FIN 550: Final Project
EXECUTIVE SUMMARY

Your Team Name (be creative): **In Data We Trust**

Select whether this is an individual or group submission. **No more than 3 members per group.** Beyond the fact that all group members may submit the same answers, each submission must be separate work.

☐ Group Submission. Group member names:

Javeria Malik, Ahmed Bilal, Shehzadi Mahum Agha

Case Overview

The Cook County Assessor's Office (CCAO) is responsible for determining the fair market value of over 1.8 million properties annually, a process critical for property tax revenue generation. Historically, valuation methods lacked transparency and precision, leading to public criticism. To address these challenges, CCAO embraced data science and machine learning techniques to enhance valuation accuracy.

As a data scientist on the team, the objective is to predict the residential property values of 10,000 homes as accurately as possible. This involves leveraging historical property sales data to develop models that minimize the Mean Squared Error (MSE) of these predictions, providing actionable insights and improvements for property assessments. We have tested multiple machine-learning models including Linear Regression, Lasso Regression, Forward Stepwise on Linear Regression, and the Random Forest to attain the minimum possible MSE to choose maximum accuracy to further test property sales data.

Methodology

For predicting household valuations, we adopted a systematic approach that involved multiple modeling techniques to ensure accuracy and reliability. The methodology involved three key steps: initial data preparation, model development, and model selection based on performance.

Data Preparation

The first step in the process was cleaning and preparing the data. This included handling missing values, ensuring that categorical variables were properly encoded, and normalizing continuous variables to ensure consistency across different scales. Any outliers or inconsistencies were addressed through appropriate data transformations, allowing the model to perform optimally. The dataset for most of the categorical variables had many missing values which made us drop those variables for our model to remain consistent and precise.

Model Development and Selection

We experimented with several modeling techniques to identify the best approach. Initially, we applied **linear regression** to establish a baseline model. To ensure a precise and uncomplicated analysis of our variables, we picked those predictors for our analysis that were most appropriate logically and were expected to explain the

sale price of houses well. This model allowed us to assess the overall relationship between the features and the target variable, providing a simple and interpretable starting point. To test the model's fit we calculated its cross-validated MSE using the k-fold cross-validation method. The CVMSE received was around 15952188508 as shown in **Fig 1.** The predictors we started with were: (meta_certified_est_bldg, meta_certified_est_land, char_bldg_sf, char_age, char_rooms, meta_class, char_bsmt, char_gar1_area, econ_midincome, geo_tract_pop, geo_black_perc, geo_asian_perc, econ_tax_rate, char_fbath, char_beds, char_frpl). The R-square for this initial model was around 83.46%. To decrease the CV MSE of the current model, we explored additional techniques, including lasso regression, forward stepwise selection, and random forest.

We used **Lasso regression**, after choosing the lambda with the least CV MSE as our penalty parameter we tried to shrink our model to drop those variables that were less important and were causing us to have a very high CV MSE. After running our lasso model the CV MSE received was 15914348180(**Fig 2**) and similar R-square as before. This number was lower than what we received for our initial linear model. The lasso regression took out those variables that were not important for our predictions. The following predictors were shrunk by lasso regression: (char_age, char_rooms, char_bsmt, geo_black_perc). This made us realize that Lasso is a better fit as compared to the linear model given the type of data we had.

Later on, to achieve an even better cross-validated MSE for our model and to ensure that we only include important and statistically relevant predictors in our model, we next implemented **forward stepwise subset selection**. This method systematically adds variables to the model based on their statistical significance. It helped identify a smaller set of features that were most influential in predicting household valuations, reducing the complexity of the model without compromising predictive power. After running this model, variables received were: (meta_certified_est_bldg, meta_certified_est_land, econ_midincome, char_fbath, geo_black_perc, geo_tract_pop, char_bsmt, char_beds, econ_tax_rate, char_gar1_area, meta_class, geo_asian_perc).

Running our linear regression model using the relevant variables mentioned above, we received a cross-validated MSE of around 15883616317 and an R-square of around 83.23% (Fig 3).

Finally, we modeled a random forest model using our initial predictors of interest just to have a more holistic analysis. The Random Forest technique is an ensemble learning method that builds multiple decision trees and combines their outputs to improve prediction accuracy and reduce overfitting. This model gave us a CVMSE of 16994333626 and R2 82.31% (Fig 4). The CV MSE for this model was higher than what we received for our linear regression after adjusting it with variables recommended by the forward stepwise technique.

At the end we used the linear regression model with the reduced set of features recommended by the forward stepwise method, since it gave us the lowest CV MSE and ensured that the model not only performed well but was also efficient and easy to interpret, making it ideal for deployment and further analysis. This approach strikes a balance between model accuracy and interpretability, aligning with the project's goal of making reliable and understandable predictions of household valuations.

Conclusion

The final model predicted residential property values with significant variability across the dataset. Key summary statistics of the assessed values are as follows: Minimum Value: $0, 1st Quartile: $109,002, Median: $230,757, Mean: $285,045, 3rd Quartile: $359,631, Maximum Value: $6,207,654.

The high maximum value suggests the presence of extreme outliers or properties with exceptional characteristics. The CV MSE of the model was minimized to 15883616317, demonstrating a balance between accuracy and model complexity. This ensures that the predictions are reliable for practical use.

Our recommendation to the Cook County Assessor's Office is to use a linear model with forward stepwise selection to minimize CVMSE and achieve accurate predictions. In our view, probabilistic models are not advisable in this scenario given our variables of importance. CCAO can use the predictions given to enhance fairness in property taxation by addressing historical biases and inaccuracies in valuation. The results also align with the project's goal of improving transparency and public trust in the assessment process.

Appendix

Fig 1 Linear:

```
> # Get the RMSE (Root Mean Squared Error) and calculate MSE (Mean Squared Error)
> cat("Cross-validated RMSE:", cv_results$RMSE, "\n")
Cross-validated RMSE: 126302
> cat("Cross-validated MSE:", cv_results$RMSE^2, "\n")
Cross-validated MSE: 15952188508
```

Linear model result:

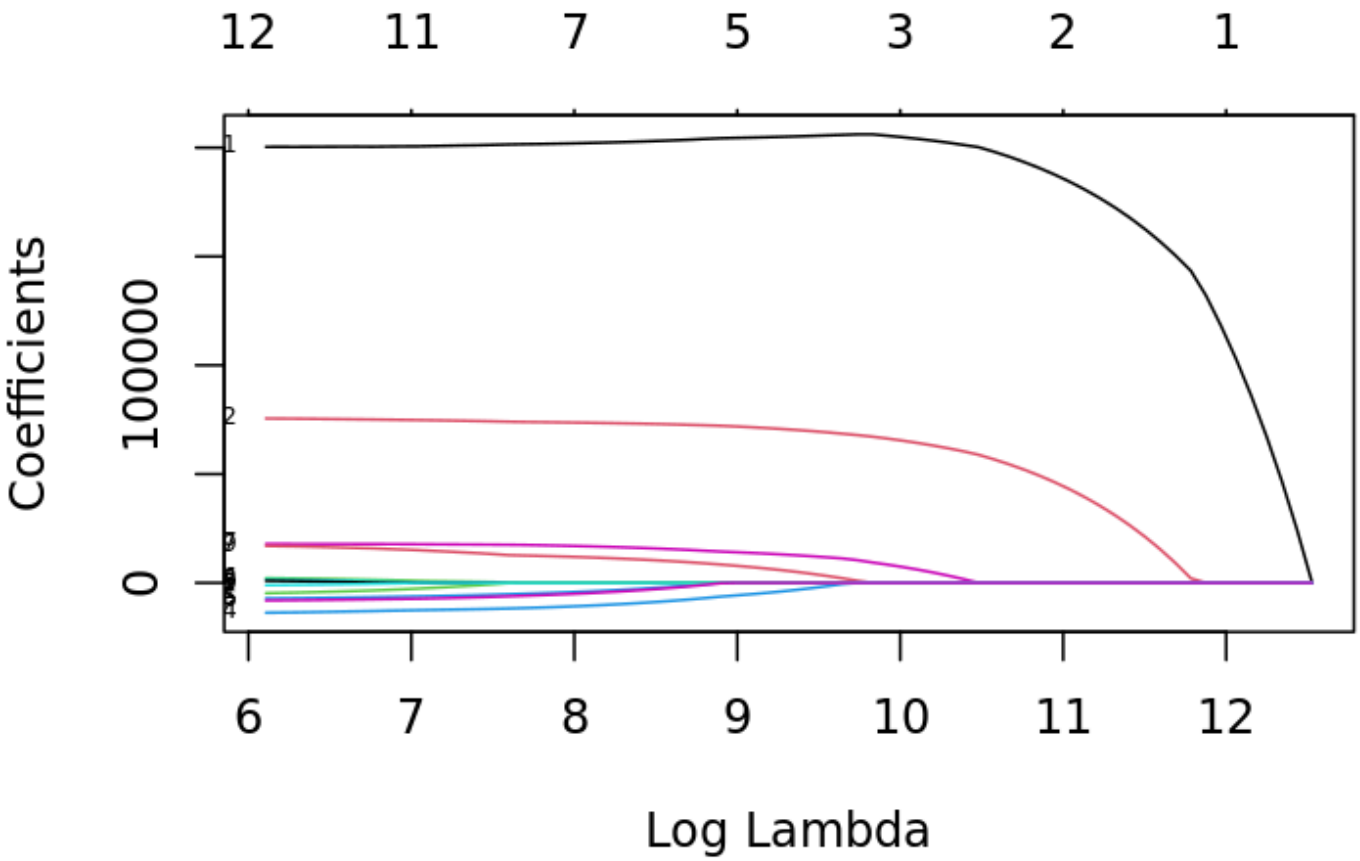
```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -4.921e+04  9.820e+03  -5.012 5.41e-07 ***
meta_certified_est_bldg   9.322e-01  5.233e-03 178.126  < 2e-16 ***
meta_certified_est_land   1.578e+00  1.871e-02  84.315  < 2e-16 ***
char_bldg_sf              1.055e+00  1.493e+00   0.707 0.479639
char_age                  1.920e+01  2.790e+01   0.688 0.491207
char_rooms                2.569e+02  5.443e+02   0.472 0.636919
meta_class                1.088e+02  2.751e+01   3.957 7.60e-05 ***
char_bsmt                -7.801e+03  6.476e+02 -12.048  < 2e-16 ***
char_gar1_area            8.823e+03  2.369e+03   3.725 0.000196 ***
econ_midincome            5.089e-01  2.401e-02  21.195  < 2e-16 ***
geo_tract_pop            -4.993e+00  3.839e-01 -13.005  < 2e-16 ***
geo_black_perc           -4.518e+04  2.356e+03 -19.176  < 2e-16 ***
geo_asian_perc           -2.079e+04  8.339e+03  -2.493 0.012679 *
econ_tax_rate             4.318e+02  1.507e+02   2.866 0.004161 **
char_fbath                2.179e+04  1.274e+03  17.108  < 2e-16 ***
char_beds                -5.668e+03  1.048e+03  -5.411 6.31e-08 ***
char_frpl                 1.680e+02  1.272e+03   0.132 0.894916
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 126100 on 42693 degrees of freedom
Multiple R-squared:  0.8346,    Adjusted R-squared:  0.8346
F-statistic: 1.347e+04 on 16 and 42693 DF,  p-value: < 2.2e-16
```

Fig 2:

Lasso:

```
> cat("Cross-validated MSE:", cv_results2$RMSE^2, "\n")
Cross-validated MSE: 15914348180
> cat("Cross-validated RMSE:", cv_results2$RMSE, "\n")
Cross-validated RMSE: 126152.1
```

lasso plot:




```
Best lambda: 449.5587
```

```
Important variables:
> print(important_vars)
 [1] "meta_certified_est_bldg" "meta_certified_est_land" "meta_class"
 [4] "char_bsmt"               "char_gar1_area"          "econ_midincome"
 [7] "econ_tax_rate"           "char_fbath"              "char_beds"
[10] "geo_black_perc"          "geo_asian_perc"          "geo_tract_pop"
```

Lasso model results:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -4.921e+04  9.820e+03  -5.012 5.41e-07 ***
meta_certified_est_bldg  9.322e-01  5.233e-03 178.126  < 2e-16 ***
meta_certified_est_land  1.578e+00  1.871e-02  84.315  < 2e-16 ***
char_bldg_sf             1.055e+00  1.493e+00   0.707 0.479639
char_age                 1.920e+01  2.790e+01   0.688 0.491207
char_rooms               2.569e+02  5.443e+02   0.472 0.636919
meta_class               1.088e+02  2.751e+01   3.957 7.60e-05 ***
char_bsmt               -7.801e+03  6.476e+02 -12.048  < 2e-16 ***
char_gar1_area           8.823e+03  2.369e+03   3.725 0.000196 ***
econ_midincome           5.089e-01  2.401e-02  21.195  < 2e-16 ***
geo_tract_pop           -4.993e+00  3.839e-01 -13.005  < 2e-16 ***
geo_black_perc          -4.518e+04  2.356e+03 -19.176  < 2e-16 ***
geo_asian_perc          -2.079e+04  8.339e+03  -2.493 0.012679 *
econ_tax_rate            4.318e+02  1.507e+02   2.866 0.004161 **
char_fbath               2.179e+04  1.274e+03  17.108  < 2e-16 ***
char_beds               -5.668e+03  1.048e+03  -5.411 6.31e-08 ***
char_frpl                1.680e+02  1.272e+03   0.132 0.894916
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 126100 on 42693 degrees of freedom
Multiple R-squared:  0.8346,    Adjusted R-squared:  0.8346
F-statistic: 1.347e+04 on 16 and 42693 DF,  p-value: < 2.2e-16
```

Stepwise:

```
> cat("Final Model RMSE:", final_rmse, "\n")
Final Model RMSE: 126030.2
> cat("Final Model MSE:", final_rmse^2, "\n")
Final Model MSE: 15883616317
```

Fig 3:
Stepwise variables:

```
> best_model_formula
sale_price ~ meta_certified_est_bldg + meta_certified_est_land +
    econ_midincome + char_fbath + geo_black_perc + geo_tract_pop +
    char_bsmt + char_beds + econ_tax_rate + char_gar1_area +
    meta_class + geo_asian_perc
```

stepwise model result:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -4.608e+04  8.764e+03  -5.258 1.47e-07 ***
meta_certified_est_bldg  9.329e-01  4.833e-03 193.035  < 2e-16 ***
meta_certified_est_land  1.582e+00  1.826e-02  86.601  < 2e-16 ***
econ_midincome           5.073e-01  2.359e-02  21.508  < 2e-16 ***
char_fbath               2.232e+04  1.160e+03  19.246  < 2e-16 ***
geo_black_perc          -4.496e+04  2.335e+03 -19.250  < 2e-16 ***
geo_tract_pop           -5.004e+00  3.760e-01 -13.310  < 2e-16 ***
char_bsmt               -7.897e+03  6.395e+02 -12.350  < 2e-16 ***
char_beds               -4.853e+03  6.657e+02  -7.290 3.15e-13 ***
econ_tax_rate            4.103e+02  1.472e+02   2.788 0.005302 **
char_gar1_area           8.588e+03  2.302e+03   3.731 0.000191 ***
meta_class               1.030e+02  2.494e+01   4.131 3.61e-05 ***
geo_asian_perc          -2.105e+04  8.320e+03  -2.530 0.011403 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 126000 on 42697 degrees of freedom
Multiple R-squared:  0.8346,    Adjusted R-squared:  0.8346
F-statistic: 1.796e+04 on 12 and 42697 DF,  p-value: < 2.2e-16
```

Fig 4:
random forest

```
> # View the Random Forest model summary
> print(rf_model)

Call:
 randomForest(formula = formula_rf, data = df2, ntree = 10, mtry = 5,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 10
No. of variables tried at each split: 5

          Mean of squared residuals: 16994333626
                    % Var explained: 82.31
```

summary stats

```
> print(summary_stats)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0  109002  230757  285045  359631 6207654
```