**Case Study – Analytics**

**Objective**

In this case study you are required to explore historical loan records collected by a bank. The primary objective of this case study is to use this dataset to determine whether these loans will be repaid.

To solve this problem, you should prepare and analyse the dataset provided and present your approach and findings.

**Dataset**

The loan dataset **loan.csv** contains the data required for this analysis.

**Deliverables**

* Summary of findings and observations (recommended 1 page)
* Your approach – which includes list of assumptions, limitations or other considerations
* Solution files (all codes/scripts used for cleaning, preparing and modelling the data)

**Next steps**

* Discuss your approach and findings in the next interview. You’ll be required to bring your laptop for demonstration.
* Email materials discussed to [jack.grundling@macquarie.com](mailto:jack.grundling@macquarie.com) for further review post the interview.

Approach:

1. Exploratory data analysis - univariate analysis, density plots, bar charts, correlation analysis.
2. Data clean up: transformations and missing value imputations. Dropped features that had too many missing values, unary level, too many levels i.e. zip\_code. Features with too many levels is likely to contribute to over-training and lower generalisation.
3. Feature engineering: augmented data with additional features based on intuition.
4. Modelling was completed h2o.ai. GLM algorithm was used to model a binary (link = binomial/logit; which is a logistic regression) target to identify factors linked to loan being defaulted.
5. Model training/validation was completed using 5-fold cross-validation using 90% of the data. The remaining 10% was set aside for calculating mean encoding estimates for high cardinality categorical features i.e. addr\_state (50 levels) & sub\_grade (35 levels)
6. Evaluation metric used was AUC of the ROC chart. This is a robust measure which plots true-positive-rate against false-positive-rate for evaluating performance of a logistic model.

Assumptions:

* In the absence of any unique identifier in the data, assume that each observation is independent for purposes of modelling.
* “dti” = debt-to-income-ratio; revol\_util = revolving balance utilisation. Is this balance as a percentage of the credit limit?

Findings & Observations:

* Data set is made up of 10,000 observations of which 12.95% loans were defaulted (i.e. positive target response being modelled).
* The following features are significantly linked to the likelihood to default on a loan:
  1. Purpose
  2. Interest rates
  3. Length of employment
  4. Grade
  5. Term of loan
  6. Annual income
  7. Revolving utilisation rate
  8. Instalment amount
  9. Address state
  10. Public records
  11. Open accounts

Standardized coefficients magnitudes (feature importance):

Blue is POS effect & Orange is NEG effect on P (default=Y | data).



Preliminary modelling appears robust and consistent based on the AUC results over 5-fold cv. Training = 0.695459 and validation = 0.678502.

Next steps (“Phase 2”):

* Observed a few cases where the stated “purpose” does not match the detail mentioned in the “desc”. Explore “desc” for more detail using text mining analysis.
* Better deal with multi-collinearity within the features based on further statistical testing.
* Explore further data transformations & interactions amongst features to “smooth” out other non-linearities in the data.
* Interesting observations relating to following features:
  + dti behaves differently to revol\_util and interest rate. Why?
  + Length of employment, verification\_status\_income
* Check for additional nested/linked data on borrowers, their product holdings and usage at the bank.
* Check for external data i.e. credit bureau related, demographic modelling data.
* Apply other types of meta-learners (GBM, DRF, XGBoost, Deep learning) with hyper-parameter tuning to achieve performance uplift e.g. **recent Kaggle competition winning team achieved an AUC ~0.80 on a similar data set.**