Outlining

Friday, November 30, 2018 3:20 PM

Model approach:

1. Exploratory Analysis - Monthly, quarterly, Annual & overall summary analysis;
2. Correlation Analysis - Monthly, quarterly, Annual & overall Pearson's; chi-squared; ANOVA
   It depends on what sense of a correlation you want. When you run the prototypical Pearson's product moment correlation, you get a measure of the strength of association and you get a test of the significance of that association. More typically however, the significance test and the measure of effect size differ.
   Significance tests:
   - Continuous vs. Nominal: run an ANOVA. In R, you can use ?aov.
   - Nominal vs. Nominal: run a chi-squared test. In R, you use ?chisq.test.

   Effect size (strength of association):
   - Continuous vs. Nominal: calculate the intraclass correlation. In R, you can use ?ICC in the psych package; there is also an ICC package.
   - Nominal vs. Nominal: calculate Cramer's V. In R, you can use ?assocstats in the vcd package.
3. Feature Engineering - Monthly, quarterly, Annual & overall variable binning; woe binning
4. Logistic model - building model; K-S, ROC, Gini based evaluation on train and test
5. Scorecard model - building model with standard evaluation coupled with evaluation on Red, Yellow, Green status
6. Design of methodology for strategic targeting

Use case approaches

Thursday, December 6, 2018 10:56 AM

Pre-meeting brainstorming

1. Bucketing/segmentation:
    a. Driven by business understanding:
        i. Potential ones can be - under reported, good or ideal, Defaulters - habitual and real, abonrmalities, etc.
2. Registration - SSM company registration data - external data
    a. Can be a dashboard with drilldown starting from location estimate to industry by location
    b. Further industry can be pinpointed to the exact firm name through databases such as yellow pages, registration data
3. Under-reporting
    a. Can be a predictive model given the features from industry other players
    b. Shall be a model which is able to give the size of the company given industry, location, revenue and other relevant factors
    c. Shall be another model which is able to give the contribution amount of the company given industry, location, revenue, number of employees and other relevant factors
4. Good employers turning defaulters
    a. Can be one model where we predict the probability of default given other features
    b. To start with, this can focus on MIA 0-2 as the defaulters
5. Categorizing defaulters in to real and habitual
    a. Model for identifying the two

Finalized use cases:

1. Propensity to pay scoring model
2. Probability of default model - already built up for MIA 1 & 2
3. Employers lack information and hence should be approached by the collector/advisor accordingly - a dashboard for this

Few ideas:

1. Include whether or not a quarter month as an attribute as there seems to be a seasonality in the data

Detailed Steps

Tuesday, December 11, 2018 4:12 PM

1. Ideal ratio of good to bad in the model input for logistic regression
2. Prepare scripts to create visualizations

Features to engineer:

1. Binning
    a. To check how to identify optimal bins in data
2. Month as quarter end or not
3. Sampling techniques to narrow down to the time and observations in each time period
4. Payment date near to cycle end date

Modelling:

1. Incremental validation and ensembling of the logistic regression model using historical data

Model Explained & Requirements

Thursday, December 13, 2018 11:41 AM

1. For variables, need to do aggregations over performance window in the second set of iterations if not at first
2. For multicollinearity, the variable with higher IV stays
3. Data input to variable filter shall be in a similar format as the one for GermanCredit
4. The response variable should be good and bad for the filter input
5. Output from filter shall be 1 for bad and 0 for good
6. How to interpret high score ?
7. How to solve the problem due to exogenous variable like Quarter end Month ?

Expenses: 20551608
Time: 20552438

Business Notes

Thursday, December 6, 2018 2:16 PM

1. Employer journey broken in to pre, during and post - meaning?

Contacts

Thursday, December 13, 2018 9:52 AM

Hassimi - Infrastructure
Chin - Data mapping and direction on the sources
Teh - overall mobilisation
Robi - Project Team
Matahir - Project lead overall
Nordin/Hanafi - Data extraction

Roadmap ahead thoughts

Thursday, December 13, 2018 11:25 AM

1. Archetyping of Employers based on collection scores
2. Usage of Machine learning to further finetune the model
3. Productionizing or the implementation plan for the model

Meeting Notes

Thursday, December 13, 2018 1:55 PM

8 - Jan Meeting with JPK HoD:

- Difference from Tandem PoC
- External data brought in to talk
- Tendency from enforcement to advisory
- Recognition programme at branches right now
- Employer advisory and employer services instead of first two heads in slide 4
- Ambition is to create a pooled list and where you can create a system
- Assignment list to be created with some actions as recommendation
- Recognition can be helped through this
- Sharifa is most experienced in the quality but going to retire
- Faizal is co-ordinator for the enforcement queries
- Involve the team from enforcement more rigorously

Activities for analysis

Tuesday, January 8, 2019 10:00 AM

1. t-test results compilation - code
2. Equivalent of t test for categorical variables - research & code
3. Correlation coefficients, identical ratio & statistical differences to calculate
4. Final evaluation filtered list to generate
5. What should be optimal ratio of good and bad for the model development - research
6. How to preserve the original data statistics in the sampled dataset - research

Send over the results to Chin and team with the summary table
Caveat is to put t test as one column and another is the identical value ratio

Key points - data preparation

Tuesday, January 8, 2019 2:16 PM

1. For the overall data with just active employer filter:
    a. **Quarter end indicator to derive**
    b. **Check on the way to calculate the full payments in the evaluation**
    c. **Check on the way to get the payment channel which creates duplicates**
    d. **Cheque status to be included or can take the information from DDHIST new fields or DDICTM table**

Key Points - Business Discussion

Monday, January 14, 2019 10:59 AM

1. The scope is to prepare the preventive actions
2. Flexibility of the model/system to show at the end of PoC
3. Show some segments to be taken forward for the discussion
4. Can show trajectory of few employers who have been falling bad and actually defaulted later

Research

Tuesday, January 8, 2019 2:27 PM

1. How to output file faster using data.table?
2. Using bigtabulate, etc for faster aggregations, merging, etc
3. Using parallel, foreach for month level processing

Running Questions

Monday, January 14, 2019 4:15 PM

- Confirm on the payment dates

-- Correlation analysis through the use of correlation measures
- Prepare Var. Selection Report
- Check for the duplicates that have been created in the process
- List dropped variables with reasons of dropping:
  Payment channels - created too many duplicates
  Industry - Didn't find a variable to map
  Employer status - used to filter and hence same for all in dataset
  CTML Payments - Can't join this data

- For EDA:
  - Plot each variables' histogram or CDF
  - Clearly articulate the messages on each slide
  - Do the histogram or stacked bars for frequency for the categorical variables and show it for the ones with limited numerical values as well
  - Check if any relation can be more clear using cdf, etc.

Wrapping up

Thursday, January 17, 2019 6:04 PM

1. Choo's Idea:
    a. Finding & Insights
    b. Strategy & Actions
    c. Model Performance
2. Take clues from UT project and put together visuals, etc.
3. Take clues from Choo's shared folder and try putting together something from there

KIV Points

Friday, January 18, 2019 2:31 PM

1. We ignored all date outliers for liability age
2. We ignored all null value rows for ABC Code
3. We ignored all null values for the field REGTYPE

Reference points

Wednesday, January 23, 2019 3:19 PM

1. Summary of delays:
    a. Delay due to infrastructure setup: 3 days (3-7 Jan)
    b. Delay due to infrastructure outage: 2 days (22 Jan- 2nd half + 24-25 Jan)
    c. Delay due to medical absence (Rag): 1 day (02 Jan)
2. Summary of fields that are removed
    a. Employee count: Not complete dataset in DDMAST reported under the field NOEMPE
3. Summary of data filters applied at each stage

Scripts_model_building

Friday, January 25, 2019 2:48 PM

1. Creating the dependent variable:

```
data = data %>%
 mutate( creditability = ifelse( creditability == 'bad', 1, 0 )
       , creditability = as.factor(creditability) )
summary(data)
```

From <https://rstudio-pubs-static.s3.amazonaws.com/376828_032c59adbc984b0ab892ce0026370352.html>

1. Remove the variables due to singularity
2. Running the model with Lasso and cross validation: Identify a way to build model using a smaller sample size and then augment it by using the bigger dataset
3. Shortlisted variables after Lasso regression in a formula format
4. Shorlist variables for correlation as well

Model Building steps

Wednesday, January 30, 2019 2:08 PM

1. Removed singular bins/aliased coefficients - Impact due to the perfect multi collinearity
2. Variables with VIF more than 4 to be ignored - none found post 1

Lasso Regression outputs:

1. AUC: 82.3
2. KS: 54.1
3. Accuracy at
    a. 0.1: Train - 78.4 % ; Test - 78.3 %
    b. 0.2
    c. 0.3
    d. 0.4
    e. 0.5

Logistic Regression outputs:

1. AUC: 82.3
2. KS: 54.1
3. Accuracy at
    a. 0.1
    b. 0.2
    c. 0.3
    d. 0.4
    e. 0.5

08-Feb actions

Thursday, February 7, 2019 4:56 PM

1. Build the outcome distribution by performing PSI
2. Check on frequency of values update among the variables, like new categories adding up

Find out the expected results due to actions, etc.

See if PSI can be done now or plan for it in model refinements
Check on reasons for NA; start from analyzing Bins information; most probably due to payment ageing in last month - shall need to change its type in the binning code

Analyze the reasons for improved performance, see if it's because of the Y variable being passed in as well

Data seems correct till October only

Revise train and test score bands

Out of total 560K active employer accounts

Employer accounts lost:
Around 12K had their liability commencing from a month post 2018 OCT
Around 7K mismatch between CFMAST and DDMAST in active employer accounts

485K form submission accounts in 201808-10
Probable reasons can be:-

1. The Employer data we considered is 2.5 years and some employers might have paid contributions beforehand

Need to put an example to the payment ageing explanations