Executive Summary of Key Findings

1.Problem statement

Begun in 1997, Genpact is a global professional services firm delivering digital transformation by putting digital and data to work to create competitive advantage. On November 5, 2020, CMMI Institute has appraised Genpact on the Capability Maturity Model Integration (CMMI) V2.0 for Maturity Level 5. The highly complex appraisal is a significant milestone in modernizing the quality and delivery of Genpact's technology services. Also, according to the financial report of quarter 3, 2020, total revenue for this year is now expected to be $3.68 to $3.695 billion, up from the prior range of $3.63 to $3.67 billion.

Like other companies under B2B mode, strategies of Genpact relies increasingly on customers' ageing bucket and binary measures of credit worthiness. Also, the growing need for companies' financial health requires better segmenting customers and leveraging predictive models. Overall, the primary goal of this case analysis is to create a predictive model for the pay of customers studying patterns from the historical payments, which may help Genpact to generalize a healthier development pattern in the future.

2. Methodology

The dataset provided by Genpact consists of 45,841 records with 28 columns. Our team deleted several columns with consistent value or with more than half missing values. Also, we extracted the weekdays of document created date and contract due date since they had high correlation with overdue days. Based on the frequency, the city and region were regrouped. Payment methods and payment terms were regrouped in terms of basic business assumptions. To better reflect the customer behavior pattern, we created mean delays, mean contract account and frequency based on each customer. After data cleaning and feature engineering, we choose to derive from two paths: one was to create a multinomial logistic regression model and the other was to utilize machine learning techniques, such as random forest and XGBoost.

3. Key findings

3.1. Exploratory Data Analysis

- "Frequency" and "Total amount" have slightly stronger correlation.
  Before calculating the correlation among variables, we created "mean_Amount" and "Means_delays" , aiming to see if average amount and average delay days of customers would have more obvious correlations with other variables. Correction map **(Exhibit 1)** shows that variable "Frequency" and "Total amount" have slightly stronger correlation compared with other variables.
- Top 4 customers contributed 77.94% of the total amount.
  According to our first finding, we dig into the aspect of customer level. **Exhibit 2** shows a dramatic result that top 5 customers' ID contributed to more than 80% percent of the total amount.
- 84% of the payments were from Direct Debits.

After ranking the popularity of each payment method, **Exhibit 3** shows that the most popular payment method is direct debit, followed by no payment method and regulatory. Wire and Third party payment are not so popular among customers.

- Customers were less likely to pay the bill in time in Quarter 2 of a year.
  Box plot in **Exhibit 4** shows that the 2nd quarter of the year generalized the highest amount of overdue delay days, which means that customers were more likely to pay their bills after due date in Quarter 2 of a year

3.2. Model

1) Variable selection :

- “Region”, “City” and “Zipcode”

These three variables are all encrypted categorical variables. Among the three variables, “City” and “Zipcode” are highly correlated, so we decided to keep “City”. We kept cities and regions which appeared most frequently. By doing one-way ANOVA tests, we compared the means of days overdue Between different regions and cities, and found that there is a significant difference between groups.

- “Payment term bin”

Through EDA, we found that most payment periods fall in 4 different groups: A week (7 days); Two months (60 days), Half a year (180 days) and More than half a year (>180 days), thus we group “Payment Term” into 4 groups

2) Hierarchical Clustering to filter out outlier

We use hierarchical clustering to see whether customers can be classified into different groups, and then adding this variable to our model. Results showed that the difference between each cluster is really small. We removed the outlier customer “5039221137” who has only 2 orders while the average number of overdue days is more than 700.

3) Result of model

- Parameter

Parameter Among all the models, Xgboost generalized the best result of 92.81% accuracy and 0.86 Kappa score. We set 'learning_rate' to 0.1, 'max_depth' into 5, 'num_boost_round' into 200, 'objective' into 'multi:softmax','random_state' into 27, 'silent' into 0 and 'num_class' into 4.

- Feature Importance and Validation

The XGBoost model generated feature importance for all 35 variables, the important features were mostly consistent with descriptive statistics findings (Exhibit 6), the three most important features are “Amount”, “weekday.NetdueDate”, and “Payment term”.

We used Accuracy and kappa scores to evaluate the performance of models. For the validation result, XGBoost model has 92.81%, Kappa’s scores as 0.86.

4. Conclusions and Recommendations

Based on the feature importance and descriptive statistics, we classify all the features into three types: cyclic features, customer features, and contract features.

We found that there is a strong seasonality effect on customer pay patterns, past data have shown that customers were more likely to overdue in Quarter 2. At the same time, seasonality related variables such as "Net Due Date", "Quarter"s, "Weekday doc. Date" are important features for prediction. We recommend choosing a nice day to make the contract and set the due date properly, and pay especially attention to Quarter 2 where the delay is more likely to happen. We also suggest signing the contracts and setting the due date on weekdays. People can hardly remember to pay for the service on weekends.

Customer's behavior also plays an important role in predicting propensity to pay, especially on his or her overdue delay days, contract amount, and demographic distribution. From what we discovered, region AA127 and City AA38 are two locations that are important to delays days predictions, they also have a high frequency. The mean delay date for each customer also has a large impact on the customer's propensity to pay, if a customer tends to have a large mean delay date, the company should pay attention to that customer. We think that it is fair to allocate more resources to these areas and set aside a small team focusing on collecting more information about customers to study their paying pattern, to not only serve customers better but also have more attention to customers' propensity to pay.

At the same time, contract features such as payment terms and payment methods are critical to whether the customer will pay on time. Moreover, nearly all the money was paid by direct debit. We believe that encouraging customers to sign the contract within one week and use deposit debit as a payment method will likely to decrease the chance of late payments. At the same time, we should make sure that the payment channel is smooth and the customers are well informed to pay on time.

5. Limitations

5.1. Our XGBoost model shows powerful strength in predicting propensity to pay. However, further statistical modelling should be done to provide inference of the relationship between variables. More research could focus on the different effects of weekdays on customer's tendency to pay, the payment methods' effect on decision making, the regional research on specific locations, the optimal payment term to influence paying and the customer profile that shows personal patterns.

5.2. The XGBoost model can still be improved by tuning and providing more data. There are only 69 unique customers in the provided dataset and only several columns reflected customer behavior. As we collect more data, we can create customer profiles that illustrate the customer pattern of paying, thus generating more accurate predictions and providing customized service.

Appendix

Exhibit 1. Correlations among variables.

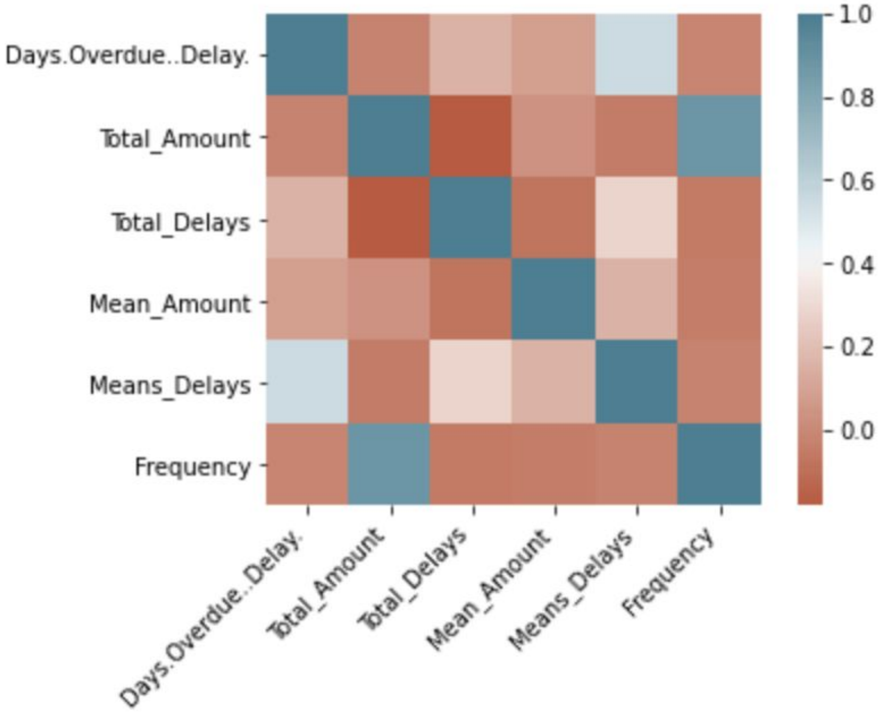


Exhibit 2 Customers contribution of total amount (by ID)

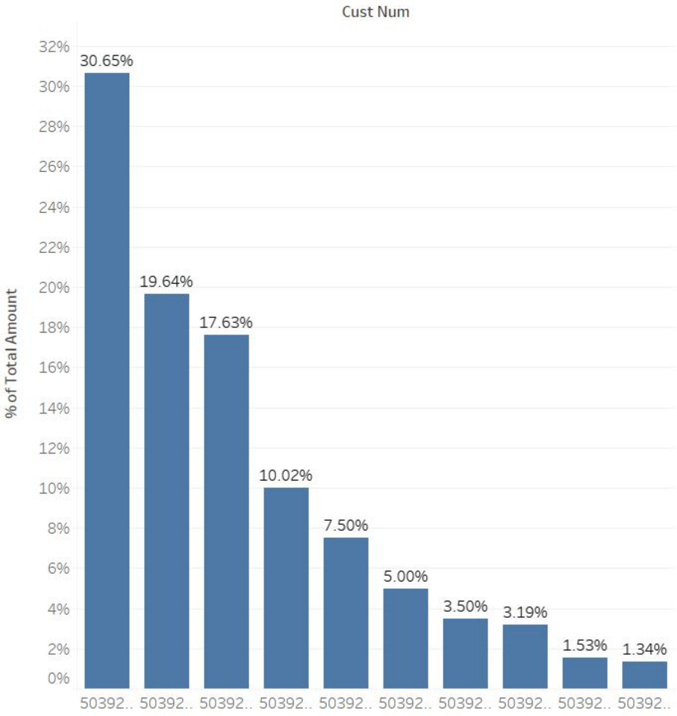


Exhibit 3 Payment method distribution

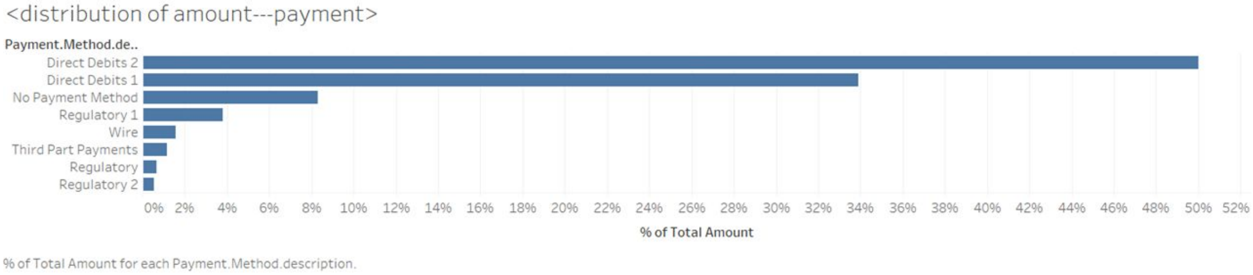
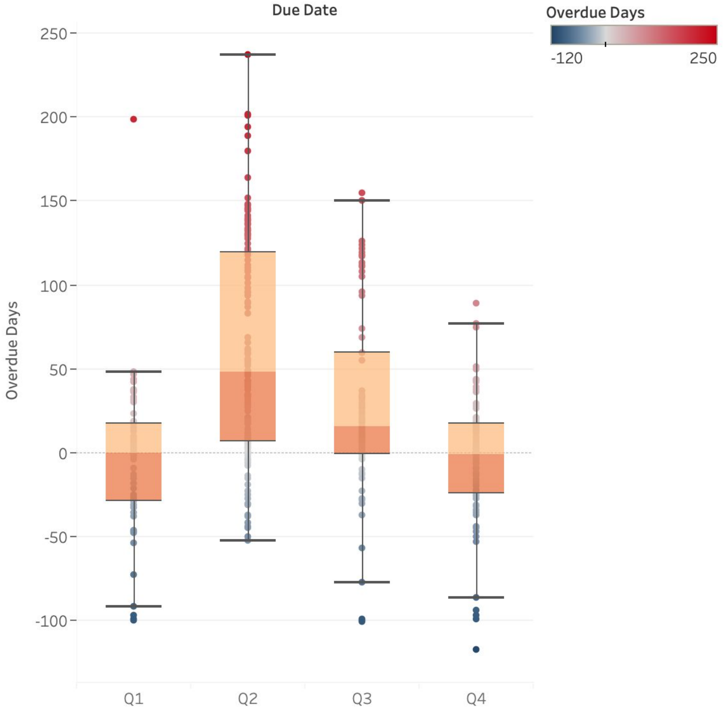


Exhibit 4 Overdue delay days by quarter

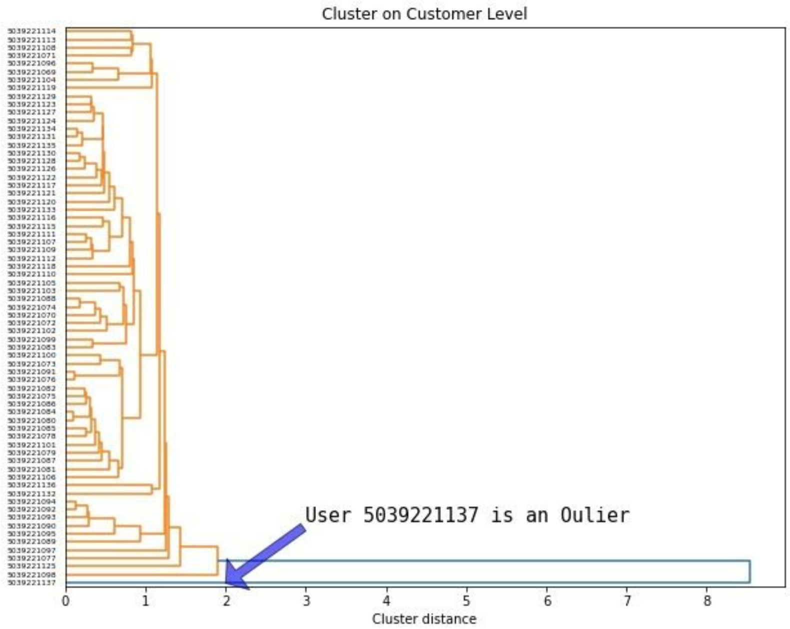


Overdue delay days by Quarter in 2012-2016

Customers were less likely to pay the bill in time in Q2.

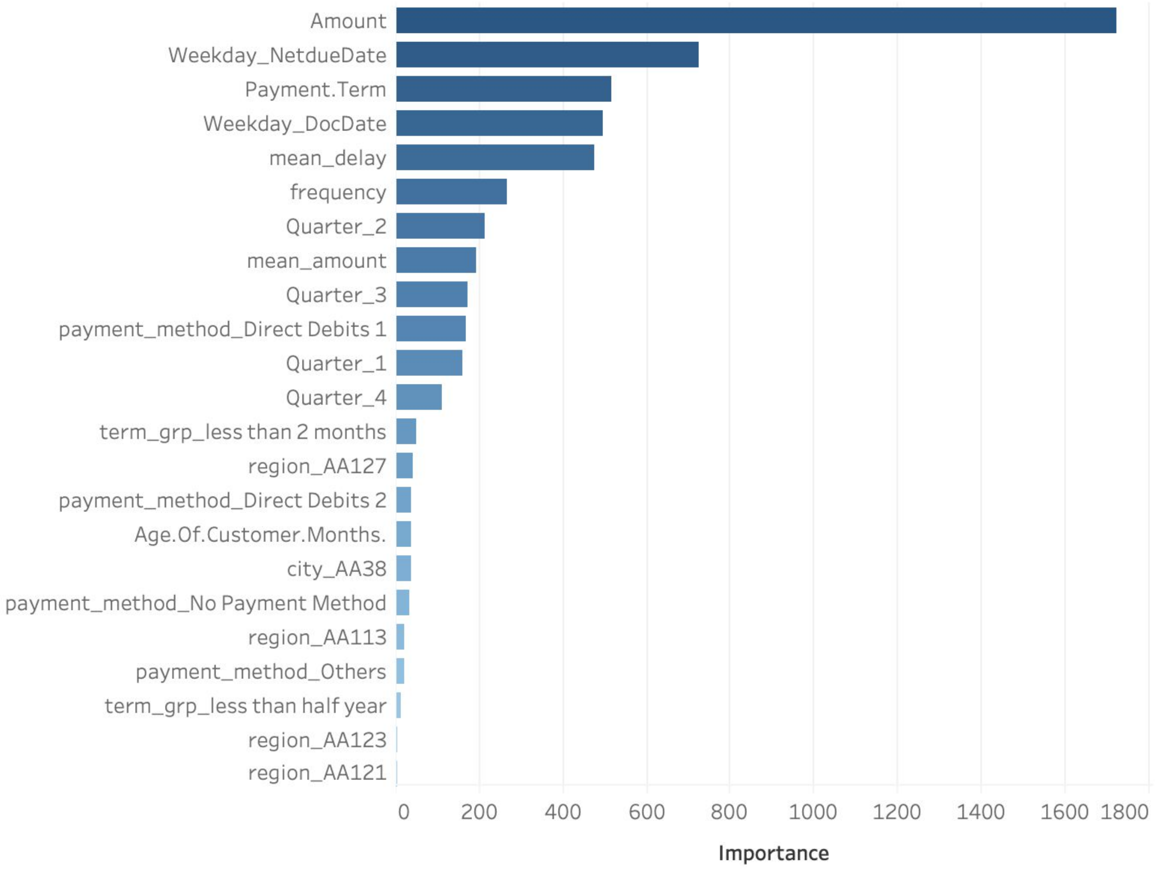
Overdue Days for each Due Date Quarter. Color shows details about Overdue Days. Details are shown for Overdue Days. The data is filtered on Due Date Year, which keeps 2012, 2013, 2014, 2015 and 2016. The view is filtered on Due Date Quarter and Overdue Days. The Due Date Quarter filter keeps Q1, Q2, Q3 and Q4. The Overdue Days filter ranges from -117 to 242.

Exhibit 5 Hierarchical Clustering to filter out outlier



User 5039221137 is an Oulier

Exhibit 6 Feature Importance


Amount
Weekday_NetdueDate
Payment.Term
Weekday_DocDate
mean_delay
frequency
Quarter_2
mean_amount
Quarter_3
payment_method_Direct Debits 1
Quarter_1
Quarter_4
term_grp_less than 2 months
region_AA127
payment_method_Direct Debits 2
Age.Of.Customer.Months.
city_AA38
payment_method_No Payment Method
region_AA113
payment_method_Others
term_grp_less than half year
region_AA123
region_AA121
Importance