'Term', 'TradesOpenedLast6Months', 'TotalInquiries', 'EmploymentStatusDuration', 'BankcardUtilization', 'InquiriesLast6Months', 'DebtToIncomeRatio', 'DelinquenciesLast7Years', 'TradesNeverDelinquent', 'BorrowerRate', 'MonthlyLoanPayment', 'OpenRevolvingAccounts', and 'AmountDelinquent'.

1.2. Mahalanobis distance method

Table 3: Root mean square error between the predicted ARR and the real ARR with five Monte Carlo CV.

CV sets / Technique	1	2	3	4	5	Average
H3LC_LOGIT_EDM(L_2)	0.25	0.20	0.18	0.26	0.23	0.224
H3LC_RF_EDM(L_2)	0.25	0.20	0.17	0.26	0.23	**0.222**
H3LC_LOGIT_EDM(Mah)	0.25	0.22	0.19	0.26	0.23	0.23
H3LC_RF_EDM(Mah)	0.26	0.21	0.18	0.26	0.22	0.226

Table 3 shows the root mean square error between the predicted ARR and the real ARR. We can observe that with the same underlying classification method the proposed technique with Euclidean distance outperforms the

Table 4: Root mean square error between the predicted standard deviation and the real standard deviation with five Monte Carlo CV.

Technique \ CV sets	1	2	3	4	5	Average
H3LC_LOGIT_EDM(L_2)	0.19	0.14	0.12	0.19	0.17	**0.162**
H3LC_RF_EDM(L_2)	0.18	0.15	0.12	0.20	0.17	0.164
H3LC_LOGIT_EDM(Mah)	0.19	0.16	0.12	0.19	0.15	**0.162**
H3LC_RF_EDM(Mah)	0.21	0.18	0.13	0.21	0.17	0.18

technique with Mahalanobis distance on average. Table 4 compares the mean saquare error of standard deviation of loans for the proposed technique with Euclidean distance and the technique with Mahalanobis distance. On average, the lowest value of mean square error of standard deviation is achieved by using H3LC_LOGIT_EDM(L_2) and H3LC_LOGIT_EDM(Mah). In fact, with the same underlying classification method, our proposed techniques with Euclidean distance show better result than the technique along with Mahalanobis distance. In particular, the results in Table 3 and Table 4, together, show that Euclidean distance is sufficient to compute the similarity between two different loans, and Euclidean distance is actually better than Mahalanobis distance.

Table 5: Error between the predicted portfolio ARR and the real portfolio ARR with five Monte Carlo CV

Technique \ CV test	1	2	3	4	5	Average
H3LC_LOGIT_EDM(L_2)	-0.016	-0.023	-0.018	-0.022	-0.007	**-0.0172**
H3LC_RF_EDM(L_2)	0.032	0.039	0.030	0.033	0.014	0.0296
H3LC_LOGIT_EDM(Mah)	-0.036	-0.037	-0.022	-0.028	-0.008	-0.0262
H3LC_RF_EDM(Mah)	0.056	0.076	0.042	0.048	0.051	0.0546

Table 6: Error between the predicted portfolio standard deviation and the real portfolio standard deviation with five Monte Carlo CV.

Technique \ CV test	1	2	3	4	5	Average
H3LC_LOGIT_EDM(L_2)	-0.031	-0.023	-0.018	-0.034	-0.027	-0.0266
H3LC_RF_EDM(L_2)	-0.031	-0.030	-0.016	-0.032	-0.025	-0.0268
H3LC_LOGIT_EDM(Mah)	-0.029	-0.024	-0.021	-0.028	-0.020	**-0.0244**
H3LC_RF_EDM(Mah)	-0.047	-0.056	-0.024	-0.048	-0.043	-0.0436

The error between the *predicted* portfolio returns and the real portfolio returns for test datasets is shown in Table 5. The closest portfolio ARR is predicted by technique H3LC_LOGIT_EDM(L_2) with absolute error 0.0172. It can be observed that the absolute errors of H3LC_RF_EDM(Mah) is much higher than that of H3LC_RF_EDM(L_2). This is due to that when predicting the ARR, H3LC_RF_EDM(Mah) usually predict ARR higher than that of H3LC_RF_EDM(L_2) for the same loan. Particularly, with the same underlying classification method, the proposed technique with Euclidean distance results in smaller absolute error than Mahalanobis distance.

Table 6 displays the error between the *predicted* portfolio standard deviation (risk) and the real portfolio standard deviation. One can observe that the smallest error with value -0.0244 is achieved by using the proposed technique with Mahalanobis distance. However, compared to the error with value -0.0266 by using H3LC_LOGIT_EDM(L_2), Mahalanobis distance does not results in significant additional performance (-0.0244 Vs -0.0266).

1.3. Efficiency frontiers

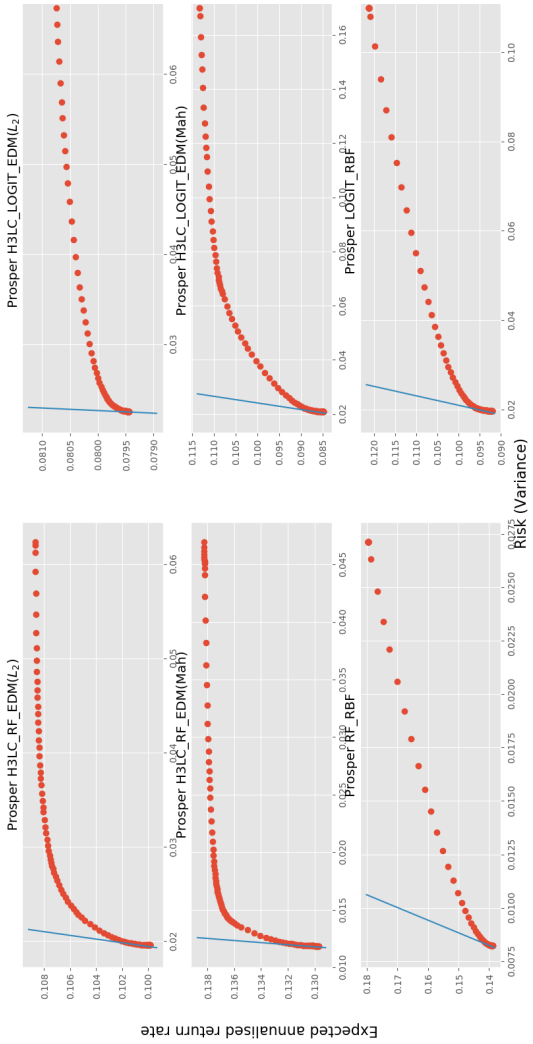

Figure 1: **The efficiency frontiers with respect to different technique on Prosper dataset** .