FP_Group9_HCDR_P1

April 1, 2024

1 FP Phase 1 - Home Credit Default Risk

Spring 2024

Team Members: - Glen Colletti - Alex Bordanca - Paul Miller

1.1 Abstract

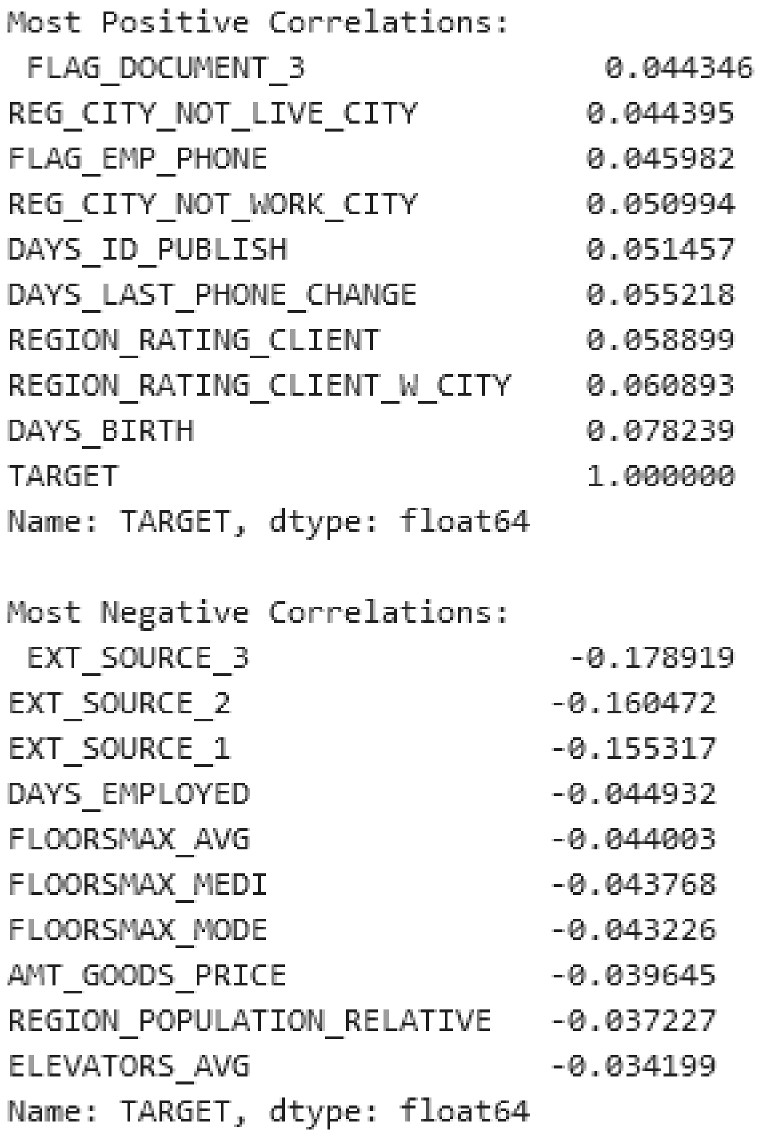
> "The project is based on the Home Credit Default Risk (HCDR) Kaggle Competition . The goal of the competition is to predict whether or not a client will repay a loan. In order to make sure that people who struggle to get loans due to insufficient or non-existent credit histories have a positive loan experience, Home Credit makes use of a variety of alternative data–including telco and transactional information–to predict their clients' repayment abilities."

The team will investigate features in the training data and other files provided for additional insight for reliable predictors of repayment. Feature selection will be conducted, investigation will include forward feature selection and backward selection. Pipelines will be implemented to standardize inputs of both categorical and numerical features. Optimal hyperparameters will be determined for model fitting. The data will be fit using a variety of machine learning algorithms to include XG_Boost, Logistic Regression, and KNN classifier.

1.2 Data Description

The HCDR dataset contains 122 feature columns and about 307,000 records. The target variable for prediction is called "TARGET", it is a binary vector. The 0 case represents full repayment, while the 1 case represents an unpaid loan.

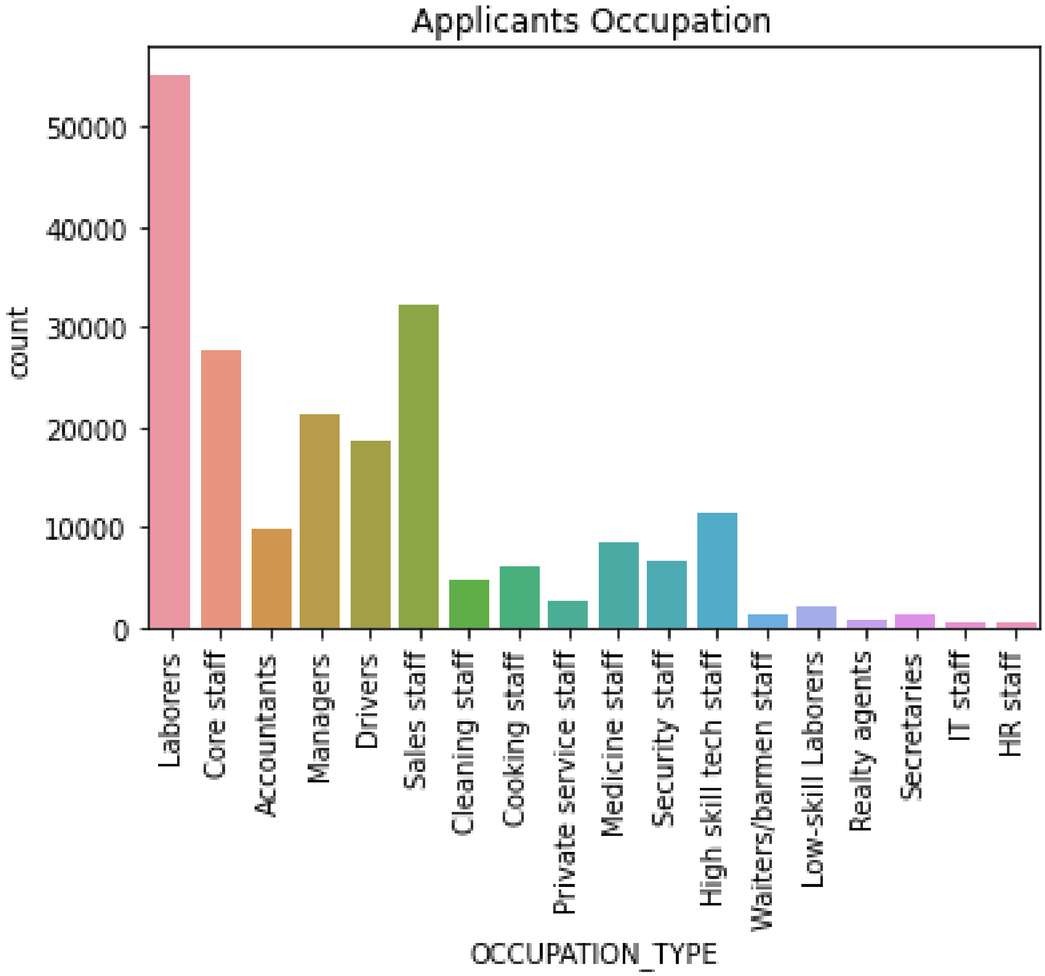
Correlating features to the target variable, we can see that most features have minimal correlations. The top ten positive correlations range from approximately 8% to about 4%. The negative correlations are somewhat stronger ranging from about 18% to 3% absolute value.



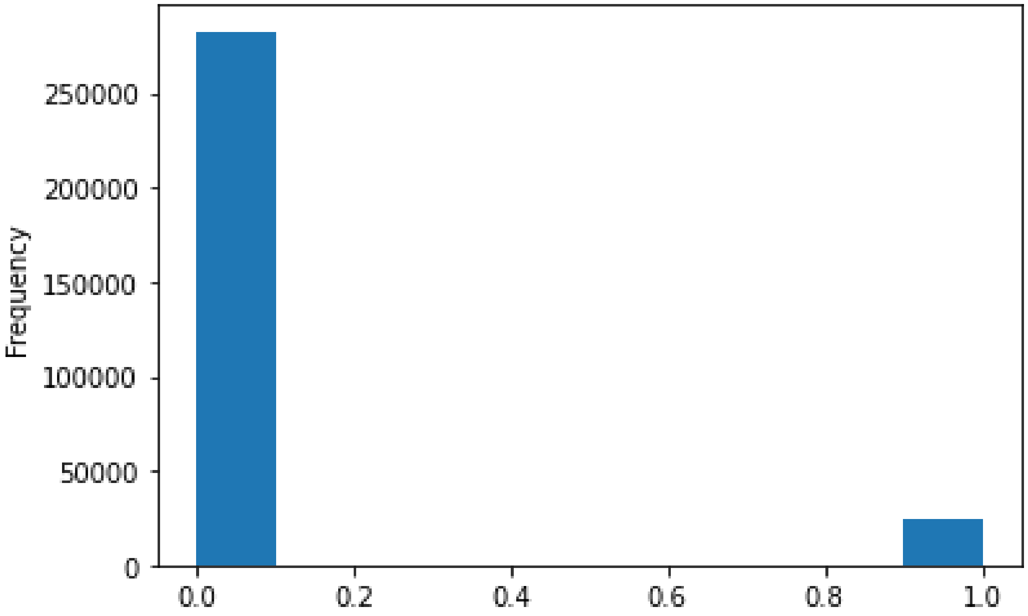
```
Most Positive Correlations:
 FLAG_DOCUMENT_3                0.044346
REG_CITY_NOT_LIVE_CITY         0.044395
FLAG_EMP_PHONE                 0.045982
REG_CITY_NOT_WORK_CITY         0.050994
DAYS_ID_PUBLISH                0.051457
DAYS_LAST_PHONE_CHANGE         0.055218
REGION_RATING_CLIENT           0.058899
REGION_RATING_CLIENT_W_CITY    0.060893
DAYS_BIRTH                     0.078239
TARGET                         1.000000
Name: TARGET, dtype: float64

Most Negative Correlations:
 EXT_SOURCE_3                  -0.178919
EXT_SOURCE_2                  -0.160472
EXT_SOURCE_1                  -0.155317
DAYS_EMPLOYED                 -0.044932
FLOORSMAX_AVG                 -0.044003
FLOORSMAX_MEDI                -0.043768
FLOORSMAX_MODE                -0.043226
AMT_GOODS_PRICE               -0.039645
REGION_POPULATION_RELATIVE    -0.037227
ELEVATORS_AVG                 -0.034199
Name: TARGET, dtype: float64
```

By occupation the majority of applicants are laborers, sales staff, or core staff.



Most of the loans let in the training data were repaid.



Of the 122 features, 58 were missing 12% or more of their data. The remaining 64 features have less than 1% missing data. The largest share of features missing maxed out at about 69%.

1.3 Machine Algorithms and Metrics

1.3.1 Models

- **Logistic Regression**: A linear model for binary classification that predicts the probability a datapoint belongs to a particular class. It is implemented with the sigmoid function which forces the output to an interval of [0,1], thus representing the probability of the positive class. The loss function use is log loss/cross entropy loss, which analyzes the difference between the predicted probabilities and the actual binary class, penalizing wrongly classified predictions that have high predicted probability.
- **XGBoost**: XGBoost uses a gradient boosted framework, fitting each new tree to the negative gradient of the loss function (logistic loss, in our case with and without regularization, both L1 and L2) of the entire ensemble. The parallelization works to prevent overfitting.
- **KNNClassifier**: Predicts the class of a particular datapoint based on the majority class of the K nearest neighbors of the point in feature space. It doesn't work directly to optimize a loss function, with its performance instead being evaluated and optimized using metrics like accuracy, F-1, etc.

1.3.2 Metrics

- **F-1 Score**: The harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = 2 \cdot \frac{tp}{2tp + fp + fn}$$

- **RMSE (Root Mean Squared Error)**: The square root of the mean of the squared differences between observed values and predicted values, i.e., residuals.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{y})^2}$$

- **R2 Score (Coefficient of Determination)**: The proportion of the variation of the dependent variable that is explained by or predictable from the independent variable(s).

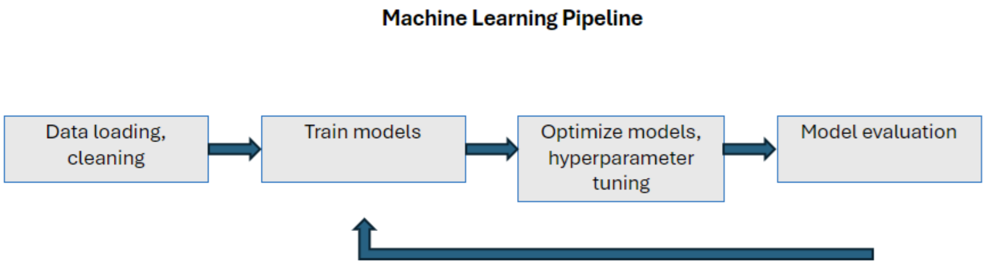
$$R^2 = 1 - \frac{\sum(y_{\text{true}} - y_{\text{pred}})^2}{\sum(y_i - \bar{y})^2}$$

- **Accuracy**: Defines how often a model correctly predicts the outcome.

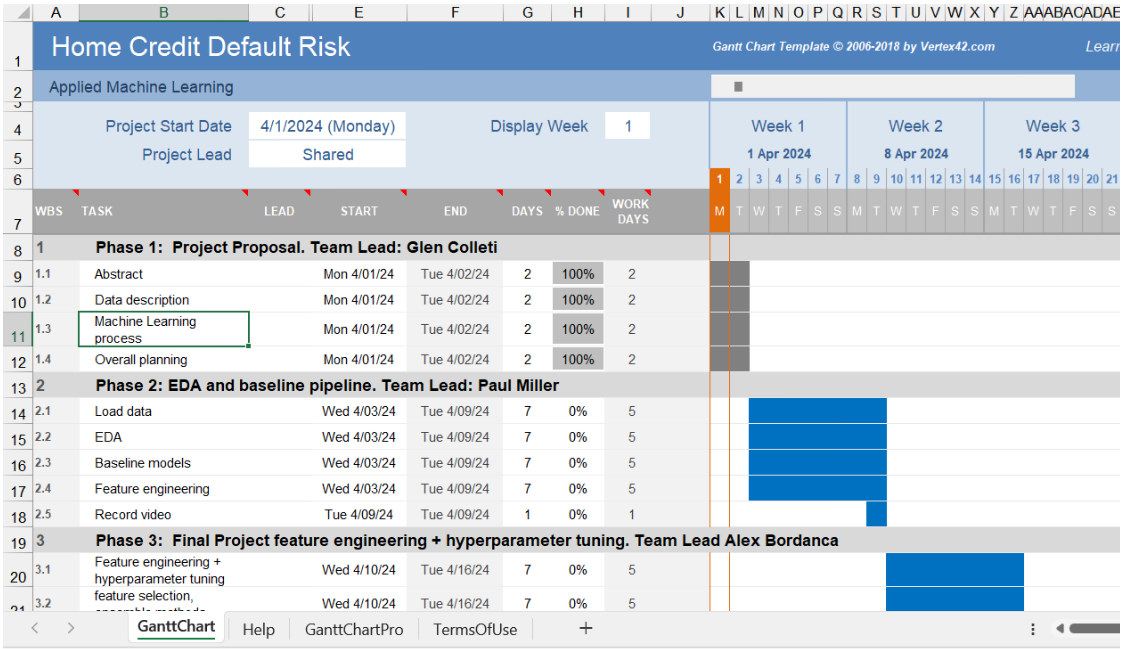
$$\text{acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

1.4 Pipelines

We plan to implement the following steps as described by this block diagram:


Machine Learning Pipeline
Data loading, cleaning
Train models
Optimize models, hyperparameter tuning
Model evaluation

1.5 Gantt Chart & Phase Leader Plan



Home Credit Default Risk

Gantt Chart Template © 2006-2018 by Vertex42.com

Applied Machine Learning

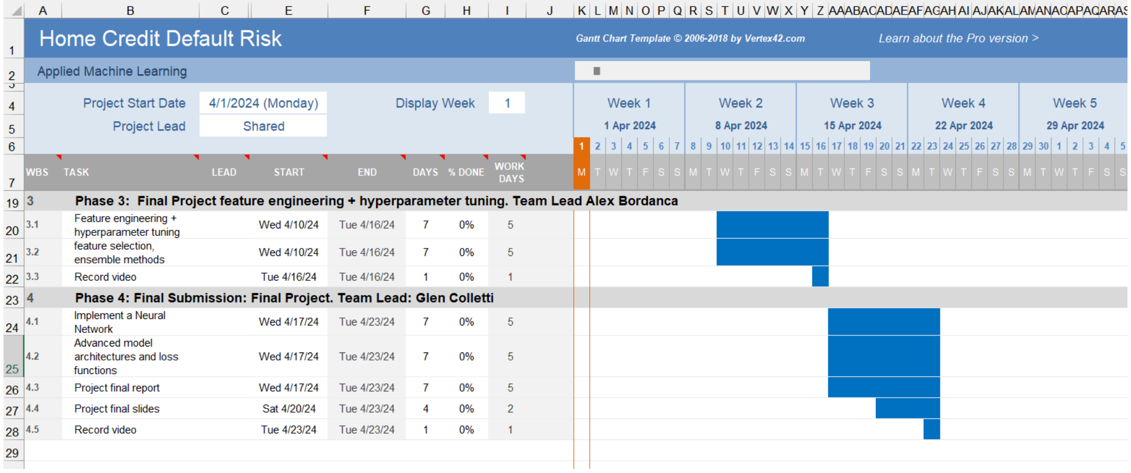
Project Start Date: 4/1/2024 (Monday)

Project Lead: Shared

Display Week: 1

Week 1 — 1 Apr 2024; Week 2 — 8 Apr 2024; Week 3 — 15 Apr 2024

WBS	TASK	LEAD	START	END	DAYS	% DONE	WORK DAYS
1	**Phase 1: Project Proposal. Team Lead: Glen Colleti**						
1.1	Abstract		Mon 4/01/24	Tue 4/02/24	2	100%	2
1.2	Data description		Mon 4/01/24	Tue 4/02/24	2	100%	2
1.3	Machine Learning process		Mon 4/01/24	Tue 4/02/24	2	100%	2
1.4	Overall planning		Mon 4/01/24	Tue 4/02/24	2	100%	2
2	**Phase 2: EDA and baseline pipeline. Team Lead: Paul Miller**						
2.1	Load data		Wed 4/03/24	Tue 4/09/24	7	0%	5
2.2	EDA		Wed 4/03/24	Tue 4/09/24	7	0%	5
2.3	Baseline models		Wed 4/03/24	Tue 4/09/24	7	0%	5
2.4	Feature engineering		Wed 4/03/24	Tue 4/09/24	7	0%	5
2.5	Record video		Tue 4/09/24	Tue 4/09/24	1	0%	1
3	**Phase 3: Final Project feature engineering + hyperparameter tuning. Team Lead Alex Bordanca**						
3.1	Feature engineering + hyperparameter tuning		Wed 4/10/24	Tue 4/16/24	7	0%	5
3.2	feature selection,		Wed 4/10/24	Tue 4/16/24	7	0%	5



Home Credit Default Risk

Gantt Chart Template © 2006-2018 by Vertex42.com

Learn about the Pro version >

Applied Machine Learning

Project Start Date: 4/1/2024 (Monday)

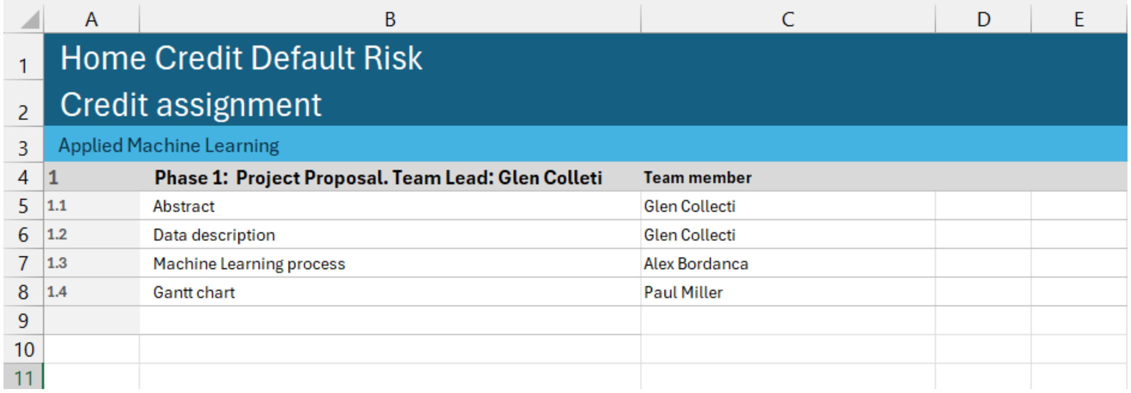
Project Lead: Shared

Display Week: 1

Week 1 — 1 Apr 2024; Week 2 — 8 Apr 2024; Week 3 — 15 Apr 2024; Week 4 — 22 Apr 2024; Week 5 — 29 Apr 2024

WBS	TASK	LEAD	START	END	DAYS	% DONE	WORK DAYS
3	**Phase 3: Final Project feature engineering + hyperparameter tuning. Team Lead Alex Bordanca**						
3.1	Feature engineering + hyperparameter tuning		Wed 4/10/24	Tue 4/16/24	7	0%	5
3.2	feature selection, ensemble methods		Wed 4/10/24	Tue 4/16/24	7	0%	5
3.3	Record video		Tue 4/16/24	Tue 4/16/24	1	0%	1
4	**Phase 4: Final Submission: Final Project. Team Lead: Glen Colletti**						
4.1	Implement a Neural Network		Wed 4/17/24	Tue 4/23/24	7	0%	5
4.2	Advanced model architectures and loss functions		Wed 4/17/24	Tue 4/23/24	7	0%	5
4.3	Project final report		Wed 4/17/24	Tue 4/23/24	7	0%	5
4.4	Project final slides		Sat 4/20/24	Tue 4/23/24	4	0%	2
4.5	Record video		Tue 4/23/24	Tue 4/23/24	1	0%	1

1.6 Credit Assignment Plan



Home Credit Default Risk

Credit assignment

Applied Machine Learning

1	**Phase 1: Project Proposal. Team Lead: Glen Colleti**	**Team member**
1.1	Abstract	Glen Collecti
1.2	Data description	Glen Collecti
1.3	Machine Learning process	Alex Bordanca
1.4	Gantt chart	Paul Miller

```
[ ]:
```