HOME CREDIT DEFAULT RISK

PROJECT PROPOSAL FIRST PRESENTATION

HONG WEI KOH & SIVA KRISHNA THOTA

© 2021 National University of Singapore. All Rights Reserved

Agenda

- Business problem
- Approach
- Data / Data wrangling
- Exploratory data analysis
- Predictive modeling
- Conclusion
- Future work

© 2021 National University of Singapore. All Rights Reserved

Business Problem

Many people struggle to get loans due to insufficient or non-existent credit histories

This population is often taken advantage by untrustworthy lenders.

Home credit tries to include the unbanked population to support their economic needs.

© 2021 National University of Singapore. All Rights Reserved

Goal

Identify if a new client shows a high risk for loan default

How can this help?

Reduce Uncertainty

Proportional Disbursement

Risk Reduction

© 2021 National University of Singapore. All Rights Reserved

Data Set

The Data is taken from Kaggle's competition publicly available

application_train.csv

- 307511 Records, 122 Columns
- Imbalanced Target Labels.
- Source for training machine Learning models.
- Target Labels :
- 0's – 282686, 1's - 24825

application_test.csv

- 48744 Records, 121 Columns
- No Target Labels.
- Source for testing the performance of machine Learning models.
- Target Labels :
- None – need to predict.

© 2021 National University of Singapore. All Rights Reserved

Understanding Data Set

There are 122 columns with different data types:

- 65 floating point numbers.
- 41 integer numbers – some are numerical categories.
- 16 Strings – Categories involved

© 2021 National University of Singapore. All Rights Reserved

Data Clean

Cleaning steps can be many, relative to the approach taken & judgement calls.

Since the ML model can't inherently deal with text, the data must be converted to appropriate numbers. Any significant distortion/noise in the model must be removed as much as possible.

- Converting string categorical columns into numerical – Label encoding.
- Converting string categorical columns into numerical and adding new columns to indicate the presence of categorical variables – One hot encoding.
- Replacing illogical outliers with empty values (NAN values).
- Imputing empty cells with the median of the values. In some cases, imputation is approached with a certain grouping.
- Dealing with a few anomalies.
- Changing invalid entries into valid entries.

© 2021 National University of Singapore. All Rights Reserved

Train Data Set Top 15 Columns with Missing values

	Missing Count	Missing Count Ratio	Missing Count %
COMMONAREA_MEDI	214865	0.698723	69.9
COMMONAREA_AVG	214865	0.698723	69.9
COMMONAREA_MODE	214865	0.698723	69.9
NONLIVINGAPARTMENTS_MODE	213514	0.694330	69.4
NONLIVINGAPARTMENTS_AVG	213514	0.694330	69.4
NONLIVINGAPARTMENTS_MEDI	213514	0.694330	69.4
FONDKAPREMONT_MODE	210295	0.683862	68.4
LIVINGAPARTMENTS_MODE	210199	0.683550	68.4
LIVINGAPARTMENTS_AVG	210199	0.683550	68.4
LIVINGAPARTMENTS_MEDI	210199	0.683550	68.4
FLOORSMIN_AVG	208642	0.678486	67.8
FLOORSMIN_MODE	208642	0.678486	67.8
FLOORSMIN_MEDI	208642	0.678486	67.8
YEARS_BUILD_MEDI	204488	0.664978	66.5
YEARS_BUILD_MODE	204488	0.664978	66.5

© 2021 National University of Singapore. All Rights Reserved

Test Data Set Top 15 Columns with Missing values

	Missing Count	Missing Count Ratio	Missing Count %
COMMONAREA_AVG	33495	0.687161	68.7
COMMONAREA_MODE	33495	0.687161	68.7
COMMONAREA_MEDI	33495	0.687161	68.7
NONLIVINGAPARTMENTS_AVG	33347	0.684125	68.4
NONLIVINGAPARTMENTS_MODE	33347	0.684125	68.4
NONLIVINGAPARTMENTS_MEDI	33347	0.684125	68.4
FONDKAPREMONT_MODE	32797	0.672842	67.3
LIVINGAPARTMENTS_AVG	32780	0.672493	67.2
LIVINGAPARTMENTS_MODE	32780	0.672493	67.2
LIVINGAPARTMENTS_MEDI	32780	0.672493	67.2
FLOORSMIN_MEDI	32466	0.666051	66.6
FLOORSMIN_AVG	32466	0.666051	66.6
FLOORSMIN_MODE	32466	0.666051	66.6

© 2021 National University of Singapore. All Rights Reserved

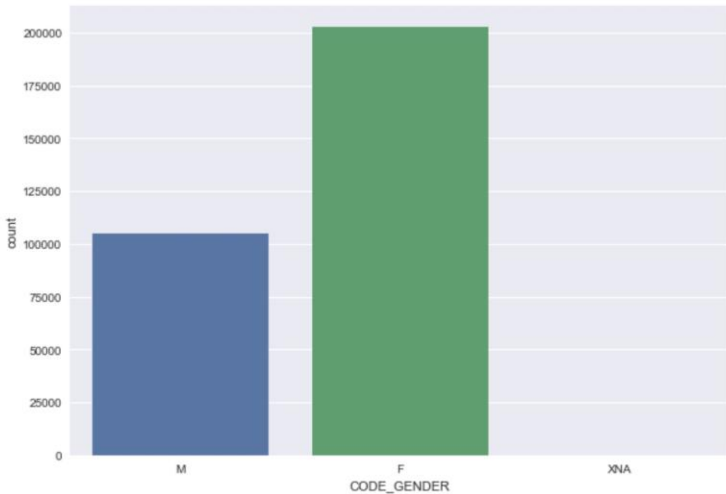
Different kinds of classes in every categorical column

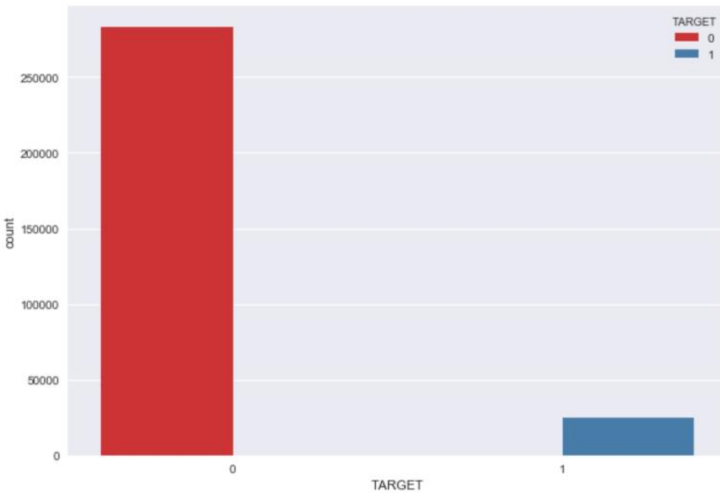
	Categorical Column	Count
0	NAME_CONTRACT_TYPE	2
1	CODE_GENDER	3
2	FLAG_OWN_CAR	2
3	FLAG_OWN_REALTY	2
4	NAME_TYPE_SUITE	7
5	NAME_INCOME_TYPE	8
6	NAME_EDUCATION_TYPE	5
7	NAME_FAMILY_STATUS	6
8	NAME_HOUSING_TYPE	6
9	OCCUPATION_TYPE	18
10	WEEKDAY_APPR_PROCESS_START	7
11	ORGANIZATION_TYPE	58
12	FONDKAPREMONT_MODE	4
13	HOUSETYPE_MODE	3
14	WALLSMATERIAL_MODE	7
15	EMERGENCYSTATE_MODE	2

© 2021 National University of Singapore. All Rights Reserved

Target Column

EDA suggests that most people returned the money





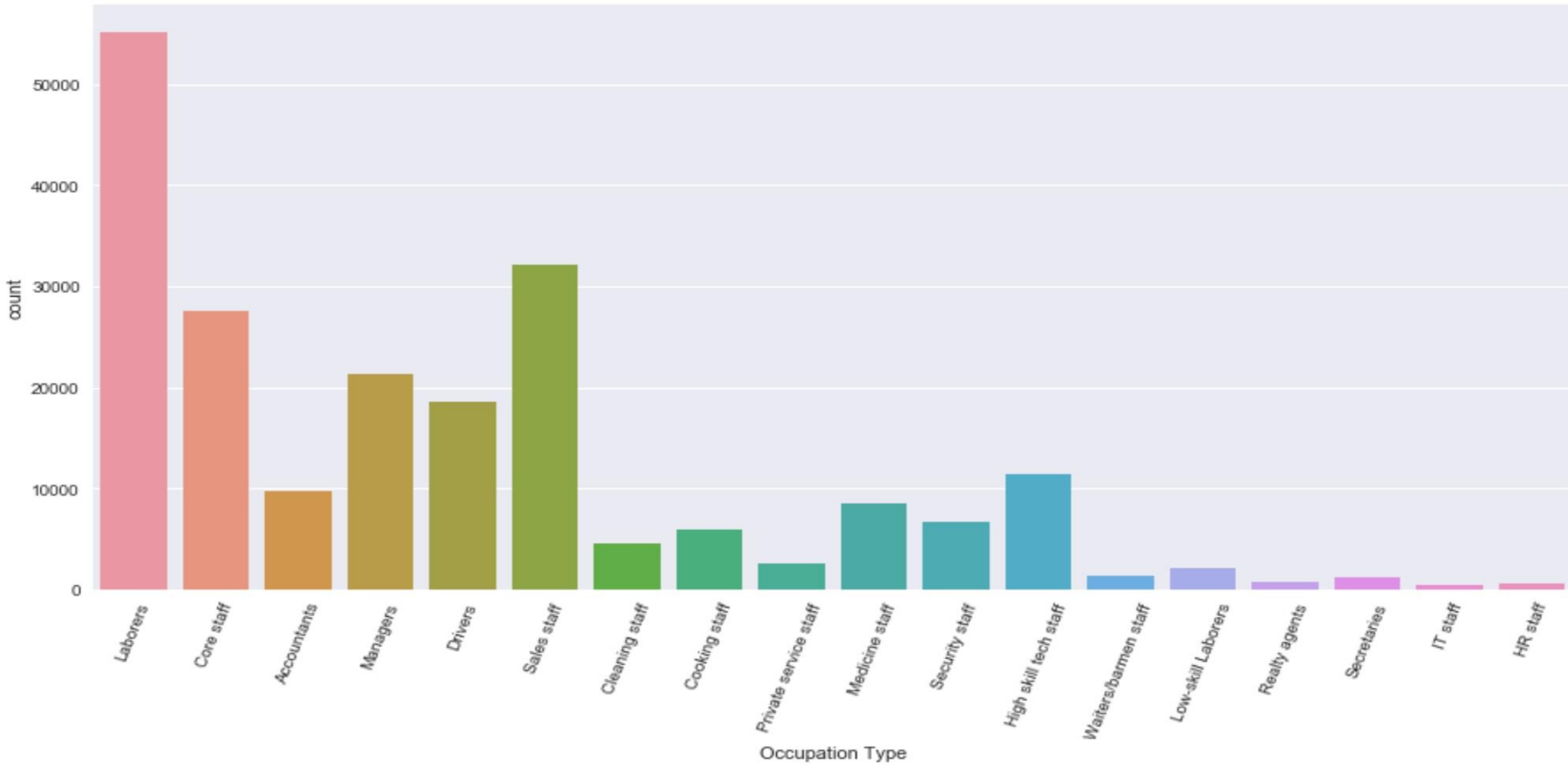
Females are the highest borrowers with counts:
F - 202448
M - 105059
XNA - 4

Most people returned the borrowed money:
0 - 282686
1 - 24825

© 2021 National University of Singapore. All Rights Reserved

Occupation type vs Borrowers

Laborers - occupation type were the most borrowers.



Most of the clients are laborers and the least of the clients are IT Staff.

© 2021 National University of Singapore. All Rights Reserved

ML Models

Predictive Modeling – Outcome of the model is expected to identify the potential that someone will default on a loan

Expected Target Outcome: 0 or 1, 0 – Not a defaulter, 1 – potential defaulter.

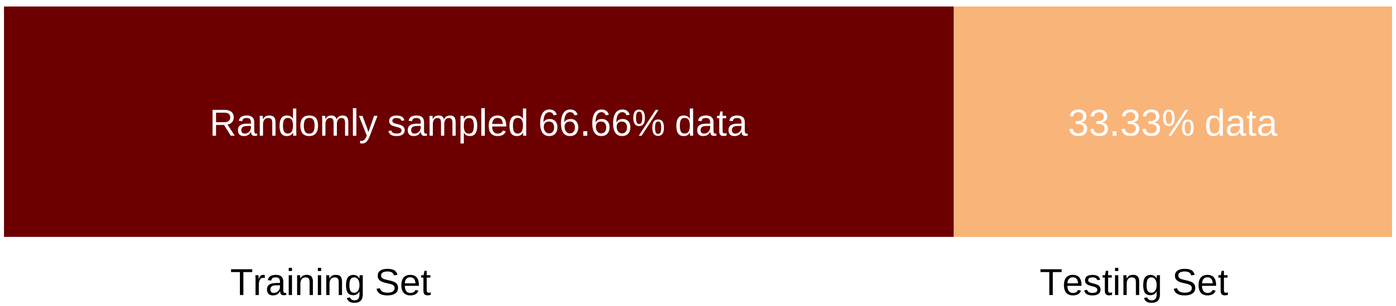
Performance Metrics : Accuracy.

ML Models : Logistic regression, Random forest, XGBoost, LightGBM, Naïve bayes, ensemble.

© 2021 National University of Singapore. All Rights Reserved

Data Partition and Preparation

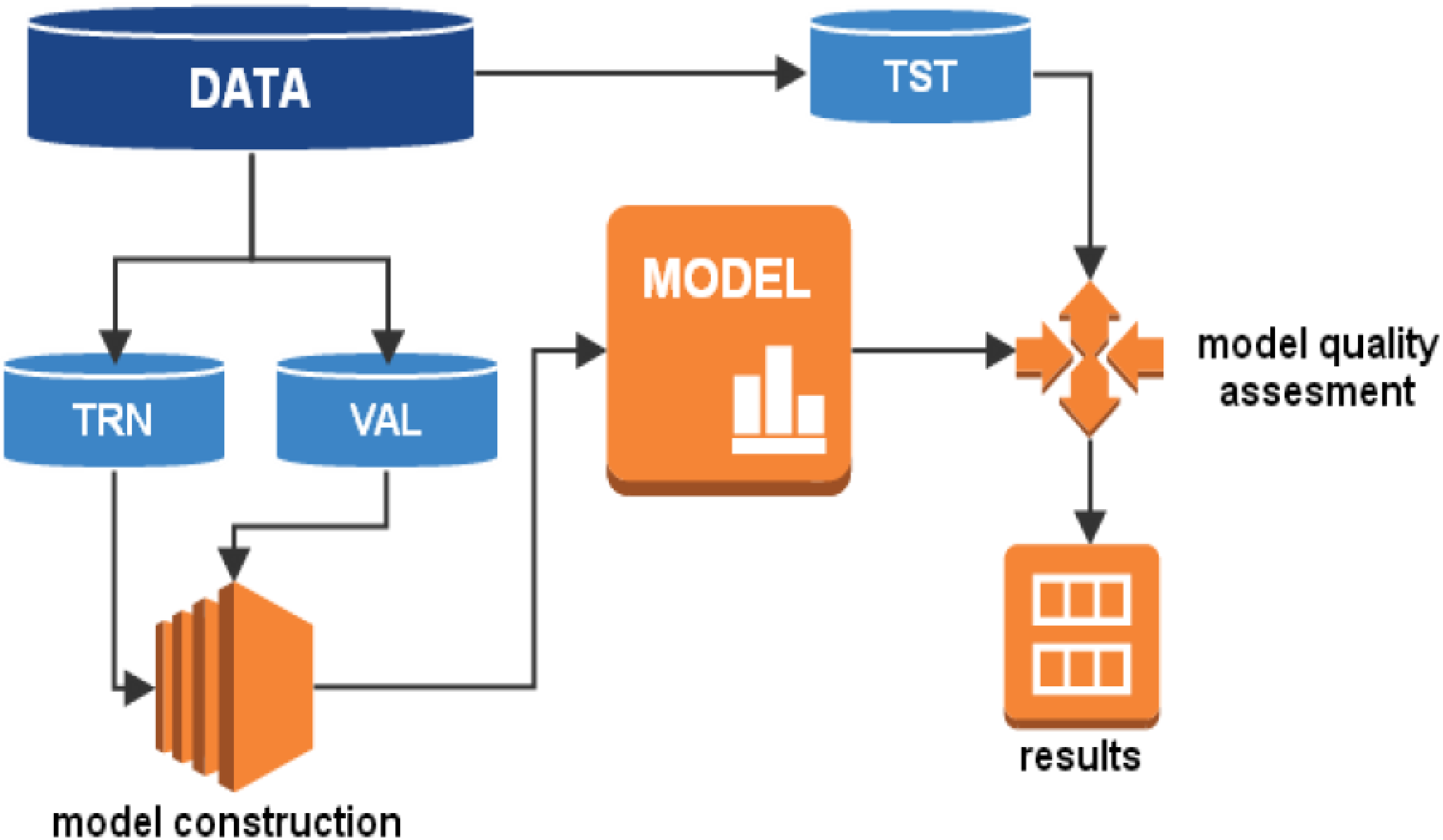
Training and Testing datasets were subjected to the same feature engineering to evaluate the model.



- Out of the main training dataset, a certain percentage is kept untrained to test the model's performance.
- Training set and validation set are split in following percentages: 66.66% : 33.33%.
- On the testing set, the target labels are hidden, until the performance is evaluated.

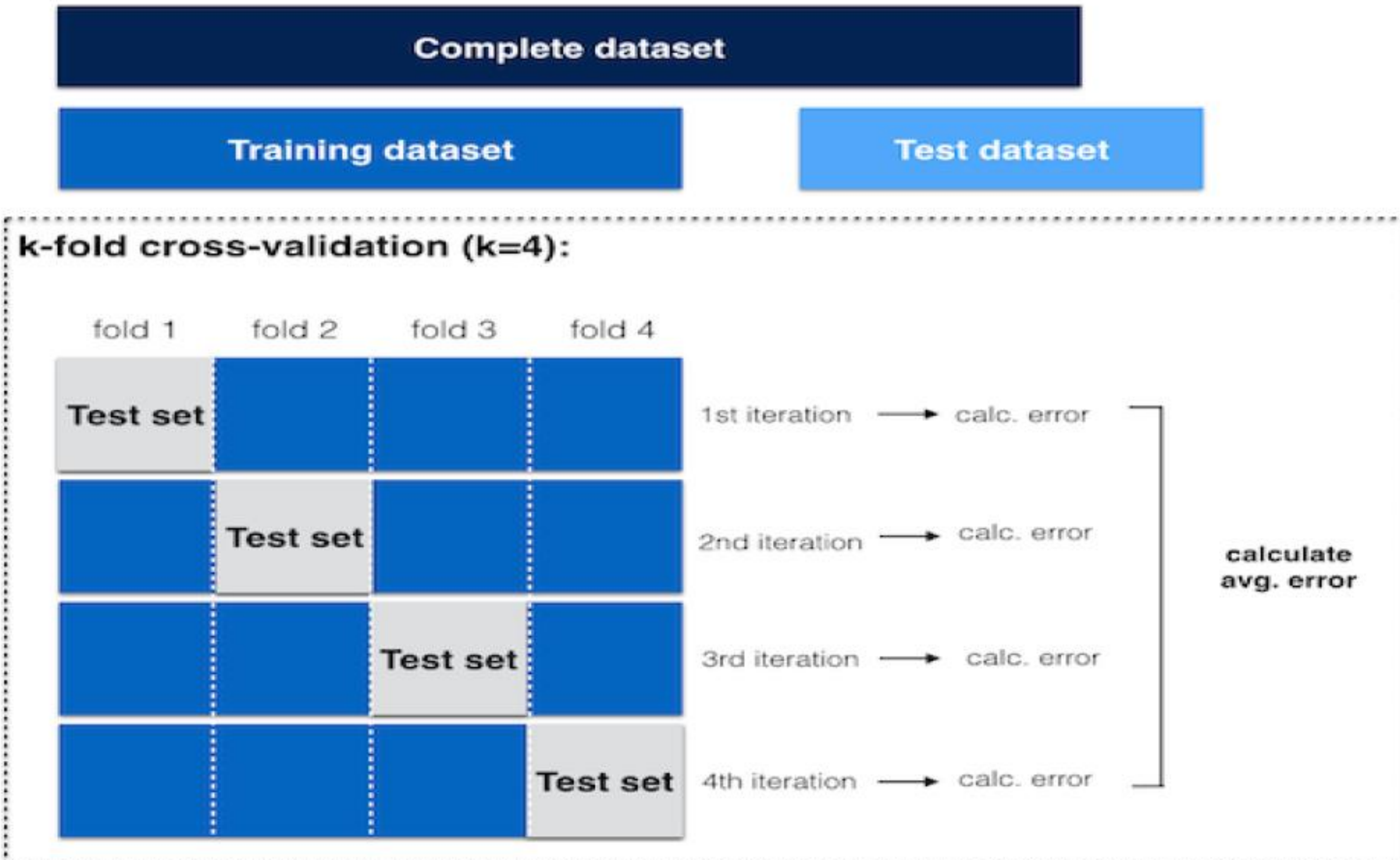
© 2021 National University of Singapore. All Rights Reserved

Data Partition and Preparation



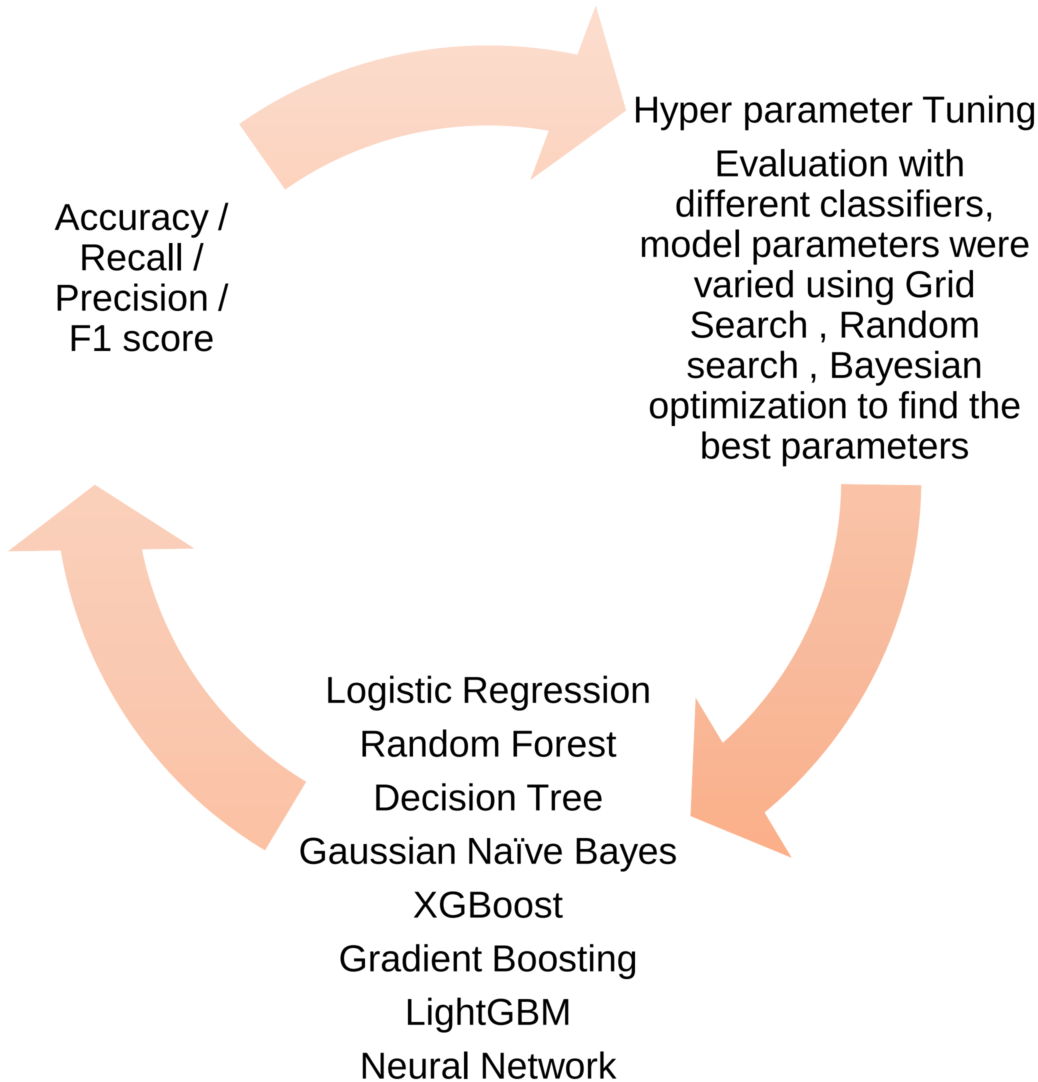
© 2021 National University of Singapore. All Rights Reserved

Cross Validation



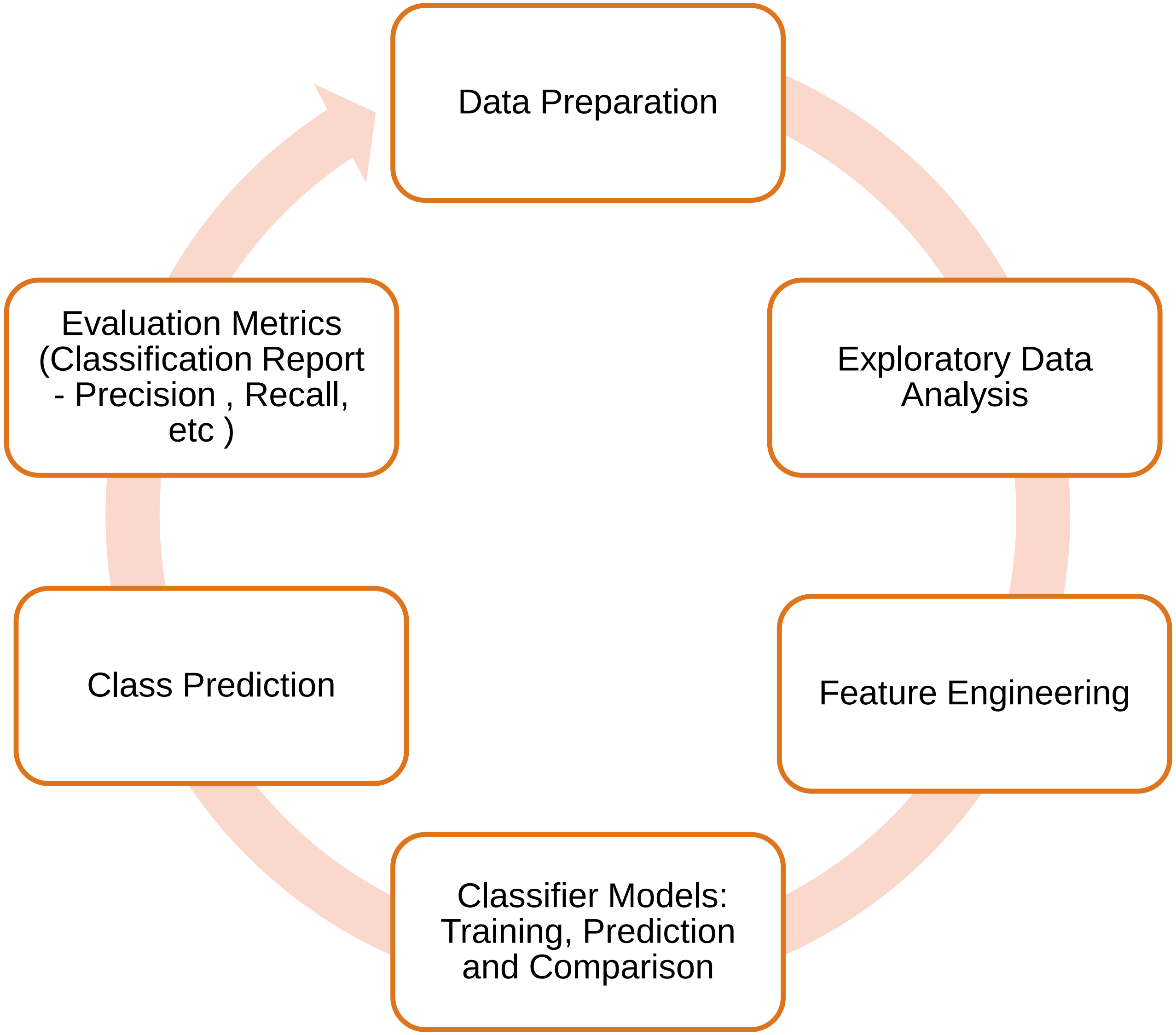
© 2021 National University of Singapore. All Rights Reserved

Technical Approach – ML Techniques



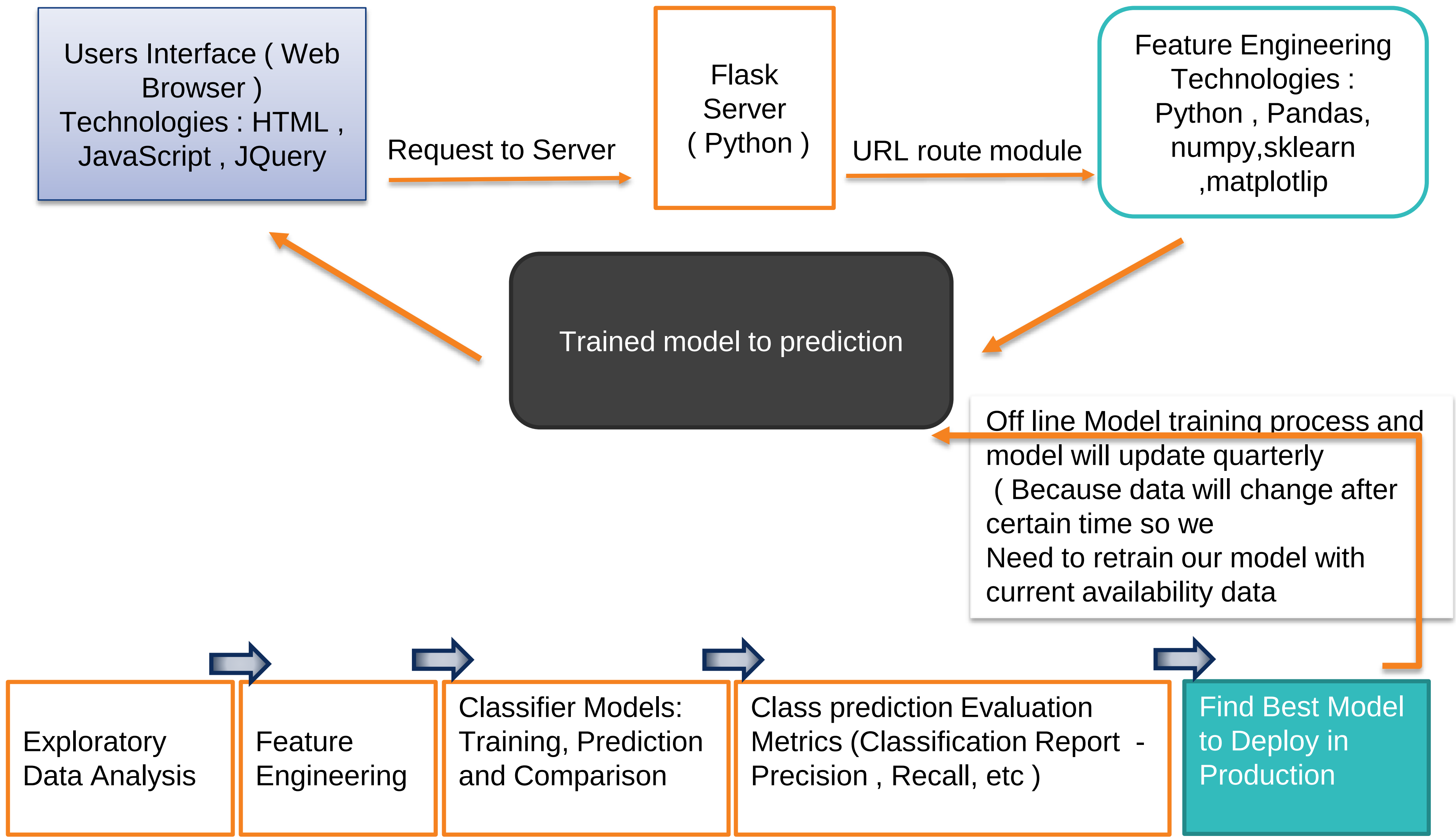
© 2021 National University of Singapore. All Rights Reserved

Project Scope



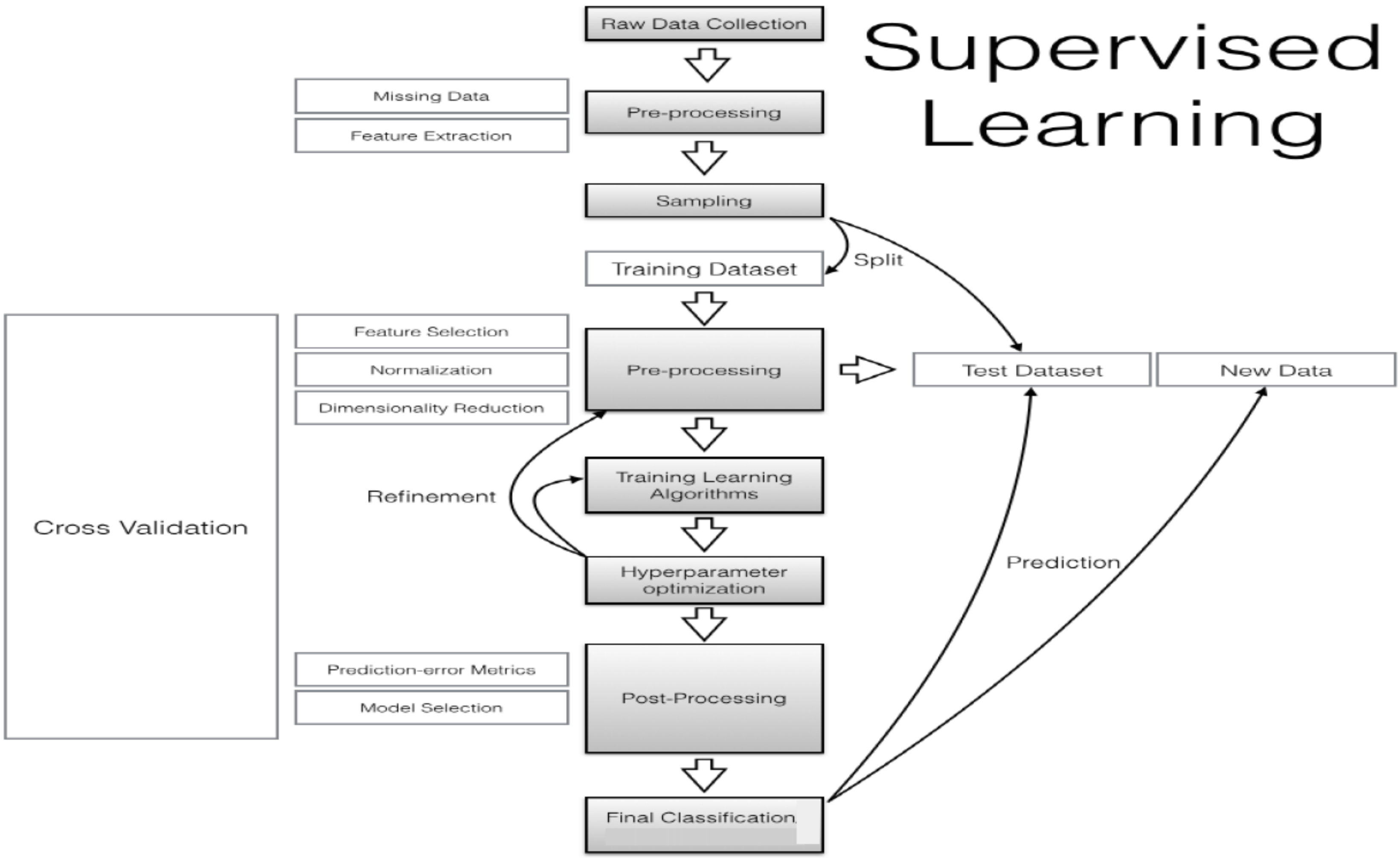
© 2021 National University of Singapore. All Rights Reserved

Technical Approach – System Design



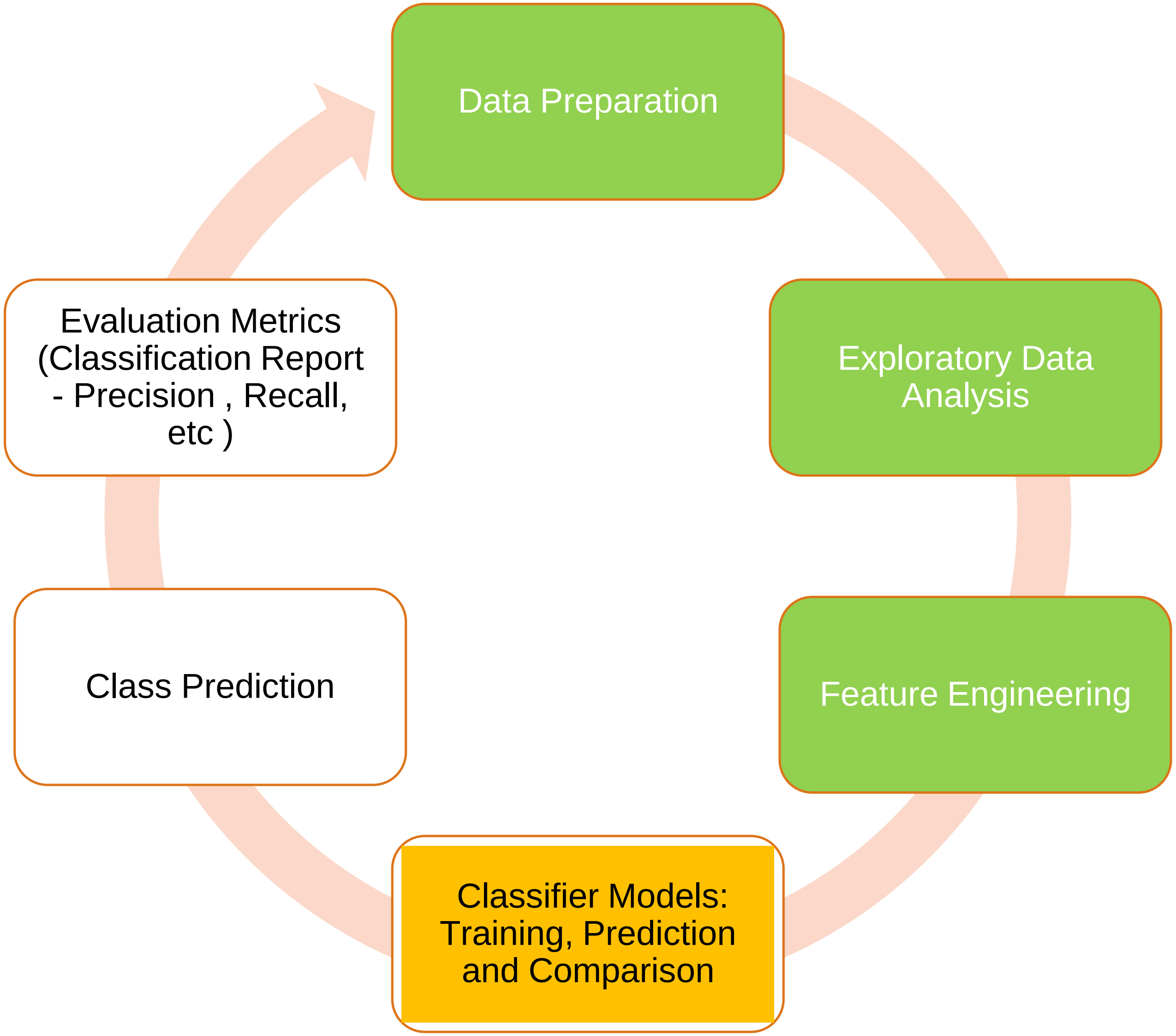
© 2021 National University of Singapore. All Rights Reserved

Supervised Learning Approach



© 2021 National University of Singapore. All Rights Reserved

Status Update



Legend:

Completed
In progress

© 2021 National University of Singapore. All Rights Reserved

Value add to this project

- Performed data wrangling / cleaning, setting up the data for analysis and model building.
- Dealt with data having anomalies.
- Added Interaction variables.
- Performed hyper parameters optimization.
- Incorporated Domain Feature engineering.
- Performed Exploratory Data Analysis.
- Discovered patterns in data.
- Built bagging based ensemble model.

© 2021 National University of Singapore. All Rights Reserved

Q&A

Thank you for your attention.

© 2021 National University of Singapore. All Rights Reserved

References

- Home Credit Default Risk Competition (2018). Kaggle. https://www.kaggle.com/c/home-creditdefault-risk/overview
- Bagherpour, A. (2017). Predicting mortgage loan default with machine learning methods. University of California/Riverside.
- Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machinelearning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.
- Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electronic Commerce Research and Applications, 31, 24-39.
- He, H., & Ma, Y. (Eds.). (2013). Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons
- Lipton, Z. C., Elkan, C., & Naryanaswamy, B. (2014, September). Optimal thresholding of classifiers to maximize F1 measure. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 225-239). Springer, Berlin, Heidelberg.
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... & Ivanov, P. (2016, May). Jupyter Notebooks-a publishing format for reproducible computational workflows. In ELPUB (pp. 87-90).
- Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).
- Poulos, J., & Valle, R. (2018). Missing data imputation for supervised learning. Applied Artificial Intelligence, 32(2), 186-196.
- Kanter, J. M., & Veeramachaneni, K. (2015, October). Deep feature synthesis: Towards automating data science endeavors. In 2015 IEEE international conference on data science and advanced analytics (DSAA) (pp. 1-10). IEEE.
- Bellman, R. (1966). Dynamic programming. Science, 153(3731), 34–37.
- Wright, R. E. (1995). Logistic regression.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.
- Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. R news, 2(3), 18-22.
- Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.
- Zhang H. (2004). The optimality of Naive Bayes. Proc. FLAIRS.
- Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. Update, 1(1),2007.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
- Bergstra, J., Yamins, D., & Cox, D. (2013, February). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In International conference on machine learning (pp. 115-123).
- A Gentle Introduction. (2018). Kaggle. https://www.kaggle.com/willkoehrsen/start-here-a-gentleintroduction
- Introduction to Automated Feature Engineering. (2018). Kaggle. https://www.kaggle.com/willkoehrsen/automated-feature-engineering-basics

© 2021 National University of Singapore. All Rights Reserved