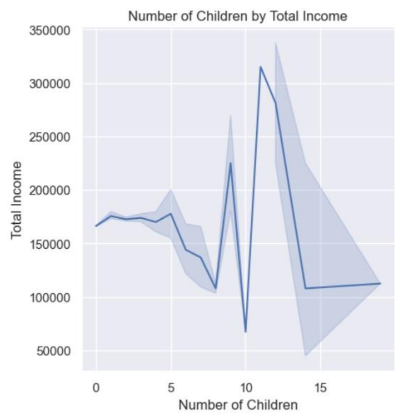
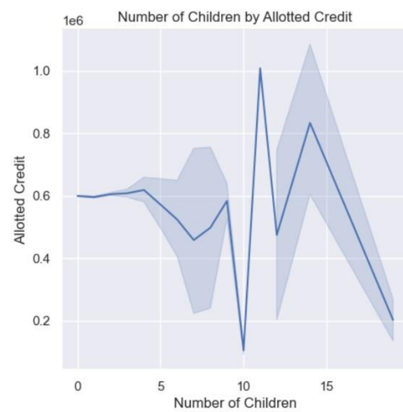
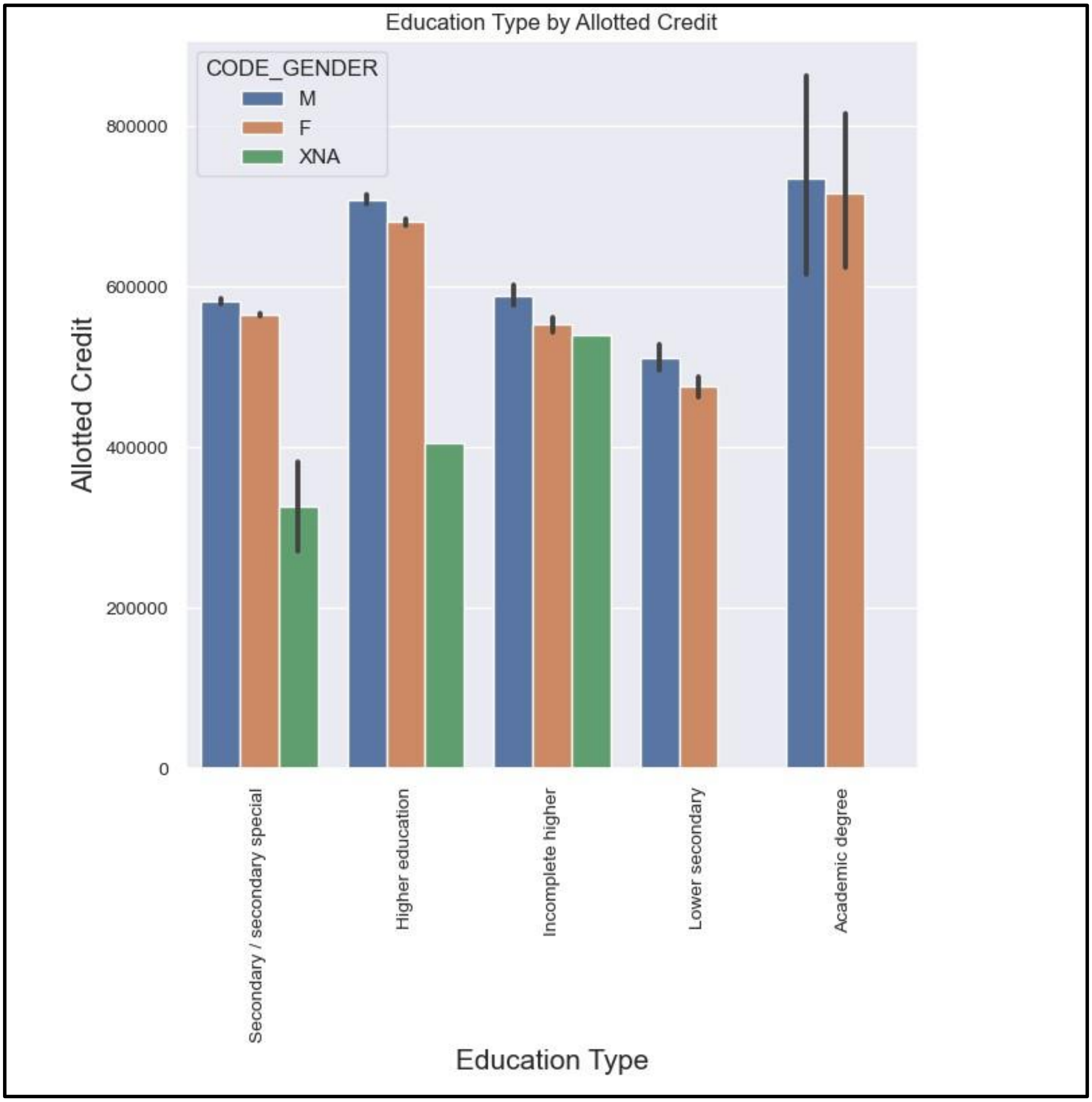
Measuring the Effect of Guardianship on the Likelihood of Loan: Key Preliminary Finding Excerpt

These are statistical graphs and metrics created to visualize the features effect on allotted created and total income.

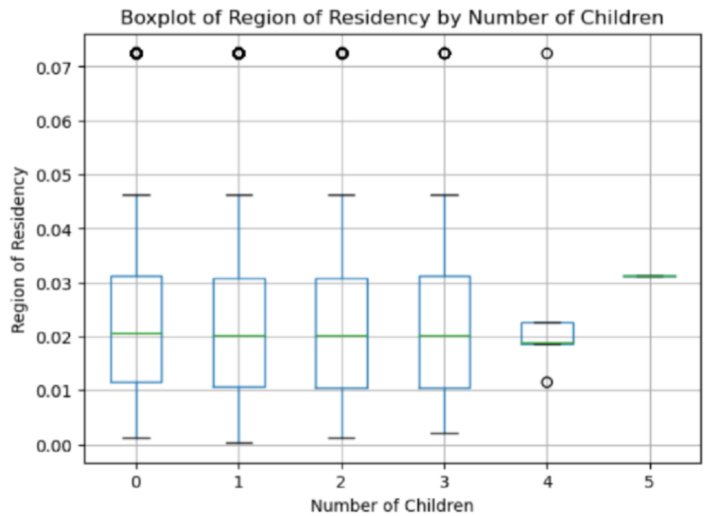


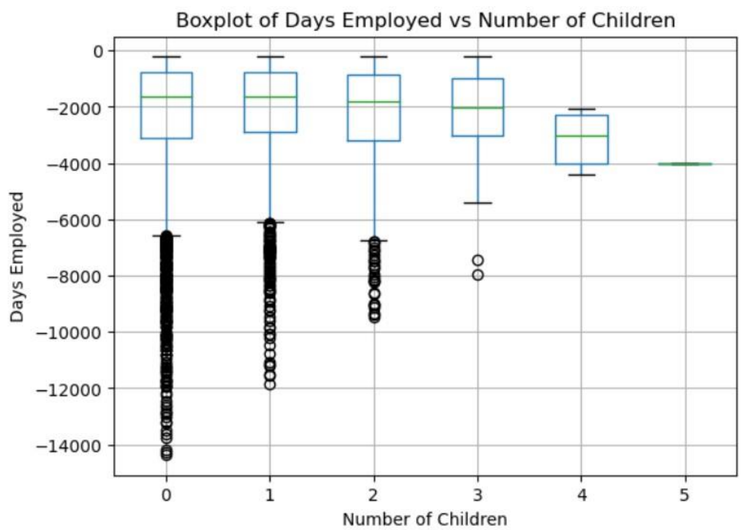




The first two graphs display the allotted credit vs the number of children. Between 150,000 and 200,000 dollars between an average of 0 to 5 children. The data peaks at around 9-10 children as individuals at nine children have above 200,000 as an income and below 100,000 at ten children. Past 12.5 children, the graph has a steep upward slope and is slightly above 100,000 dollars. From the bar chart, one can see that those with an academic degree and higher education have the highest credit limit. The lowest secondary, meaning grades 6-8th, and secondary special being grades

9-12th.In all categories, Males have higher credit than females or non-gender specific.

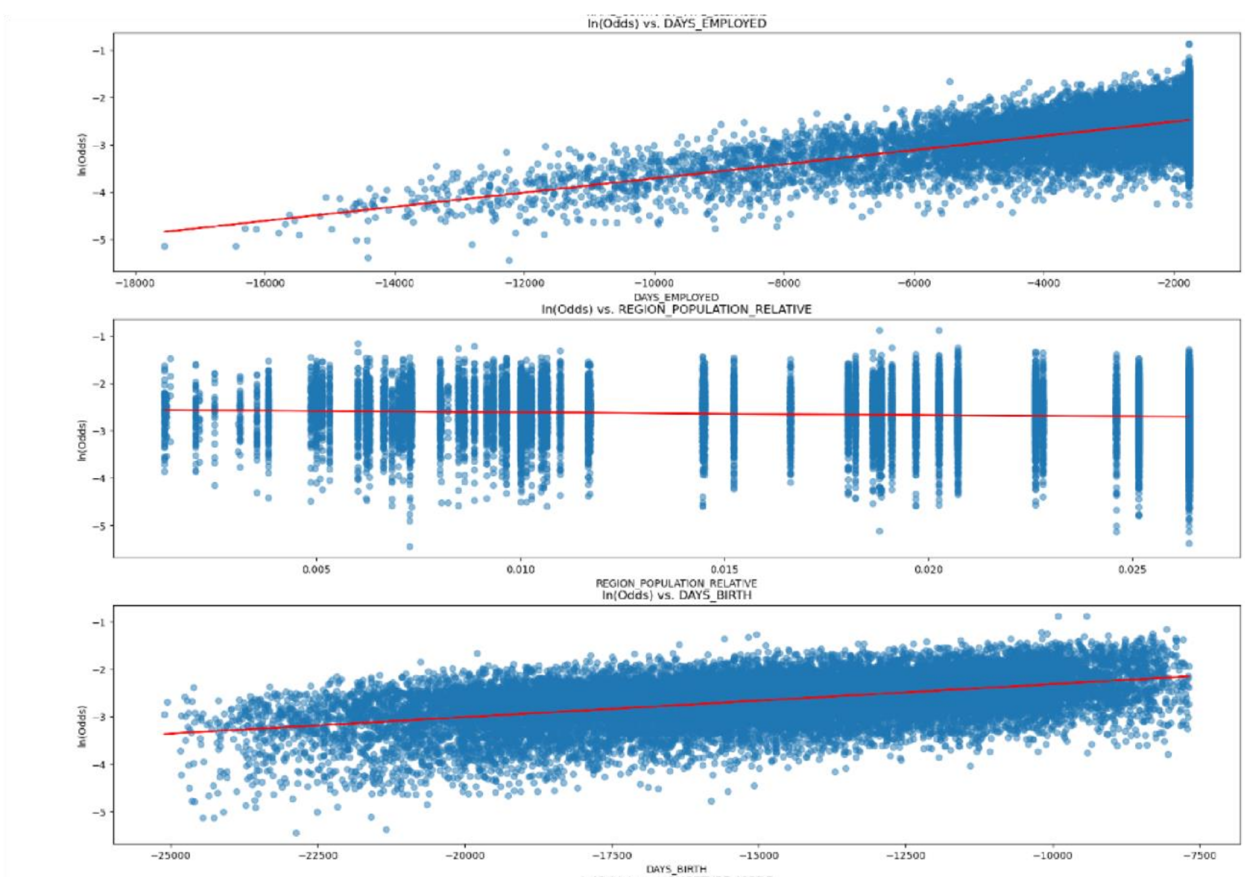




Model ideas came from a similar study conducted by the Journal of Big Data; in their study they focused on Naive Bayes, Logistic regression, and Random Forest. For the first model, logistic regression, The categorical variables were mapped and zero through their respective labeling and added to the data frame for modeling. The exact process was done for binary variables such as contract type, own car, and own reality emergency state. Before any logistic modeling, predictor variables with a high number of outliers were windsored at an 80 percent threshold. These were 'AMT_ANNUITY',' 'AMT_INCOME_TOTAL,' 'AMT_CREDIT', and' DAYS_EMPLOYED'Variables were scaled in attempts to handle imbalance classes and produce better metrics. A stratified shuffle split was done with a test size of thirty, completing seven interactions with a random state of 42. The pros of this are more robust evaluation, as the model is trained and tested on multiple subsets of data. The cons of this are more computation time, and fifteen splits are not suitable for small data sets, as the training set becomes smaller with each division. The outcome variable was TARGET; 1 client had payment difficulties, or 0 did not. As one can see, there are a handful of negative coefficients and positive coefficients. The highest coefficient is HOUSETYPE_MODE, with the associated value of 0.1978, and the lowest coefficient is INCOME_TOTAL -2.206e-6. In this data set, as income increases, the difficulty paying increases. Individuals in specific residencies apartment size, number of entrances, and state of the building- are more likely to have payment difficulties. The coefficient for the number of children is -0.0835. If the coefficient were to increase, then the likelihood of not making payments would increase. The y_predict threshold was set to 0.19 instead of 0.5 to produce false positives and true negatives. Setting the y_predict at or above 3 produced no false positives or true negatives.

The results display a 93% accuracy. The confusion matrix presents 16391 True positives, as in difficulties paying back, 1259 false positives, 89 false negatives, and 19 true negatives. The precision score is 25% and can only predict 25 of the true positives, those with difficulties paying back.

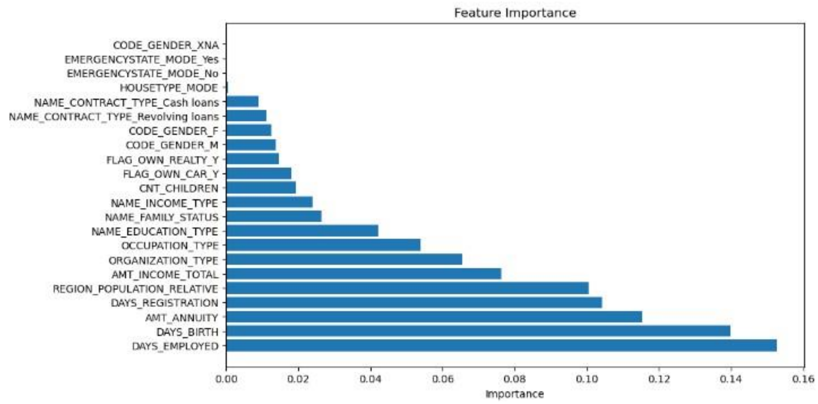
```
(58995, 19) (58995,)
Optimization terminated successfully.
         Current function value: 0.251324
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:                41296
Model:                          Logit   Df Residuals:                    41278
Method:                           MLE   Df Model:                           17
Date:                Sun, 12 Nov 2023   Pseudo R-squ.:                 0.03060
Time:                        11:48:25   Log-Likelihood:                -10379.
converged:                       True   LL-Null:                       -10706.
Covariance Type:            nonrobust   LLR p-value:                2.763e-128
=====================================================================================================
                                         coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                                 -0.5764   1.03e+06   -5.6e-07      1.000   -2.02e+06    2.02e+06
CODE_GENDER                           -0.3748      0.043     -8.723      0.000      -0.459      -0.291
CNT_CHILDREN                          -0.0835      0.028     -2.989      0.003      -0.138      -0.029
NAME_INCOME_TYPE                      -0.1224      0.022     -5.515      0.000      -0.166      -0.079
NAME_CONTRACT_TYPE_Cash loans         -0.0269   1.03e+06  -2.62e-08      1.000   -2.02e+06    2.02e+06
DAYS_EMPLOYED                          0.0001   1.32e-05      8.943      0.000     9.2e-05       0.000
REGION_POPULATION_RELATIVE            -1.1223      2.565     -0.438      0.662      -6.149       3.904
DAYS_BIRTH                          6.219e-05   6.11e-06     10.180      0.000    5.02e-05    7.42e-05
HOUSETYPE_MODE                         0.1978      0.055      3.612      0.000       0.090       0.305
DAYS_REGISTRATION                   2.334e-06   6.17e-06      0.379      0.705   -9.75e-06    1.44e-05
NAME_CONTRACT_TYPE_Revolving loans    -0.5495   1.03e+06  -5.34e-07      1.000   -2.02e+06    2.02e+06
OCCUPATION_TYPE                       -0.0106      0.005     -2.270      0.023      -0.020      -0.001
AMT_ANNUITY                         9.249e-06   3.78e-06      2.445      0.014    1.84e-06    1.67e-05
FLAG_OWN_CAR                          -0.2475      0.043     -5.772      0.000      -0.332      -0.163
FLAG_OWN_REALTY                        0.0741      0.042      1.782      0.075      -0.007       0.156
AMT_INCOME_TOTAL                   -2.206e-06   7.16e-07     -3.080      0.002   -3.61e-06   -8.02e-07
ORGANIZATION_TYPE                     -0.0045      0.002     -2.451      0.014      -0.008      -0.001
NAME_EDUCATION_TYPE                   -0.3367      0.037     -9.159      0.000      -0.409      -0.265
NAME_FAMILY_STATUS                     0.0460      0.018      2.619      0.009       0.012       0.080
=====================================================================================================
Model Performance Metrics:
Accuracy: 0.93
precision: 0.25
recall: 0.01
f1: 0.02
Confusion Matrix:
[[16391    30]
 [ 1268    10]]
```




```
Out-of-Bag Perf: 0.8314
Out-of-Bag (OOB) Error: 0.1686
Train Accuracy: 0.8868162261831591
Accuracy: 0.8323653832365383
precision: 0.16
Confusion Matrix (with Labels):
                        Predicted 0(Make Payments)  Predicted 1(Difficulties)
Actual 0(Make Payments)                      11311                       1576
Actual 1(Difficulties)                         762                        298
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.88      0.91     12887
           1       0.16      0.28      0.20      1060

    accuracy                           0.83     13947
   macro avg       0.55      0.58      0.55     13947
weighted avg       0.88      0.83      0.85     13947
```




```
Feature Importance (sorted by importance):
DAYS_EMPLOYED: 0.1527
DAYS_BIRTH: 0.1398
AMT_ANNUITY: 0.1155
DAYS_REGISTRATION: 0.1042
REGION_POPULATION_RELATIVE: 0.1005
AMT_INCOME_TOTAL: 0.0763
ORGANIZATION_TYPE: 0.0655
OCCUPATION_TYPE: 0.0540
NAME_EDUCATION_TYPE: 0.0421
NAME_FAMILY_STATUS: 0.0264
NAME_INCOME_TYPE: 0.0240
CNT_CHILDREN: 0.0193
FLAG_OWN_CAR_Y: 0.0182
FLAG_OWN_REALTY_Y: 0.0146
CODE_GENDER_M: 0.0138
CODE_GENDER_F: 0.0125
NAME_CONTRACT_TYPE_Revolving loans: 0.0110
NAME_CONTRACT_TYPE_Cash loans: 0.0090
HOUSETYPE_MODE: 0.0006
EMERGENCYSTATE_MODE_No: 0.0001
EMERGENCYSTATE_MODE_Yes: 0.0000
CODE_GENDER_XNA: 0.0000
```

The second model was a random forest. The model uses the sample predictive variables; however, GENDER_CODE, EMERGENCYSTATE_MODE, and FLAG_OWN_CAR were broken down into binary and categorical outcomes. The settings for the hyperparameter tuning are as follows: 115 estimators, max depth equal to None. The minimum samples per leaf and the minimum samples per split were set to 25. The model used balanced class weight in the case of unbalanced classes. The max features were log2, and the random state was 42. From this model, it is 94% precise at finding those who can make payments, and the precision score was 16% for those who cannot. The overall accuracy is 88%. The most important feature on the list is the number of days employed and the number of children falling in the middle of the list at 0.912. The random forest appears to be the better technique between the two models.

Resources:

- LLeberi. E, Sun. Y, Wangm Z.J (2022, February 2022). *Journal of Big Data.* Retrieved from: https://doi.org/10.1186/s40537-022-00573-8.
- Kaggle.(2019) Kaggle. https://www.kaggle.com/datasets/mishra5001/credit-card