# HOME CREDIT SCORECARD MODEL

Mochammad Rendra Putra Pratama

Link to Github repo

# PROBLEM RESEARCH

Home Credit saat ini sedang menggunakan berbagai macam metode statistik dan Machine Learning untuk membuat prediksi skor kredit. Sekarang, kami meminta anda untuk membuka potensi maksimal dari data kami. Dengan melakukannya, kita dapat memastikan pelanggan yang mampu melakukan pelunasan tidak ditolak ketika melakukan pengajuan pinjaman, dan pinjaman datap diberikan dengan principal, maturity, dan repayment calendar yang akan memotivasi pelanggan untuk sukses. Evaluasi akan dilakukan dengan mengecek seberapa dalam pemahaman analisa yang anda kerjakan. Sebagai catatan, anda perlu menggunakan setidaknya 2 model Machine Learning dimana salah satunya adalah Logistic Regression.

# DATA PRE-PROCESSING

```python
def imputation(df):
  for cat in df.describe(include='object').columns:
    df[cat].fillna(df[cat].mode(), inplace=True)

  for num in df.describe().columns:
    df[num].fillna(df[num].median(), inplace=True)
```

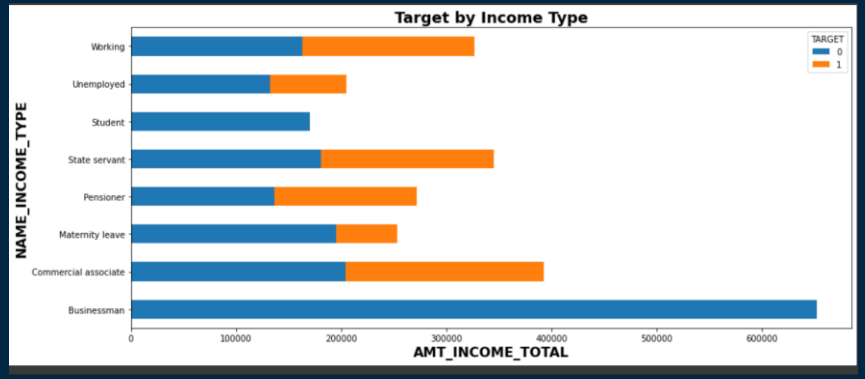
```python
def group_income(x):
  if x < 112500.0:
    return '<24%'
  elif x < 147150.0:
    return '25-49%'
  elif x < 202500.0:
    return '50-74%'
  else:
    return '75%>'
```

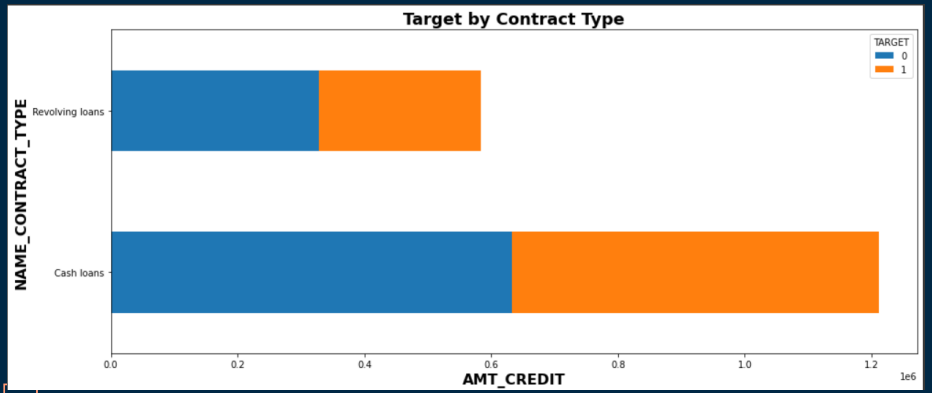
```python
def group_credit(x):
  if x < 270000.0:
    return '<24%'
  elif x < 513531.0:
    return '25-49%'
  elif x < 808650.0:
    return '50-74%'
  else:
    return '75%>'
```

```python
def feature_engineering(df):
  df['HAS_CHILDREN'] = df['CNT_CHILDREN'].apply(lambda x: True if x>0 else False)
  df['INCOME_GROUP'] = df['AMT_INCOME_TOTAL'].apply(lambda x: group_income(x))
  df['CREDIT_GROUP'] = df['AMT_CREDIT'].apply(lambda x: group_credit(x))
  df['AGE'] = (df['DAYS_BIRTH'])/(-365)
  df['EMPLOYED_YEARS'] = (df['DAYS_EMPLOYED'])/(-365)
  df['YEARS_SINCE_REGISTERED'] = (df['DAYS_REGISTRATION'])/(-365)
```

Dilakukan imputasi untuk data yang kosong. Data kategorikal diisi dengan mode dan data numerik diisi dengan median. Feature engineering membuat feature baru menggunakan feature yang sudah ada.

# DATA VISUALIZATION AND BUSINESS INSIGHT




Berdasarkan pekerjaan dan pendapatan, Businessman dari berbagai macam pendapaptan tidak memiliki masalah dalam pelunasan. Sedangkan berdasarkan jenis kontrak dan banyaknya pinjaman, rata-rata banyaknya pinjaman yang lebih rendah lebih banyak memiliki masalah dalam pelunasan.

# FEATURE SELECTION

```python
def feature_selection(df):
  df.drop(columns=['DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION'])

  corr_matrix = train.corr().abs()

  upper = corr_matrix.where((np.triu(np.ones(corr_matrix.shape), k=1) + np.tril(np.ones(corr_matrix.shape), k=-1)).astype(bool))

  to_drop = [column for column in upper.columns if any(upper[column] > 0.75)]

  df.drop(to_drop, axis=1, inplace=True)
```

Karena banyaknya kolom, dibuat fungsi untuk membuang feature yang digunakan saat feature engineering dan feature-feature yang memiliki correlation yang lebih tinggi dari 0.75.

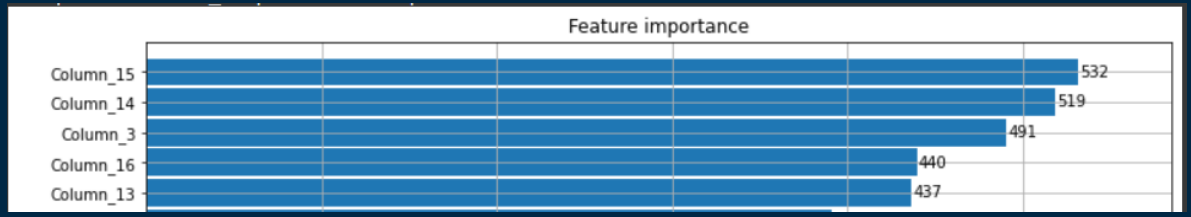
# MACHINE LEARNING IMPLEMENTATION AND EVALUATION

```python
for clf, model in fit_models.items():
    y_pred = model.predict(X_valid)
    accuracy = accuracy_score(y_valid, y_pred)
    precision = precision_score(y_valid, y_pred)
    recall = recall_score(y_valid, y_pred)
    f1 = f1_score(y_valid, y_pred)
    print(f'{clf}: Accuracy- {accuracy}, Recall- {recall}, Precision- {precision}, F1- {f1}')
```

```
lr: Accuracy- 0.9193697868396663, Recall- 0.011921600323297636, Precision- 0.4609375, F1- 0.023242072089816823
lgbm: Accuracy- 0.9194185649480513, Recall- 0.010709234188724995, Precision- 0.4690265486725664, F1- 0.020940339786645595
```

Model ditrain menggunakan Logistic Regression dan Light Gradient Boosting Machine. Berdasarkan evaluasi, Light GBM memiliki accuracy yang lebih baik oleh karena itu Light GBM digunakan untuk model prediksi ini.

# FEATURE IMPORTANCE




Berdasarkan feature importance, EXT_SOURCE_3, EXT_SOURCE_2, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE, dan EXT_SOURCE_1 adalah 5 feture paling penting.

| | EXT_SOURCE_3 | EXT_SOURCE_2 | DAYS_ID_PUBLISH | DAYS_LAST_PHONE_CHANGE | EXT_SOURCE_1 |
|---|---|---|---|---|---|
| 0 | 0.139376 | 0.262949 | -2120 | -1134.0 | 0.083037 |
| 1 | 0.535276 | 0.622246 | -291 | -828.0 | 0.311267 |
| 2 | 0.729567 | 0.555912 | -2531 | -815.0 | 0.505998 |
| 3 | 0.535276 | 0.650442 | -2437 | -617.0 | 0.505998 |
| 4 | 0.535276 | 0.322738 | -3458 | -1106.0 | 0.505998 |

# BUSINESS RECOMMENDATION

- Normalisasi score dari data eksternal, hari sejak client melakukan pengubahan identitas untuk pengajuan, dan hari sejak client mengubah nomor telepon menjadi factor paling penting dalam menentukan apakah client akan memiliki masalah dalam pembayaran.

- Normalisasi score dari data eksternal bisa menjadi acuan utama untuk menerima atau menolak pengajuan credit client untuk mengatasi banyaknya client yang bermasalah dalam pembayaran.