UCLA ANDERSON MFE APPLIED FINANCE PROJECT

WITH MOTOLEASE LLC

CREDIT ANALYSIS:

CONSUMER CREDIT DEFAULT PREDICTION USING MACHINE LEARNING

UCLA Anderson

THINK IN THE NEXT

AGENDA

1. Motivation
2. Data and Features
3. Methodology
4. Conclusion

MOTIVATION

CONSUMER CREDIT MODELING

- **Definition:**
  - Likelihood the person will default on their debt
- **Data/factors:**
  - Credit activity
    - Credit mix, credit balances, payment patterns
    - Bankruptcy filings, collection items
    - Credit inquiries
  - Financial personal information (e.g. income)
  - Non-financial personal information (e.g. education level)
- **Usage:**
  - Mortgages
  - Credit cards, installment loans, etc.
  - Automobile and other vehicle financing

TYPICAL MEASUREMENT – CREDIT SCORE

- **Definition:**
  - A numerical value based on a person's credit files' analysis in order to represent the creditworthiness of an individual

- **Types:**
  - **FICO:** 300 – 850
  - **Experian:** 330 – 830
  - **Equifax:** 300 – 850
  - **TransUnion:** 300 – 850

UCLA Anderson
THINK IN THE NEXT

EXAMPLE – FICO SCORE

- Calculated based on consumer credit files of Experian, Equifax, and TransUnion
- Used by the major banks and credit grantors
- General components:
  - 35%: payment history
  - 30%: debt burden
  - 15%: length of credit history
  - 10%: types of credit used
  - 10%: hard credit

EXAMPLE – FICO SCORE

UCLA Anderson
THINK IN THE NEXT

THE CLIENT

- **MotoLease LLC** originates and services motorcycle leases
- Significant investments into software-based credit decision systems
- Ability to make most credit decisions in 60 seconds
- 90% approval rate
- Leases offered through a network of dealers

MOTIVATION

- Typical consumer credit score models are not specialized for motorcycle financing industry
- Certain factors not considered for credit scores may be useful for predicting defaults on motorcycle leases

MAIN OBJECTIVE

- Derive and identify **2 to 5 new features** from MotoLease's data to add to their current existing model
- Ensure our model is robust to missing data
- Develop and enhance machine learning model to improve accuracy of default prediction

UCLAAnderson
THINK IN THE NEXT

DATA AND FEATURES

DATA DOMAIN

- Data on 16,000 lease applicants
- Leases were funded between March 2015 and September 2017
- Default status is determined one year after the lease was funded
  - 42% default rate
- Default status is binary and defined by MotoLease
  - Recovery rates are out of scope
- Each lease has a unique identifier for mapping to input datasets

UCLA Anderson
THINK IN THE NEXT

DATA STRUCTURE

- Output dataset
  - Default vs. non-default leases (16,000 x 9)

- Input datasets
  - Credit scores (included in Default vs. non-default leases)
  - Tradelines (210,000 x 33)
  - Inquiries (140,000 x 12)
  - Collections (65,000 x 28)
  - Credit summary (18,000 x 42)

UCLA Anderson
THINK IN THE NEXT

AN ILLUSTRATION OF OUR DATA STRUCTURE

» **Input data:**

Tradelines (shown below), Inquiries, Public Records, Collections, Credit Summary

Id	Account type	Utilization ratio
5123	Credit card	34%
5123	Credit card	12%
5123	Mortgage	N/A
5123	Auto loan	N/A
651	Mortgage	N/A
651	Mortgage	N/A
52	Credit card	60%

» **Output data:**

Default vs. Non-Default Leases

Id	Defaulted
5123	0
651	0
52	1
99	1

CHALLENGES RELATED TO DATA STRUCTURE

- Data is not provided in a flat format
  - I.e. one table with $Y, X_1, X_2, X_3, \ldots X_N$
  - Varying amounts of data are available for different applicants
- Categorical variables have hundreds of different values
  - E.g. lender name, account type, creditor industry
- There are interactions between variables
  - E.g. credit card balance has other coefficient than mortgage balance

UCLA Anderson
THINK IN THE NEXT

OUR APPROACH TO FEATURE ENGINEERING

- Create as many features as possible, take care of overfitting in a separate step
- Brainstorm within our team and research credit agency methodologies
- For categorical variables, create subtotals and counts for the most common values only
  - Include count(credit cards)
  - Exclude count('401k Loan Repayment')

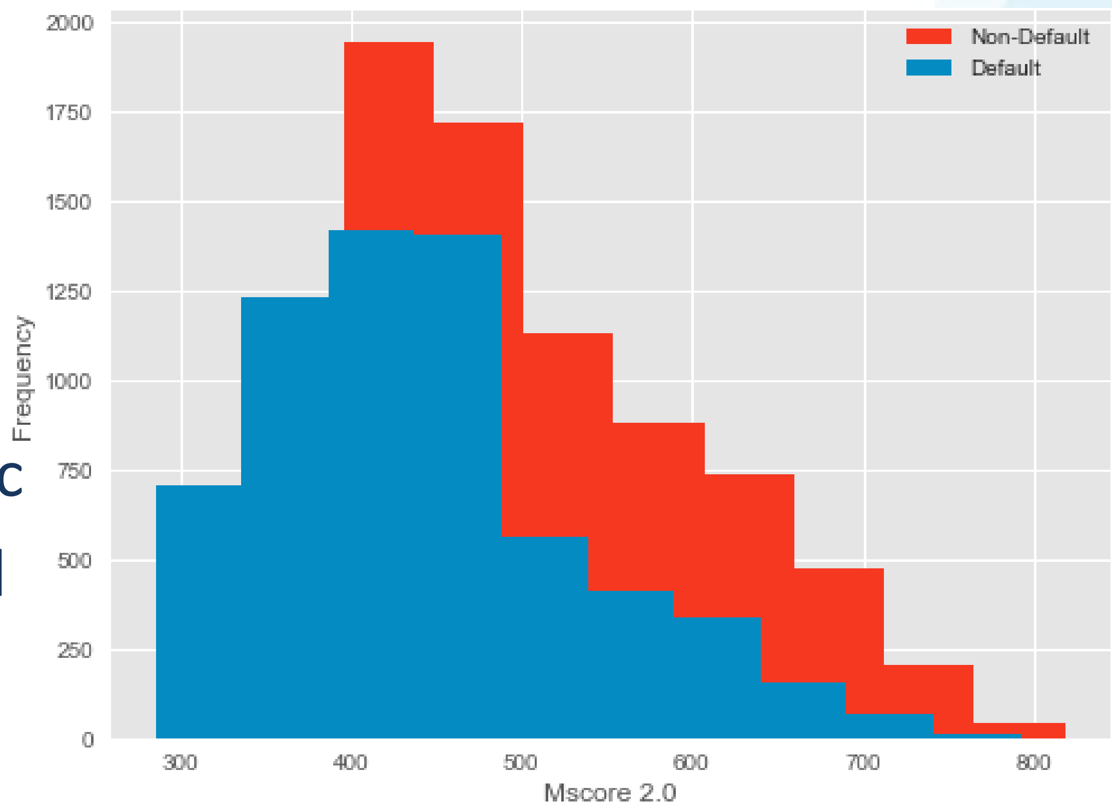
UNIVERSE OF FEATURES

# of features	% N/A	Type of features
9	0	External credit scores, payment-to-income, debt-to-income
2	0	Counts of collection items, separated by collection item status
6	0	Counts of credit inquiries, separated by type of credit
8	0	Sums/means/maxima of past due amounts, balances, limits, monthly rates, high credit
5	0	Ratios of balance sums/means/maxima to limits, high credit
18	0	Counts of tradelines, separated by tradeline status, lender types, account types, loan types, ownership types
2	0	Features on frequency and range of tradeline open dates
10	0	Payment pattern
23	4	Various utilization ratios, from alternative data source
12	4	Various counts of tradelines, public records, inquiries, from alternative data source
95	**4**	**Total**

METHODOLOGY

KOLMOGOROV-SMIRNOV (K-S) STATISTIC

- K-S test
  - determine if two data-sets differ significantly
  - Non-parametric
  - KS statistic and P value

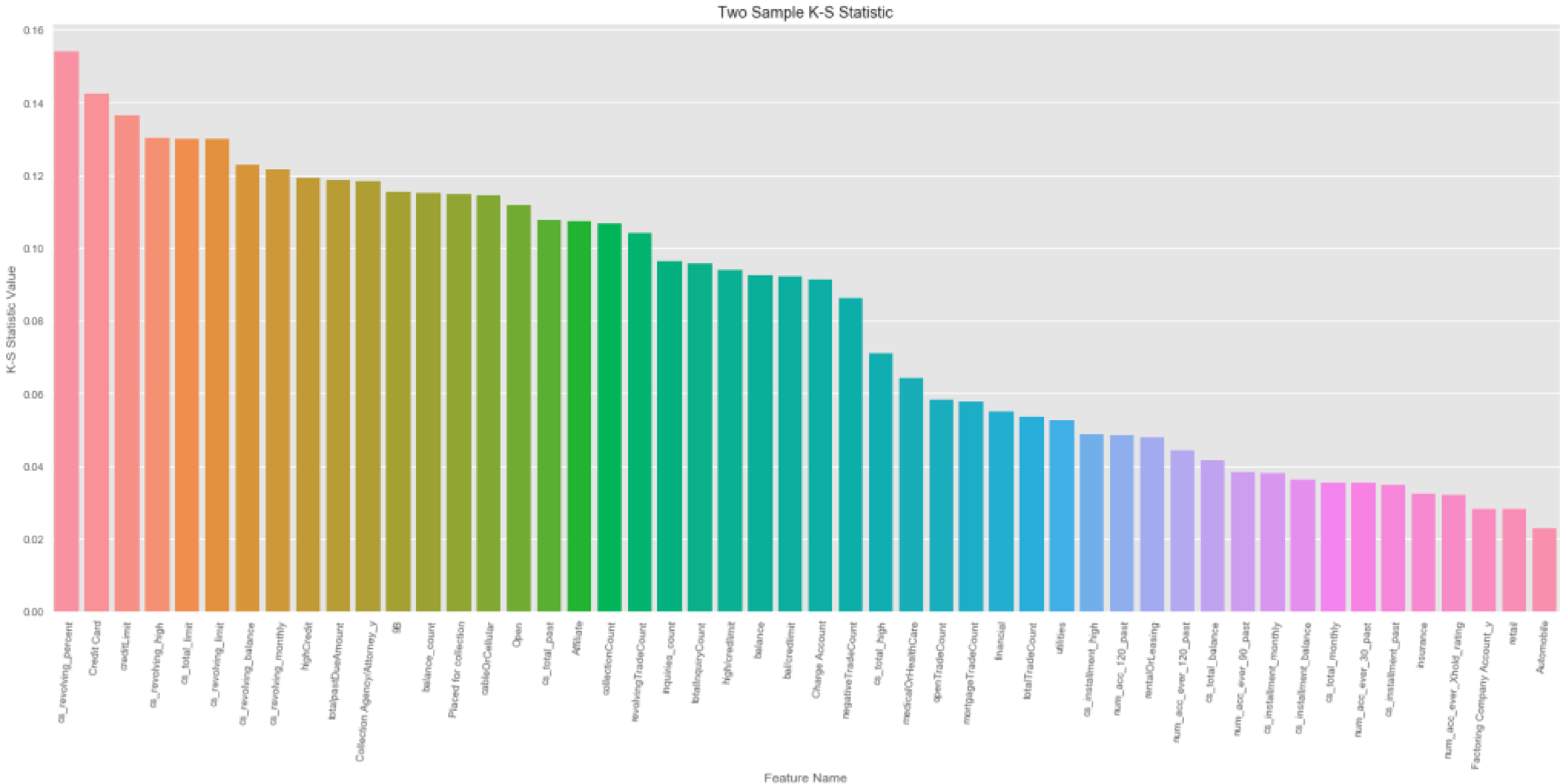


UCLAAnderson
THINK IN THE NEXT

FEATURE SELECTION

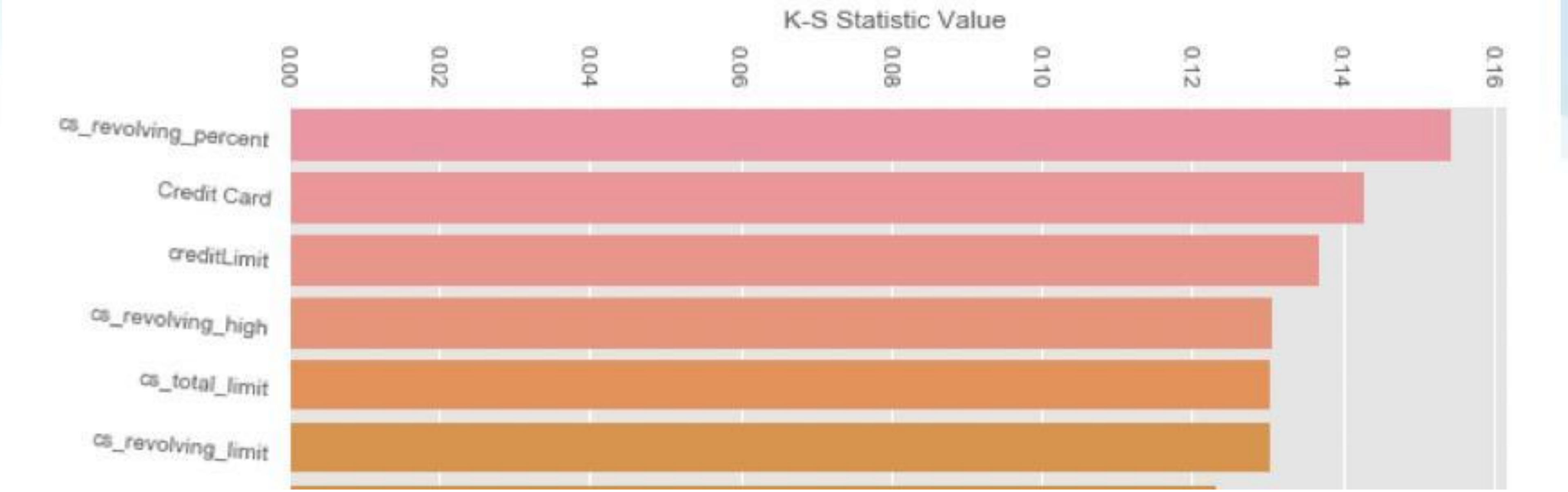
- Sort on KS statistic
- Number of tree splits

TOP FEATURES BY K-S



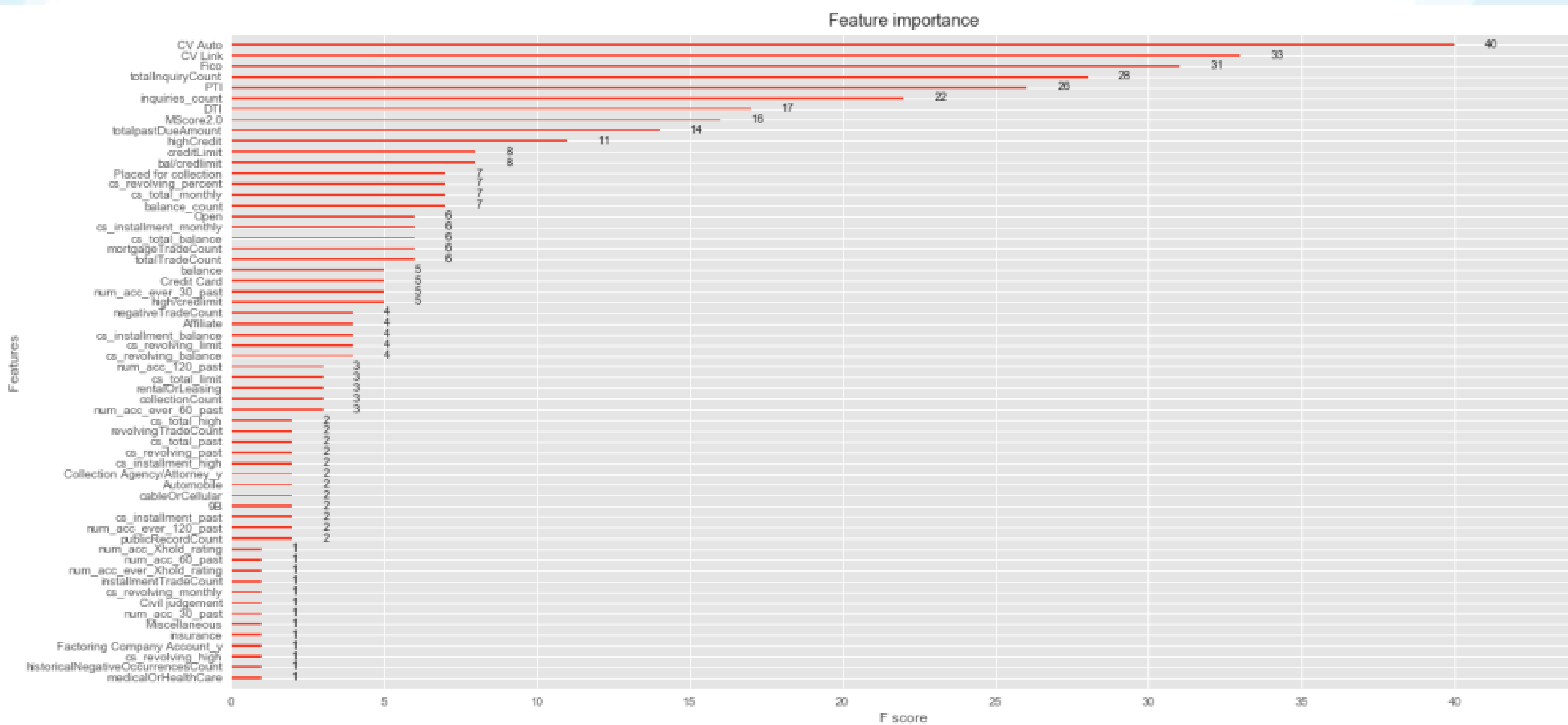
UCLA Anderson
THINK IN THE NEXT

TOP FEATURES BY K-S



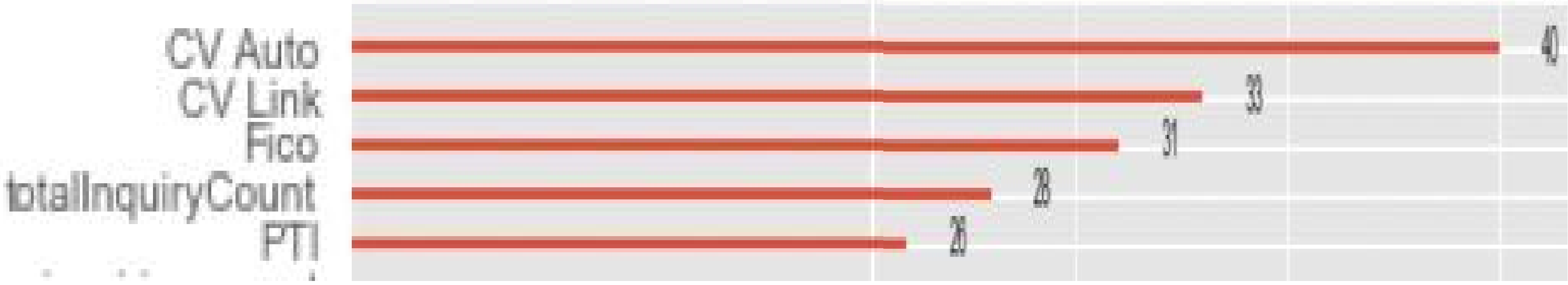
UCLA Anderson
THINK IN THE NEXT

TOP FEATURES BY TREE SPLITS



UCLA Anderson
THINK IN THE NEXT

TOP FEATURES BY TREE SPLITS



UCLA Anderson
THINK IN THE NEXT

FEATURE SELECTION FOR MSCORE

- **Goal**
  - Identify 2-5 features to improve the MSCORE
  - Metric AUC - TPR vs FPR
- **Approach**
  - Top ~25 features ranked by K-S/tree splits
  - $\binom{25}{5}$ Logistic regressions, ~53k
  - Compared AUC of top feature combinations

UCLA Anderson
THINK IN THE NEXT

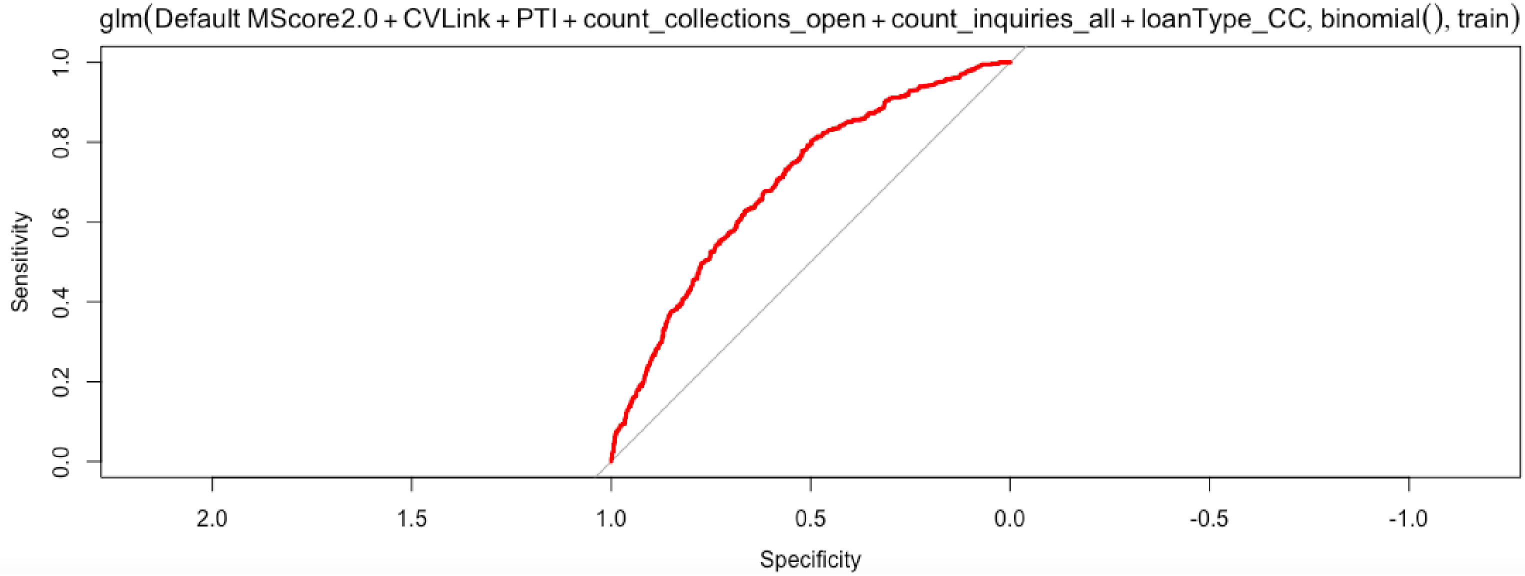
MACHINE LEARNING TECHNIQUES

- **Models**
  - Logistic Regression
  - XGBoost
- **Cross-validation**
  - Tune the number of gradient boosting rounds
- **Bayesian Optimization**
  - Tune the hyperparameters
- **Bootstrap Aggregating** (Bagging)

UCLAAnderson
THINK IN THE NEXT

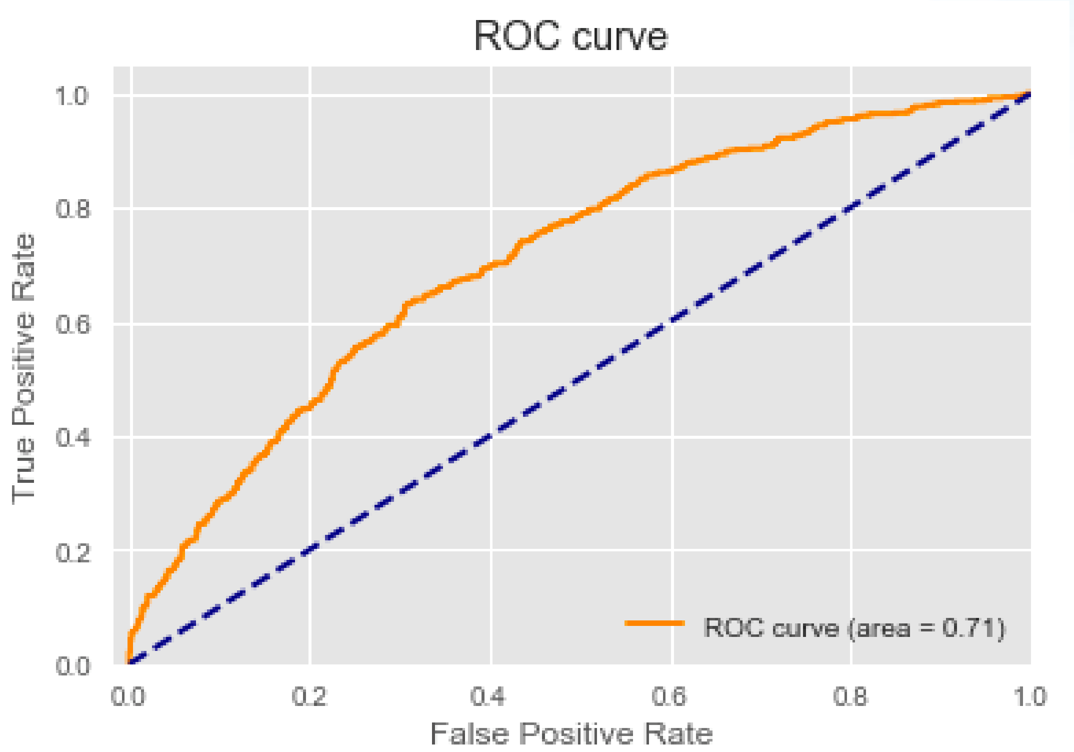
MODELS

- **Logistic regression**
  - AUC = .70



MODELS

- **XGBoost**
  - AUC = .71



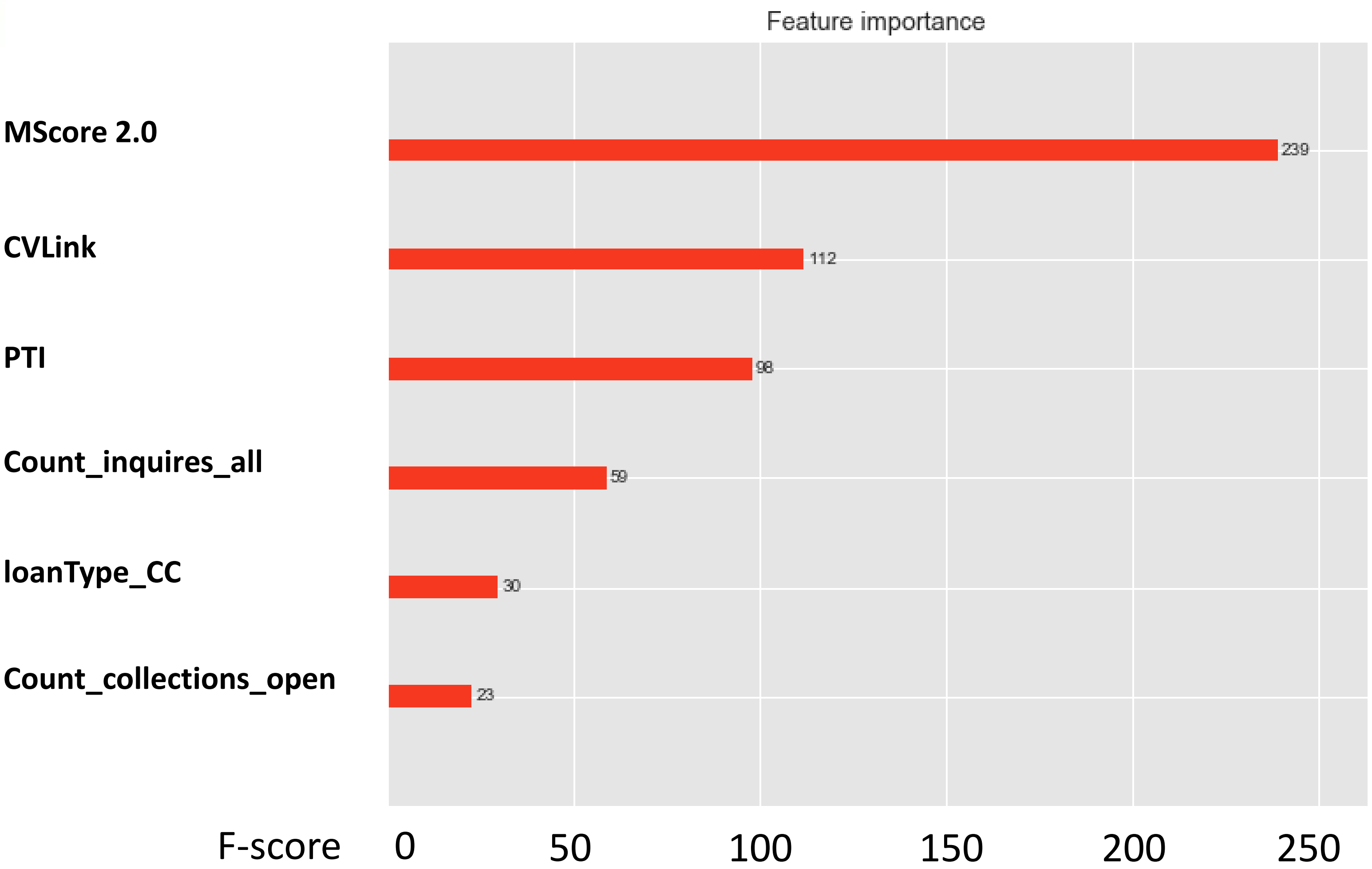
UCLA Anderson
THINK IN THE NEXT

ADDED FEATURES

- CVLINK : Credit Vision custom score based on alternative and trended data
- PTI: Payment to income ratio
- Count of inquiries
- Count of open collections
- Loan Type : No. of credit cards

UCLA Anderson
THINK IN THE NEXT

FINAL FEATURES



UCLA Anderson
THINK IN THE NEXT

RESULTS

- AUC of existing M-Score: 0.63
- AUC with added features (Logistic): 0.70
- AUC with added features (XGBoost): 0.71

CONCLUSION

INSIGNIFICANT FEATURES

- Most categorical variables (e.g. lender type)
- Payment Pattern
- Utilization Ratios: Amount of credit used

UCLA Anderson
THINK IN THE NEXT

```
Call:
glm(formula = Default ~ MScore2.0 + CVLink + PTI + count_collections_open +
    count_inquiries_all + loanType_CC, family = binomial(), data = train)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-4.800   -1.017   -0.667    1.136    2.432

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.7391279  0.1818032   9.566  < 2e-16 ***
MScore2.0             -0.0034093  0.0002281 -14.947  < 2e-16 ***
CVLink                -0.0024298  0.0003443  -7.058 1.69e-12 ***
PTI                    0.0713066  0.0041095  17.352  < 2e-16 ***
count_collections_open 0.0146020  0.0037856   3.857 0.000115 ***
count_inquiries_all    0.0314918  0.0026097  12.067  < 2e-16 ***
loanType_CC           -0.0667365  0.0085244  -7.829 4.92e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20214  on 14823  degrees of freedom
Residual deviance: 18562  on 14817  degrees of freedom
AIC: 18576

Number of Fisher Scoring iterations: 5
```

FEATURE IMPACT ON DEFAULTS

- CVLINK (Negative Impact)
- PTI (Positive Impact)
- Count of inquiries (Positive Impact)
- Count of open collections (Positive Impact)
- No. of credit cards (Negative Impact)

UCLAAnderson
THINK IN THE NEXT

THINK IN THE NEXT