Credit Card Fraud Detection Dataset:
https://www.kaggle.com/datasets/mishra5001/credit-card/data

# Introduction

In the field of credit card loans, fraud detection has always been on the radar for most banks. With the usage of credit cards being the norm among most individuals to reap benefits from cashback, miles or rewards points, and building credit scores.
Institutions such as banks that provide credit card loans tend to have an extensive Know-Your-Customer (KYC) and Due Diligence process in attaining various information about each customer, ranging from compliance laws by the government to curb money laundering, to gauging one's creditworthiness.

In this project, my goal would be to gain a better understanding of individuals and their propensity to default on credit card payments, to not only better predict but also recognize which attributes could contribute to this probability of committing fraud.

## Presentation

Initially, at the point of the presentation, I conducted dimensionality reduction by first manually selecting the ideal attributes as I recognized the computational power required to conduct stepwise model selection with 164 attributes. I then proceeded to conduct said stepwise model selection after narrowing down the attributes to 15, then continued with multicollinearity checks to reduce it further. Through the finalized variables, I've built my econometrics model, with the dependent variable being `TARGET`, which is the outcome of which individual defaulted, and the independent variable being `AMT_ANNUITY`, to understand the financial burden each individual experiences per annum.

However, the Instrument Variable (IV) chosen, `CNT_CHILDREN` was weakly correlated to the independent variable. Attempting to strengthen it through overidentification by including `AMT_CREDIT`, `NAME_INCOME_TYPE`, `REGION_POPULATION_RELATIVE`, and `OCCUPATION_TYPE`, provided a strong IV, with endogeneity being unlikely a significant issue, however fails the Sargan test, resulting in at least one instrument not being valid.

## Post-Presentation

Thus, I've decided instead to look directly for the IV with all the variables available through a correlation matrix, accounting for a threshold of >= 0.7 to ensure an IV that is strongly correlated to my main independent variable `AMT_ANNUITY`, before building onto it with a Causal Forest to understand any Heterogeneous Treatment Effects and the Importance of specific subgroups.

# Locating of IVs

After filtering by indicating a threshold of >= 0.7, I've located 2 variables that are strongly correlated with `AMT_ANNUITY` (which is not AMT_ANNUITY itself).

The other variables include:

1. `AMT_CREDIT`
2. `AMT_GOODS_PRICE`

I've ultimately decided on `AMT_GOODS_PRICE` as it does not directly indicate default risk, and thus, is more likely to fulfil the exogeneity condition.
`AMT_CREDIT` on the other hand, is very closely tied to `TARGET` and general creditworthiness assessments, and thus, is unlikely to have an impact on `TARGET` solely through AMT_ANNUITY, unlike `AMT_GOODS_PRICE`, and could directly impact the default likelihood, making it unlikely to fulfill the exogeneity condition.

## First Stage Regression

```
Call:
lm(formula = AMT_ANNUITY ~ AMT_GOODS_PRICE, data = num_app_data)

Residuals:
   Min     1Q Median     3Q    Max
-30865  -7385  -2041   4618 146403

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.302e+04  1.820e+02    71.5   <2e-16 ***
AMT_GOODS_PRICE 2.915e-02  2.358e-04   123.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10820 on 11349 degrees of freedom
Multiple R-squared:  0.5738,    Adjusted R-squared:  0.5738
F-statistic: 1.528e+04 on 1 and 11349 DF,  p-value: < 2.2e-16
```

This first stage regression for Instrument Variable regresses `AMT_ANNUITY` against `AMT_GOODS_PRICE`. The statistical significance of the intercept estimate indicates that even without any expenditure of goods (if `AMT_GOODS_PRICE` is 0), there will be a fixed or baseline cost of `AMT_ANNUITY`, at 1.301e+04. This captures all other factors that contribute to `AMT_ANNUITY`. The statistical significance of the estimate of `AMT_GOODS_PRICE` indicates that for every marginal one unit (dollar) increase in `AMT_GOODS_PRICE`, there is an associated value of  2.915e-02 increase in `AMT_ANNUITY`.

The relatively strong R^2 of 0.5738 indicates that the model is able to explain more than half the variability in `AMT_ANNUITY`, showcasing a reliable model for predicting `AMT_ANNUITY` based on `AMT_GOODS_PRICE` and supports the idea that `AMT_GOODS_PRICE` is a good instrument.

## Second Stage Regression

```
Call:
lm(formula = TARGET ~ predicted_AMT_ANNUITY, data = num_app_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.07026 -0.06607 -0.06127 -0.05581  0.96644

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           7.844e-02  6.064e-03   12.94  < 2e-16 ***
predicted_AMT_ANNUITY -5.710e-07  1.779e-07   -3.21  0.00133 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.238 on 11349 degrees of freedom
Multiple R-squared:  0.0009069, Adjusted R-squared:  0.0008189
F-statistic:  10.3 on 1 and 11349 DF,  p-value: 0.001333
```

The second stage regression of the IV analysis regresses the `TARGET` against the `predicted_AMT_ANNUITY`, which should mitigate bias of `AMT_ANNUITY` due to endogeneity by leveraging the instrumented values from the first stage regression. This would provide a more accurate estimate of the causal relationship between the treatment `AMT_ANNUITY` and outcome `TARGET`.

The intercept is statistically significant, at 7.844e-02, indicating that there is a baseline probability of `TARGET` being 1 when `predicted_AMT_ANNUITY` is 0. (in absence of the treatment effect).
The coefficient of `predicted_AMT_ANNUITY` indicates that for each marginal increase in `predicted_AMT_ANNUITY`, the probability of `TARGET` being 1 is associated with a decrease by 5.71e-07, albeit, by a very small amount. With it being significant due to the p-value being less than 0.05.

## IV Validity Tests

```
   value    numdf    dendf
15281.47     1.00 11349.00
                 df1   df2   statistic      p-value
Weak instruments   1 11349 15281.46520 0.000000e+00
Wu-Hausman         1 11348    16.72286 4.355513e-05
Sargan             0    NA          NA           NA
```

F-Statistic from the First Stage Regression indicates a value of 15281.47, which is significantly greater than 10, the threshold of what is considered high enough for the strong instruments to be highly correlated with the endogenous variable.
The Wu-Hausman p-value of 4.35e-05 indicates that there are little to no endogeneity problems, with the explanatory variable of `AMT_ANNUITY` being correlated with the error term, resulting in OLS estimates being biased and inconsistent, with `AMT_GOODS_PRICE` being justified as the IV to address the bias.
The Sargan test is NA as it is an overidentification test, with this only having one IV.

## Overidentification of IVs

Conducting an overidentification test, with both `AMT_GOODS_PRICE` and `AMT_CREDIT` being the combined IVs.

First Stage Regression:

```
Call:
lm(formula = AMT_ANNUITY ~ AMT_GOODS_PRICE + AMT_CREDIT, data = clean_app_data)

Residuals:
   Min     1Q Median     3Q    Max
-29327  -7191  -2115   4410 147309

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.290e+04  1.847e+02  69.827  < 2e-16 ***
AMT_GOODS_PRICE 2.370e-02  1.516e-03  15.633  < 2e-16 ***
AMT_CREDIT      5.094e-03  1.401e-03   3.635 0.000279 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10820 on 11348 degrees of freedom
Multiple R-squared:  0.5743,    Adjusted R-squared:  0.5743
F-statistic:  7656 on 2 and 11348 DF,  p-value: < 2.2e-16
```

Second Stage Regression:

```
Call:
lm(formula = TARGET ~ predicted_AMT_ANNUITY, data = clean_app_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06985 -0.06574 -0.06121 -0.05585  0.96587

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            7.756e-02  6.062e-03  12.795  < 2e-16 ***
predicted_AMT_ANNUITY -5.432e-07  1.778e-07  -3.055  0.00226 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2381 on 11349 degrees of freedom
Multiple R-squared:  0.0008216, Adjusted R-squared:  0.0007335
F-statistic: 9.331 on 1 and 11349 DF,  p-value: 0.002258
```

With all estimates of the first and second stage regression being statistically significant, the interpretation would be similar to what I've done previously, indicating that the combined IVs are generally accepted to be valid.

F-Statistic:

| value | numdf | dendf |
|---|---|---|
| 7655.563 | 2.000 | 11348.000 |

This is further strengthened by the F-Statistic of the First Stage Regression being above 10 as well

Overidentification Test:

| Test <chr> | df1 <dbl> | Statistic <dbl> | p_value <dbl> |
|---|---|---|---|
| Weak Instruments | 2 | 7655.56321 | 0.000000e+00 |
| Wu-Hausman | 1 | 14.85146 | 1.169592e-04 |
| Sargan | 1 | 28.18838 | 1.100634e-07 |

The overidentification test shows significant p-values for weak instruments and the Wu-Hausman test, indicating that this combined IV is likely strong with little to no endogeneity. The Sargan test being statistically significant however indicates that one or more instruments may not be valid. In this case, likely `AMT_CREDIT` as mentioned in the introduction as it likely does not have instrument exogeneity.
However, since these are the only variables correlated to the main independent variable `AMT_ANNUITY` in this dataset of 164 variables, the IV of `AMT_GOODS_PRICE` would be our best bet.

Since the Wu-Hausman test for `AMT_GOODS_PRICE` is at a lower p-value, and the Sargan test for the overidentification test indicates that one or more IVs are invalid, I will proceed with using `AMT_GOODS_PRICE` as the sole IV, without combining the IVs to include `AMT_CREDIT`.

Verifying other correlations with `AMT_GOODS_PRICE`:

```
[1] "Non-numeric Amt_Goods_Price values: 278"
```

Variables Correlated with `AMT_GOODS_PRICE`:

```
"AMT_CREDIT"       "AMT_ANNUITY"      "AMT_GOODS_PRICE"
```

Considering `AMT_GOODS_PRICE` is among the 3 variables correlated with `AMT_GOODS_PRICE`,  that can be ignored, and `AMT_ANNUITY` and `AMT_CREDIT` are majorly causally linked.
This pushes for the notion that there is Instrument Exogeneity as `AMT_GOODS_PRICE` may only impact `TARGET` through `AMT_ANNUITY`

# Causal Forest

Using a causal forest in conjunction with instrumental variables (IVs) can help locate heterogeneous treatment effects and identify subgroups with varying responses to the treatment. I will be using a Bayesian Causal Forest with Instrumental Variable (BCF-IV).

Result:

| | Length | Class | Mode |
|---|---|---|---|
| _ci_group_size | 1 | -none- | numeric |
| _num_variables | 1 | -none- | numeric |
| _num_trees | 1 | -none- | numeric |
| _root_nodes | 2000 | -none- | list |
| _child_nodes | 2000 | -none- | list |
| _leaf_samples | 2000 | -none- | list |
| _split_vars | 2000 | -none- | list |
| _split_values | 2000 | -none- | list |
| _drawn_samples | 2000 | -none- | list |
| _send_missing_left | 2000 | -none- | list |
| _pv_values | 2000 | -none- | list |
| _pv_num_types | 1 | -none- | numeric |
| predictions | 11351 | -none- | numeric |
| variance.estimates | 0 | -none- | numeric |
| debiased.error | 11351 | -none- | numeric |
| excess.error | 11351 | -none- | numeric |
| seed | 1 | -none- | numeric |
| ci.group.size | 1 | -none- | numeric |
| X.orig | 2565326 | -none- | numeric |
| Y.orig | 11351 | -none- | numeric |
| W.orig | 11351 | -none- | numeric |
| Y.hat | 11351 | -none- | numeric |
| W.hat | 11351 | -none- | numeric |
| clusters | 0 | -none- | numeric |
| equalize.cluster.weights | 1 | -none- | logical |
| tunable.params | 7 | -none- | list |
| has.missing.values | 1 | -none- | logical |

The causal forest model, comprising 2000 trees, provides a robust mechanism to analyze the heterogeneous treatment effects of `AMT_ANNUITY` on `TARGET`. By leveraging a large section of decision trees, the model offers detailed insights into how different individuals' probabilities of default vary with changes in their annuity payments.

Key Interpretations:

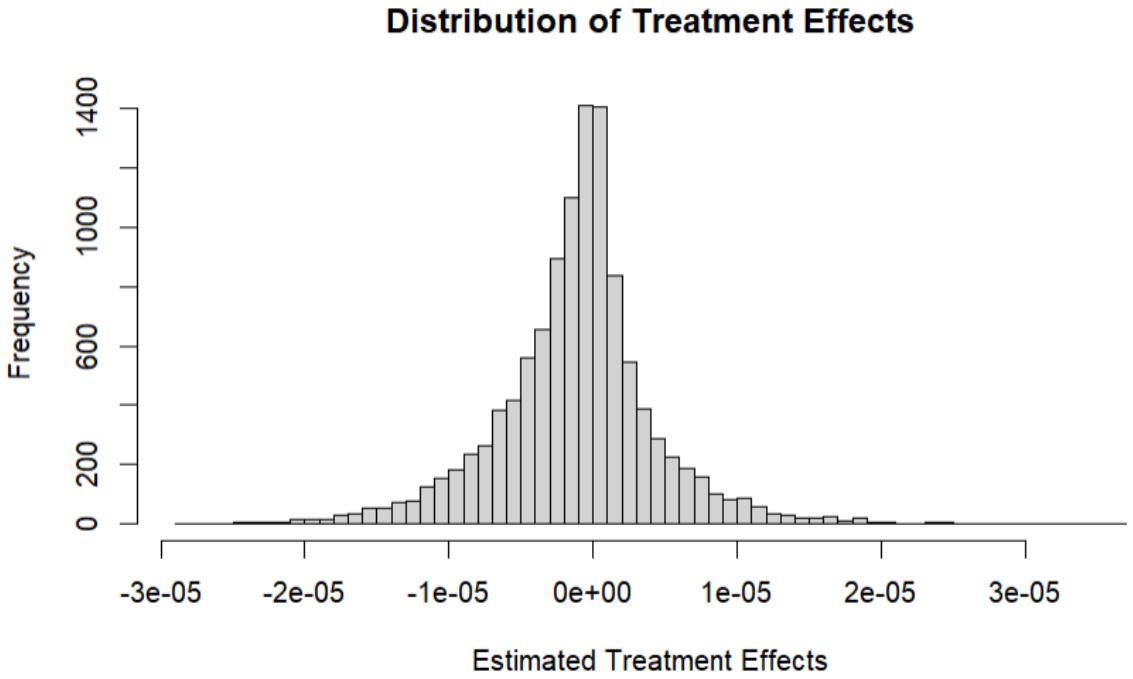
- Predictions: The model generates individualized treatment effect predictions for each of the 11,351 observations, offering a more detailed view of how `AMT_ANNUITY` interacts with the probability of individuals defaulting.
- Debiased Error: The debiased error estimates adjust for potential biases in treatment effect estimation, enhancing the accuracy of the predictions.
- Excess Error: These estimates account for additional uncertainty, providing a conservative measure of the treatment effect's variability.

Presence of Heterogeneous Treatment Effects:
Notably, the analysis reveals that certain subgroups, such as individuals with varying income levels or differing numbers of children, exhibit distinct responses to changes in `AMT_ANNUITY`. This indicates that targeted treatments could be more effective for specific segments of the population.

This implies that banks that provide loans for credit cards, could consider offering customized financial products or support for individuals that are more sensitive to changes in annuity payments, reducing the likelihood of defaulting.

## Distribution of Treatment Effects

Results:




Average Treatment Effect (ATE):

```
      estimate       std.err
-8.460701e-06  4.886466e-06
```

Interpretation:
The distribution of Treatment Effects has a general central tendency, showcasing that the majority of the treatment effect estimates are centered around zero, indicating that most of the treatment effects are close to zero. This suggests that the treatment of `AMT_ANNUITY` has little to no effect on the outcome `TARGET`. This makes sense as a unit (dollar) increase in `AMT_ANNUITY` should see little to no difference in whether an individual is more likely to commit fraud as the percentage change of a dollar is negligible compared to the average `AMT_ANNUITY`

The distribution appears to be approximately symmetric and bell-shaped, similar to a normal distribution (based on the Central Limit Theorem). This adheres to our understanding as it implies that positive and negative treatment effects are distributed evenly around the mean

The spread, however, is more skewed negatively, which means higher `AMT_ANNUITY` is associated with a lower probability of default. Although this may be counterintuitive at first glance, it could be a result of a preventive effect of creditworthiness, where higher annuity payments are dished out by those with higher overall income, allowing them to be able to more reliably pay off their loans.

# Counterfactual Outcome

Results:

| Actual <dbl> | Counterfactual <dbl> | AMT_INCOME_TOTAL <dbl> | AMT_ANNUITY <dbl> |
|---|---|---|---|
| 1 | 1.03306835 | 225000 | 31032.0 |
| 1 | 0.76935434 | 99000 | 9000.0 |
| 1 | 0.86061183 | 450000 | 47650.5 |
| 0 | 0.08243263 | 103500 | 24435.0 |
| 0 | -0.22776101 | 202500 | 16789.5 |
| 0 | -0.16330096 | 175500 | 35568.0 |

The values under the Counterfactual Column are the probability of increase or decrease (depending on the sign) of the individual if their `AMT_ANNUITY` were to double.

Generally, among these outcomes, an increase in double the `AMT_ANNUITY` would lead to an increase in the probability of default. This could be explained by the doubling of `AMT_ANNUITY` leading to a significant difference in repayment and lower disposable income, resulting in an overall increase in probability.

Observation 1 has a counterfactual probability of above 1 while the actual is already 1, this indicates that the individual may be an always-taker, continually intending to default.

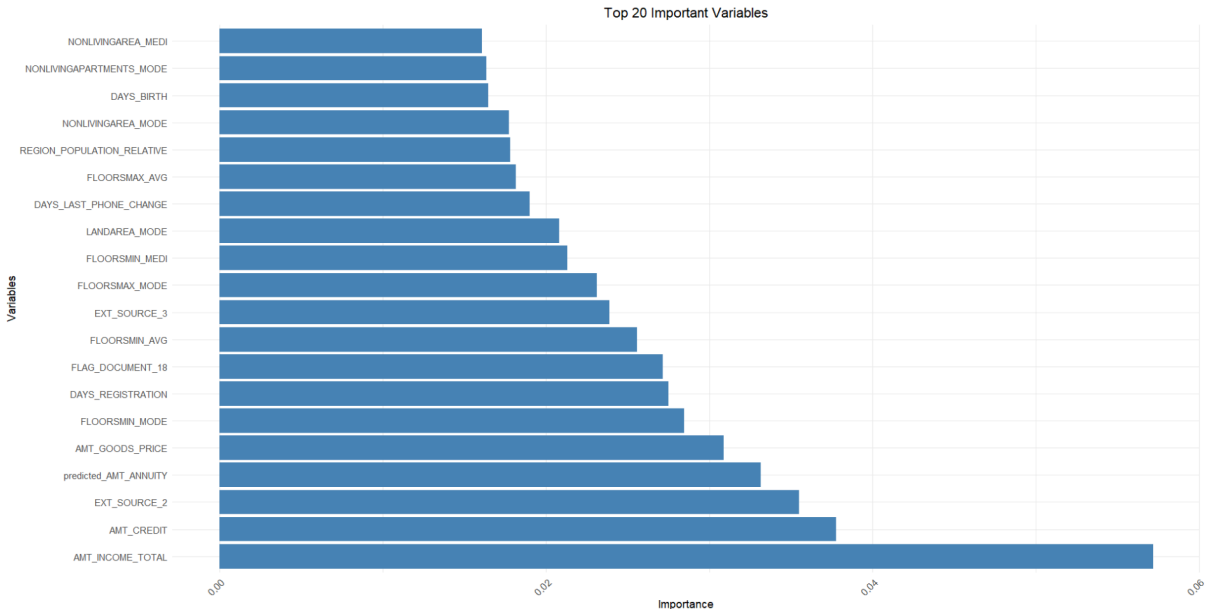
Observations 5 and 6 on the other hand, have a negative counterfactual probability, indicating that they are less likely to default despite already not defaulting, indicating that they are a never-taker (assuming Monotonicity).

# Subgrouping and Variable Importance

Results:

| | Variable <chr> | Importance <dbl> |
|---|---|---|
| 8 | AMT_INCOME_TOTAL | 0.05718735 |
| 9 | AMT_CREDIT | 0.03775413 |
| 136 | EXT_SOURCE_2 | 0.03551784 |
| 226 | predicted_AMT_ANNUITY | 0.03315967 |
| 10 | AMT_GOODS_PRICE | 0.03088516 |
| 160 | FLOORSMIN_MODE | 0.02846876 |
| 37 | DAYS_REGISTRATION | 0.02751792 |
| 216 | FLAG_DOCUMENT_18 | 0.02714404 |
| 146 | FLOORSMIN_AVG | 0.02556575 |
| 137 | EXT_SOURCE_3 | 0.02388212 |
| 159 | FLOORSMAX_MODE | 0.02313266 |
| 174 | FLOORSMIN_MEDI | 0.02130100 |
| 161 | LANDAREA_MODE | 0.02081181 |
| 200 | DAYS_LAST_PHONE_CHANGE | 0.01898624 |
| 145 | FLOORSMAX_AVG | 0.01814142 |
| 34 | REGION_POPULATION_RELATIVE | 0.01781504 |
| 165 | NONLIVINGAREA_MODE | 0.01775037 |
| 35 | DAYS_BIRTH | 0.01648082 |
| 164 | NONLIVINGAPARTMENTS_MODE | 0.01633933 |
| 179 | NONLIVINGAREA_MEDI | 0.01608064 |

Plot Graph:




Interpretation:
The variable importance scores in the causal forest model reflect the significance of each variable in determining the treatment effect heterogeneity, but they do not directly indicate the probability of committing fraud. Rather, they show how influential each variable is in the context of the treatment effect analysis.

The Variable Importance Scores indicate how frequently and effectively each variable is used to split the data within the trees of the causal forest, where higher scores imply that the variable is more crucial in creating partitions that help estimate treatment effects.

Variables like `AMT_INCOME_TOTAL`, `AMT_CREDIT`, and `EXT_SOURCE_2` are essential in explaining how the treatment effect varies across different subgroups.
These variables indicate how changes in `AMT_ANNUITY` impact the likelihood of fraud or default, providing insights into different segments of the population.

By identifying the key factors that drive treatment effect heterogeneity, financial institutions can develop tailored interventions for different subgroups, improving outcomes and reducing default rates.

# Conclusion

Overall, this project aimed to uncover the factors influencing credit card payment defaults and assess the causal relationship between `AMT_ANNUITY` and the likelihood of default, `TARGET`. Through econometrics modeling, IV analysis, and machine learning techniques like causal forests, we are able to generate various insights uncovered throughout the project.

While `AMT_ANNUITY` has a minimal marginal effect on default probability, its interactions with other attributes, especially when doubled or increased significantly, reveal significant heterogeneity. Financial institutions such as banks could leverage these insights to:

- Develop targeted credit risk mitigation strategies, tailoring products to sensitive subgroups.
- Enhance KYC processes by prioritizing critical variables identified in the analysis.
- Use advanced models like causal forests to gain nuanced insights into customer behavior and improve decision-making.

By combining econometric rigor with machine learning, this project highlights the potential to improve credit card fraud detection and loan management strategies, ensuring both financial stability for customers and reduced risk for institutions.