Prepared by Phu Dang for Professor Sun

WI24 Week 8 Update

ULI H2H - LIHTC Data/ML Analysis

Model (Negative Binomial)
n_units ~ yr_alloc + bond + credit + allocamt + dda + basis + fmha_514 + fmha_515 + fmha_538

The model applies the logarithm to the predictor in modeling the target variable using the iterative reweighted least squares optimization method

Highlights:

Bond and fmha_515 seem informative/influential as predictors for the number of units

Fmha_515 ~ FmHA (RHS) section 515 loan

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                n_units   No. Observations:                  733
Model:                            GLM   Df Residuals:                      725
Model Family:        NegativeBinomial   Df Model:                            7
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3730.7
Date:                Sat, 02 Mar 2024   Deviance:                       387.73
Time:                        06:11:04   Pearson chi2:                     412.
No. Iterations:                    11   Pseudo R-squ. (CS):             0.1040
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
yr_alloc       0.0021   9.46e-05     21.755      0.000       0.002       0.002
bond           0.9303      0.128      7.275      0.000       0.680       1.181
credit        -0.1636      0.191     -0.855      0.393      -0.539       0.211
allocamt    1.806e-08   1.13e-08      1.603      0.109   -4.02e-09    4.01e-08
dda           -0.0718      0.117     -0.615      0.539      -0.301       0.157
basis          0.0713      0.083      0.861      0.389      -0.091       0.234
fmha_514   -2.016e-17   3.63e-18     -5.548      0.000   -2.73e-17    -1.3e-17
fmha_515      -0.2475      0.135     -1.836      0.066      -0.512       0.017
fmha_538      -0.5193      0.348     -1.493      0.136      -1.201       0.163
==============================================================================
```

Model (Negative Binomial) - (Same model for low-income units)
li_units ~ yr_alloc + bond + credit + allocamt + dda + basis + fmha_514 + fmha_515 + fmha_538

The model applies the logarithm to the predictor in modeling the target variable using the iterative reweighted least squares optimization method

Highlights:

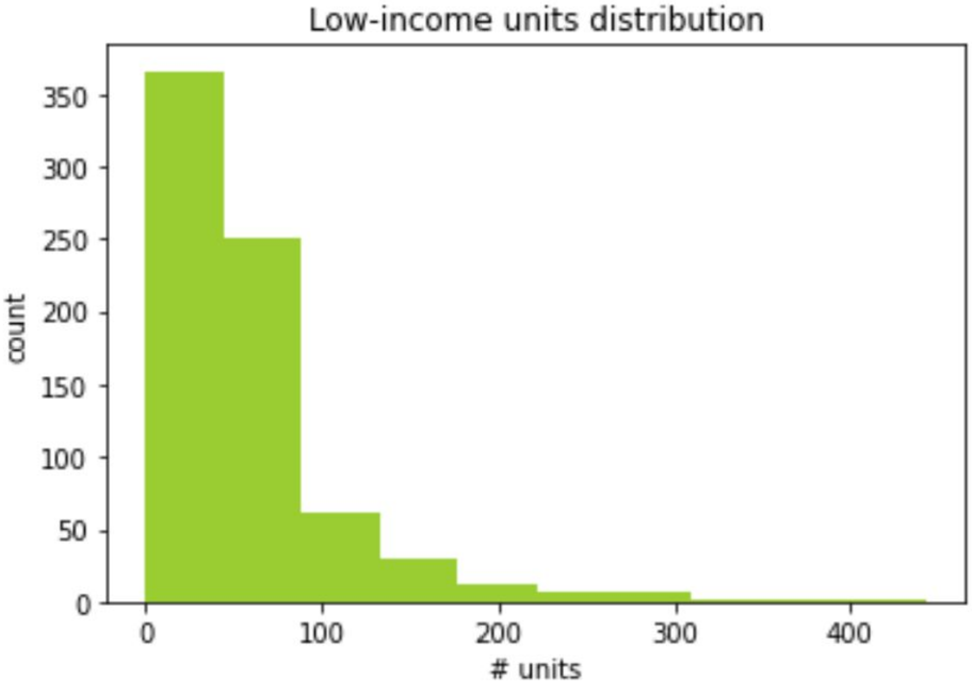
Bond and fmha_515 seem informative/influential as predictors for the number of low-income units

Fmha_515 ~ FmHA (RHS) section 515 loan

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:               li_units   No. Observations:                  733
Model:                            GLM   Df Residuals:                      725
Model Family:        NegativeBinomial   Df Model:                            7
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3662.3
Date:                Sat, 02 Mar 2024   Deviance:                       591.01
Time:                        06:11:04   Pearson chi2:                     436.
No. Iterations:                    10   Pseudo R-squ. (CS):             0.1164
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
yr_alloc       0.0020   9.46e-05     21.596      0.000       0.002       0.002
bond           0.9741      0.128      7.615      0.000       0.723       1.225
credit        -0.2411      0.191     -1.260      0.208      -0.616       0.134
allocamt    1.935e-08   1.13e-08      1.716      0.086   -2.74e-09    4.14e-08
dda           -0.0872      0.117     -0.746      0.456      -0.316       0.142
basis          0.1005      0.083      1.214      0.225      -0.062       0.263
fmha_514   -1.554e-15   3.47e-16     -4.477      0.000   -2.23e-15   -8.74e-16
fmha_515      -0.3074      0.135     -2.277      0.023      -0.572      -0.043
fmha_538      -0.4652      0.348     -1.336      0.181      -1.147       0.217
==============================================================================
```

This plot shows that it’s appropriate to use Negative Binomial regression to predict li_units as the variable is (1) non-negative, and (2) right-skewed

→ Our use case is appropriate



Model (Linear Regression) -
log(allocamt) ~ li_units + bond + credit + bias

The model uses the ordinary least squares approach to predict the log-transformed allocation amounts

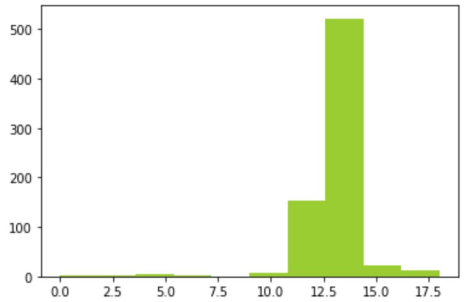
Highlights:

li_units, bond, credit

(all appears to be decent predictors of allocation amount)

RMSE ~ 1.378; the range of true target values is 18-ish

→ Distribution of true target values




```
                            OLS Regression Results
==============================================================================
Dep. Variable:               allocamt   R-squared:                       0.053
Model:                            OLS   Adj. R-squared:                  0.049
Method:                 Least Squares   F-statistic:                     13.45
Date:                Sat, 02 Mar 2024   Prob (F-statistic):           1.50e-08
Time:                        06:11:04   Log-Likelihood:                -1266.6
No. Observations:                 728   AIC:                             2541.
Df Residuals:                     724   BIC:                             2560.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
li_units       0.3335      0.058      5.772      0.000       0.220       0.447
bond          -0.4602      0.201     -2.293      0.022      -0.854      -0.066
credit        -0.1778      0.073     -2.442      0.015      -0.321      -0.035
bias          13.5026      0.177     76.485      0.000      13.156      13.849
==============================================================================
Omnibus:                      670.454   Durbin-Watson:                   1.597
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            34554.705
Skew:                          -3.948   Prob(JB):                         0.00
Kurtosis:                      35.815   Cond. No.                         11.3
==============================================================================
```

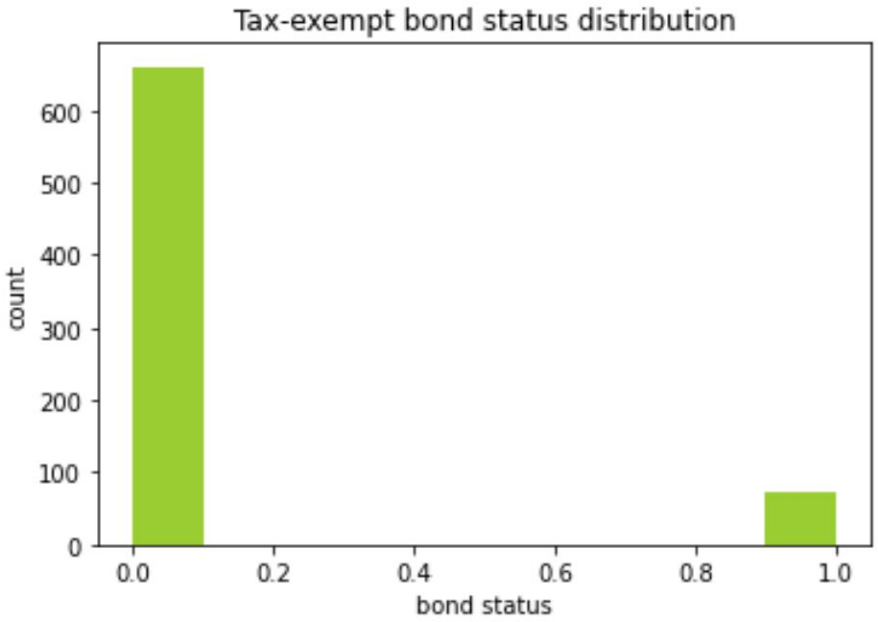
Next: Predict tax-exempt bond status (received any or not)

I first plotted the bond values to get an idea of its distribution

Notes:

Projects that received tax-exempt bond only constitute about 10% of our dataset, there is potential for issues due to data imbalance

9.822646657571624



Model (Logit)
bond ~ yr_alloc + dda + li_units + 0br + 1br + 2br + 3br + 4br
→ The #br variables are percentage of #-bedroom units

Notes:

Model structure: The logit transformation/link is the inverse of the logistic function

Takeaways:

- dda (difficulty dev. area) seems to be an informative predictor, not only based on weight but also accomodated by a lower p-value (statistically significant)

- The bedroom percentages have high coefficients but also high p-values, they may not be statistically significant, I will do more visualizations/exploratory analyses on these features to see why

- I also plan to try one-hot encoding for yr_alloc next. I'm a bit suspicious of the coefficients we're seeing because the year variable is in the thousands, which is a much higher scale than other variables, which could distort the weights

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                   bond   No. Observations:                  733
Model:                            GLM   Df Residuals:                      724
Model Family:                Binomial   Df Model:                            8
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -166.70
Date:                Sat, 02 Mar 2024   Deviance:                       333.39
Time:                        06:11:04   Pearson chi2:                     588.
No. Iterations:                     9   Pseudo R-squ. (CS):             0.1710
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       -324.8506    113.510     -2.862      0.004    -547.325    -102.376
yr_alloc       0.1564      0.056      2.788      0.005       0.046       0.266
dda            0.8725      0.226      3.858      0.000       0.429       1.316
li_units       0.0229      0.003      8.744      0.000       0.018       0.028
0br            7.5332     10.181      0.740      0.459     -12.422      27.488
1br            5.9150     10.172      0.581      0.561     -14.023      25.853
2br            5.9399     10.171      0.584      0.559     -13.994      25.874
3br            5.4855     10.187      0.538      0.590     -14.481      25.452
4br            3.5189     10.345      0.340      0.734     -16.758      23.795
==============================================================================
```

Notes:

The table to the right shows the correlations among features used in the last model. There isn't any extreme correlation I think. I will explore the relationships between the bedroom percentages, especially between 3br, 2br, and 1br → the numbers seem very interesting

	const	yr_alloc	dda	li_units	0br	1br	2br	3br	4br
const	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
yr_alloc	NaN	1.000000	-0.201236	-0.056933	-0.019778	0.023617	-0.122323	0.125489	-0.008213
dda	NaN	-0.201236	1.000000	-0.022779	-0.014489	-0.056921	-0.021319	0.071857	0.060222
li_units	NaN	-0.056933	-0.022779	1.000000	0.113240	0.052100	0.062826	-0.112445	-0.158945
0br	NaN	-0.019778	-0.014489	0.113240	1.000000	-0.058068	-0.233162	-0.139256	-0.072775
1br	NaN	0.023617	-0.056921	0.052100	-0.058068	1.000000	-0.480557	-0.538402	-0.293233
2br	NaN	-0.122323	-0.021319	0.062826	-0.233162	-0.480557	1.000000	-0.120143	-0.269329
3br	NaN	0.125489	0.071857	-0.112445	-0.139256	-0.538402	-0.120143	1.000000	0.005477
4br	NaN	-0.008213	0.060222	-0.158945	-0.072775	-0.293233	-0.269329	0.005477	1.000000

The bottom table shows # low-income units to have highest correlation with the target variable (bond status)

```
const            NaN
yr_alloc    0.046098
dda         0.093775
li_units    0.465914
0br         0.138382
1br         0.026645
2br         0.005428
3br        -0.056746
4br        -0.067888
target      1.000000
```

Model (Probit) - (Same as last model with different link function)
bond ~ yr_alloc + dda + li_units + 0br + 1br + 2br + 3br + 4br
→ The #br variables are percentage of #-bedroom units

Notes:

Model structure: The probit link function transforms the predictions into probabilities using the inverse of the CDF of the standard normal distribution

Takeaways:

Similar to the same model with logit link function

- dda (difficulty dev. area) seems to be an informative predictor, also accomodated with a lower p-value (statistically significant)

- The bedroom percentages have high coefficients but also high p-values, they may not be statistically significant, need more EDA on these features to see why

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                   bond   No. Observations:                  733
Model:                            GLM   Df Residuals:                      724
Model Family:                Binomial   Df Model:                            8
Link Function:                 Probit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -164.74
Date:                Sat, 02 Mar 2024   Deviance:                       329.48
Time:                        06:11:04   Pearson chi2:                     622.
No. Iterations:                    10   Pseudo R-squ. (CS):             0.1754
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       -161.6519     58.162     -2.779      0.005    -275.648     -47.656
yr_alloc       0.0777      0.029      2.699      0.007       0.021       0.134
dda            0.4617      0.121      3.825      0.000       0.225       0.698
li_units       0.0125      0.001      9.086      0.000       0.010       0.015
0br            3.8018      3.989      0.953      0.340      -4.016      11.619
1br            2.9132      3.977      0.732      0.464      -4.882      10.709
2br            2.9311      3.977      0.737      0.461      -4.864      10.727
3br            2.5937      3.989      0.650      0.516      -5.225      10.412
4br            1.8598      4.084      0.455      0.649      -6.145       9.865
==============================================================================
```

Our imbalance data seems to be an issue. The last model has performance analogous to simply predicting 0 for everything.

I will try resampling and other data balancing techniques to hopefully amend this problem.

```python
# Get accuracy

pred7 = pred7 > 0.5
np.mean((pred7 == y6))
```

✓ 0.0s

```
0.9045020463847203
```

```python
# Get accuracy if we predict 0 for everything
np.mean(np.array([0]*X6.shape[0]) == y6)
```

✓ 0.0s

```
0.9017735334242838
```

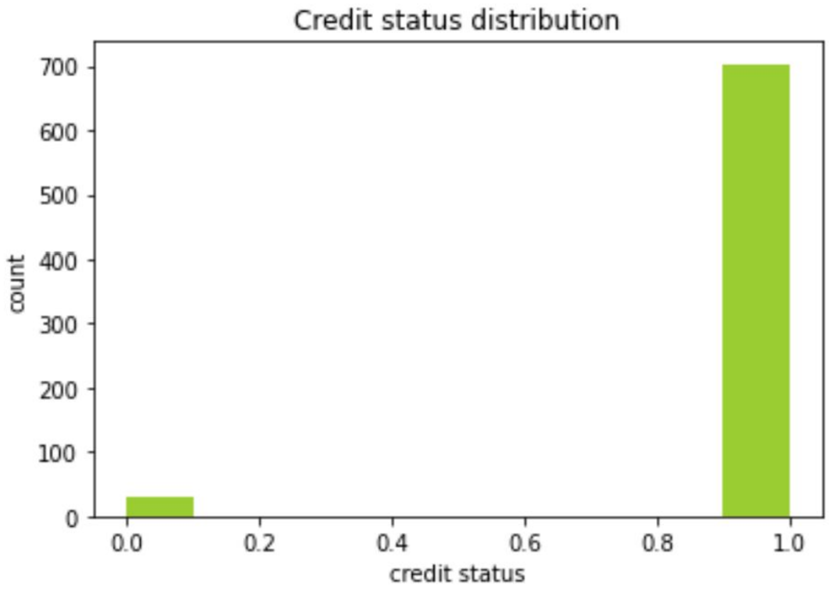
Credit has an even more extreme data imbalance issue with only about 4% of projects that are TCEP only.

We can also stay with the original credit categories:

1=30 percent present value
2=70 percent present value
3=Both
4=TCEP only

TCEP - Tax Credit Exchange Program funds

4.092769440654844



Takeaways and next steps:

In addition to the next-steps I mentioned throughout the slides, there is still a good number of variables I haven't checked out yet (and models), so I will try to get to them.

Something interesting I think we can look into is deriving a list of patterns that could potentially be informative to our "LIHTC-score" scale. For ex, in the right plot, it's pretty rare for a project to receive bond while also getting a Section 515 loan.