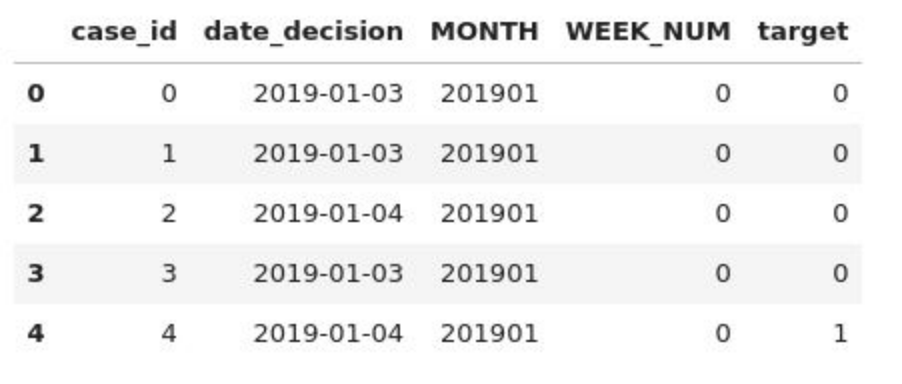
Supervised Learning Final Exam

Kaggle : Home Credit - Credit Risk Model Stability

- On-going Kaggle competition
- Goal is to predict client likely to default on their loans using bank data

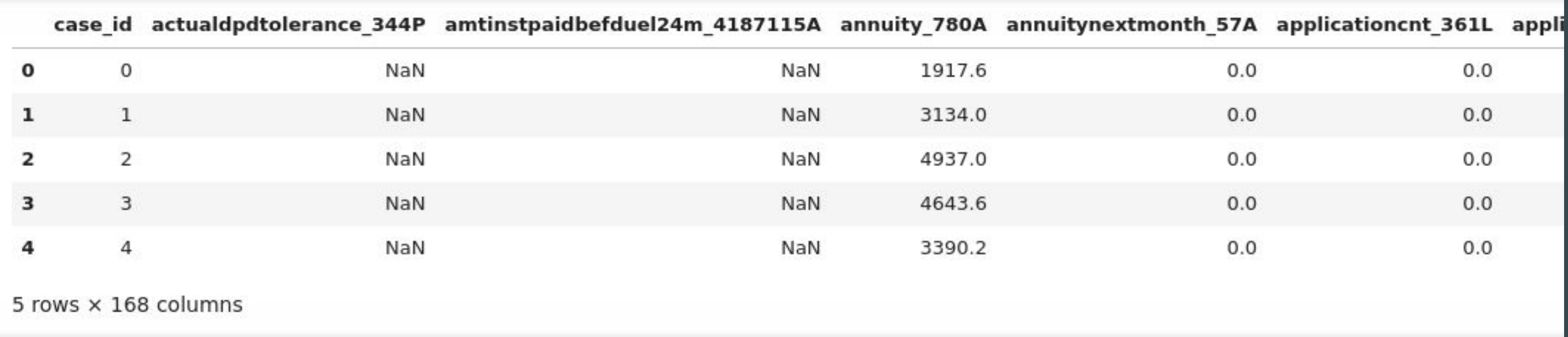
- Challenges :
  - Large amount of data (several datasets, 400+ columns, different granularity)
  - Limited amount of computing resources
  - Very imbalance classes

What’s the data look like ?



	case_id	date_decision	MONTH	WEEK_NUM	target
0	0	2019-01-03	201901	0	0
1	1	2019-01-03	201901	0	0
2	2	2019-01-04	201901	0	0
3	3	2019-01-03	201901	0	0
4	4	2019-01-04	201901	0	1

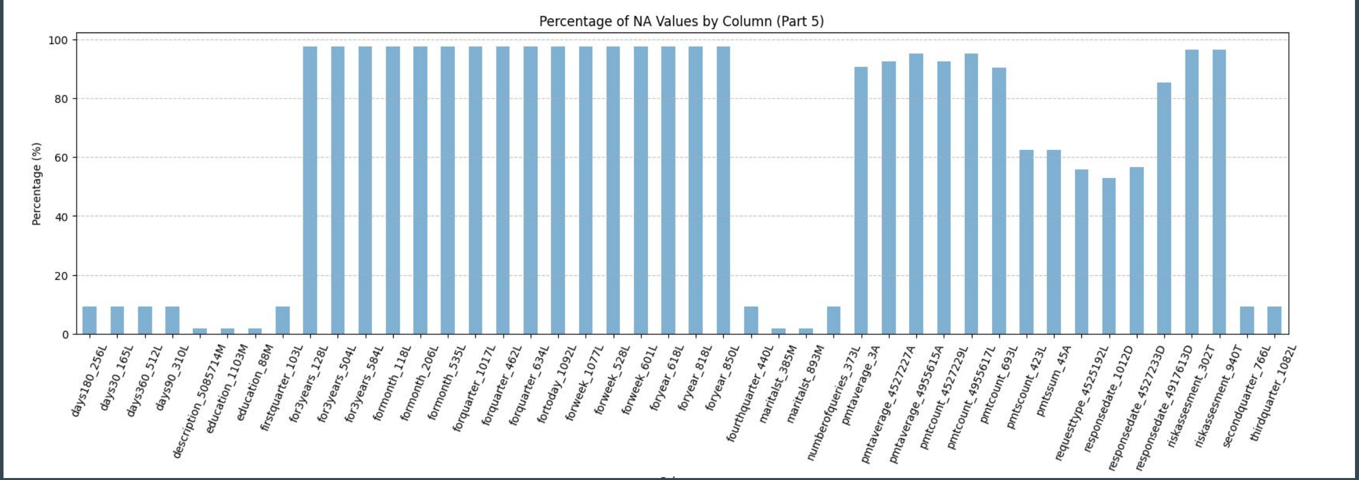
What’s the data look like ?



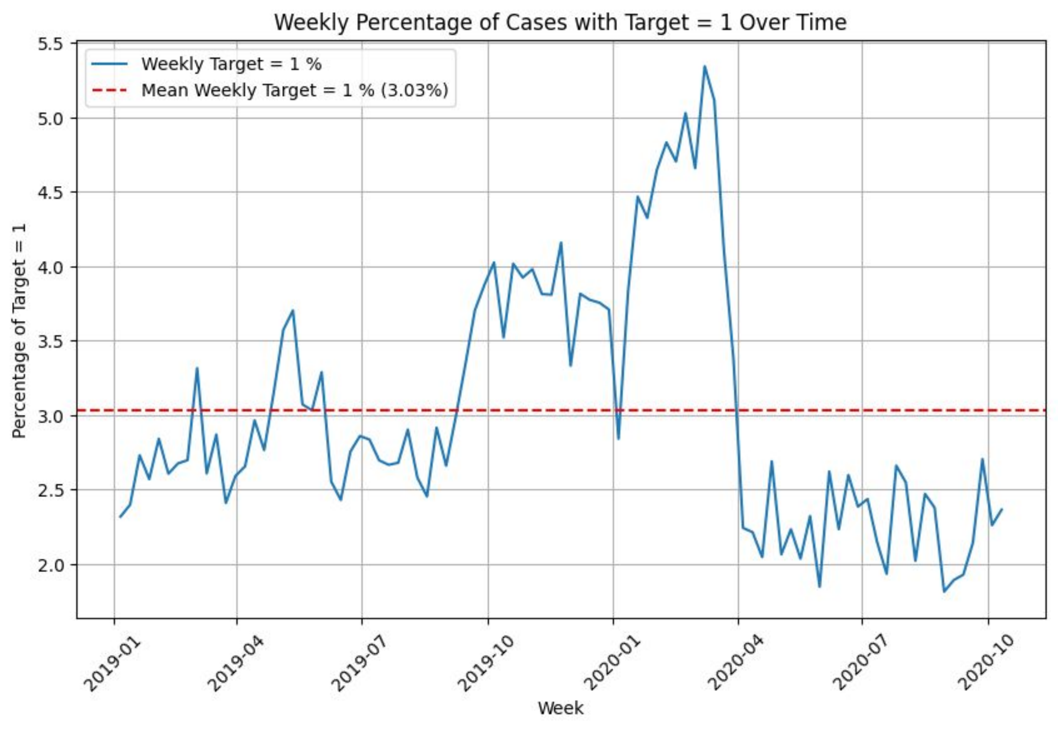
	case_id	actualdpdtolerance_344P	amtinstpaidbefduel24m_4187115A	annuity_780A	annuitynextmonth_57A	applicationcnt_361L	appli
0	0	NaN	NaN	1917.6	0.0	0.0	
1	1	NaN	NaN	3134.0	0.0	0.0	
2	2	NaN	NaN	4937.0	0.0	0.0	
3	3	NaN	NaN	4643.6	0.0	0.0	
4	4	NaN	NaN	3390.2	0.0	0.0	

5 rows × 168 columns

Data quality problem...



Class imbalance



Features selection

- We are still left with 46 columns (7 categorical, with a high number of possible values)

- To fit the computing resources constraint we need to reduce this number

Features selection

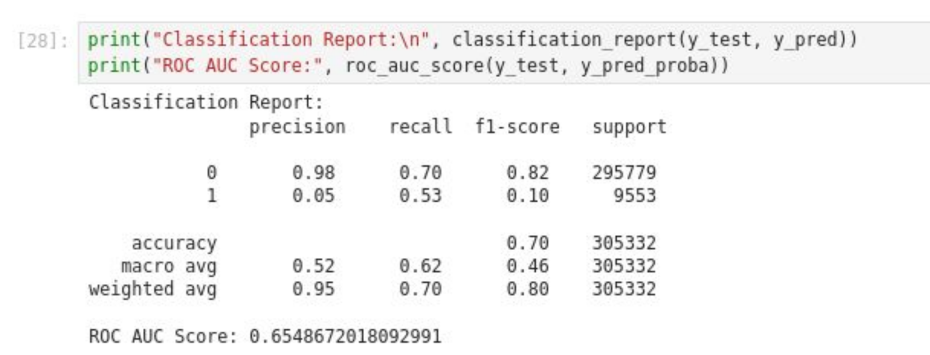
- We have use the Select K Best features approach
  - ANOVA test for continuous variables
  - Chi-square test for categorical

- Main idea is to measure the statistical significance of the difference in each features between the two class, and keep the most significant

Modelling

- As a first iteration, we use a simple Logistic Regression model to get a baseline

Results



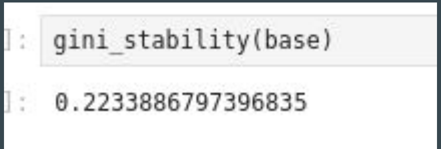
```
[28]: print("Classification Report:\n", classification_report(y_test, y_pred))
      print("ROC AUC Score:", roc_auc_score(y_test, y_pred_proba))
```

```
Classification Report:
               precision    recall  f1-score   support

           0       0.98      0.70      0.82    295779
           1       0.05      0.53      0.10      9553

    accuracy                           0.70    305332
   macro avg       0.52      0.62      0.46    305332
weighted avg       0.95      0.70      0.80    305332

ROC AUC Score: 0.6548672018092991
```



```
]: gini_stability(base)
```

```
]: 0.2233886797396835
```

Results

- Good start but not satisfying
- Top public leaderboard score are much higher, usage in production is not possible

- Areas of improvement
  - More data
  - Explore specific techniques imbalanced class (sampling, others algorithms)
  - Tune logistic regression hyperparameters
  - Try other learning algorithms
  - Handle NaN values differently (inferring missing values, etc)
  - Improve features and features selections method