Funding Club Market Entry Analysis

MSBA 6320 Team Project

December 16th, 2015

Team 8:

Ameeta Japtiwale
Manikandan Kuppuswamy
Fu Lin
Kaushik Nuvvula
John Tong

Abstract

It's December 1st, 2011, five University of Minnesota students founded an online marketplace that connects borrowers to investors, called Funding Club. The founders are graduate students who wanted to provide funding to borrowers at a more competitive rate than its biggest competitor: Lending Club. Lending Club is currently the largest online marketplace for lending. In order to compete and offer an appealing proposition to borrowers, we extracted Lending Club's customer data from 2007 to 2011 and stored it in our own database for analysis.

With the data available to us, we created a predictive model to see the type of customers who are typically approved or declined by Lending Club and then targeted these groups differentially. For the unqualified candidates, we would consider the top 5% at an annual interest rate excess of 30%. For the qualified ones, we will attract with a lower interest rate. This strategy impacts our profitability from the start, but we are willing to sacrifice profit to build the Funding Club brand and initial customer base.

Introduction

Funding Club is an online marketplace that connects borrowers to investors. The company was founded in 2011 by a group of University of Minnesota students. We founded this company to provide funding to borrowers at a more competitive rate than our competitors. We in turn also offer an opportunity for lenders to receive a meaningful return on their cash. To launch a successful startup, we extracted and analyzed data from our biggest competitor: Lending Club.

Lending Club is currently the largest online marketplace for lending. In order to compete and offer an appealing proposition to borrowers, Funding Club is willing to take on a slightly more risky portfolio. We will consider clients who were declined by Lending Club as well as compete for qualified clients by offering a lower interest rate. The team extracted Lending Club data from 2007 to 2011 and loaded the data into our data warehouse. We then performed predictive and descriptive analysis on Lending Club clients (both approved and declined) to help finalize our competitive go-to market strategy.

Requirements

Data Description

We utilized datasets from 2 different sources for our analysis.

Lending Club: https://www.lendingclub.com/info/download-data.action

- Approved Loan Data : years 2007 – 2011
- Declined Loan Data : years 2007 – 2012

FRED Economic Data:

- FRED CPI (monthly)
  (The Consumer Price Index for All Urban Consumers)
  https://research.stlouisfed.org/fred2/series/CPIAUCSL
- FRED GDP (quarterly)
  (Gross domestic product (GDP), the featured measure of U.S. output)
  https://research.stlouisfed.org/fred2/series/GDP
- FRED LoanRate (monthly)
  (Rate posted by a majority of top 25 (by assets in domestic offices) insured U.S.-chartered commercial banks)
  https://research.stlouisfed.org/fred2/series/DPRIME

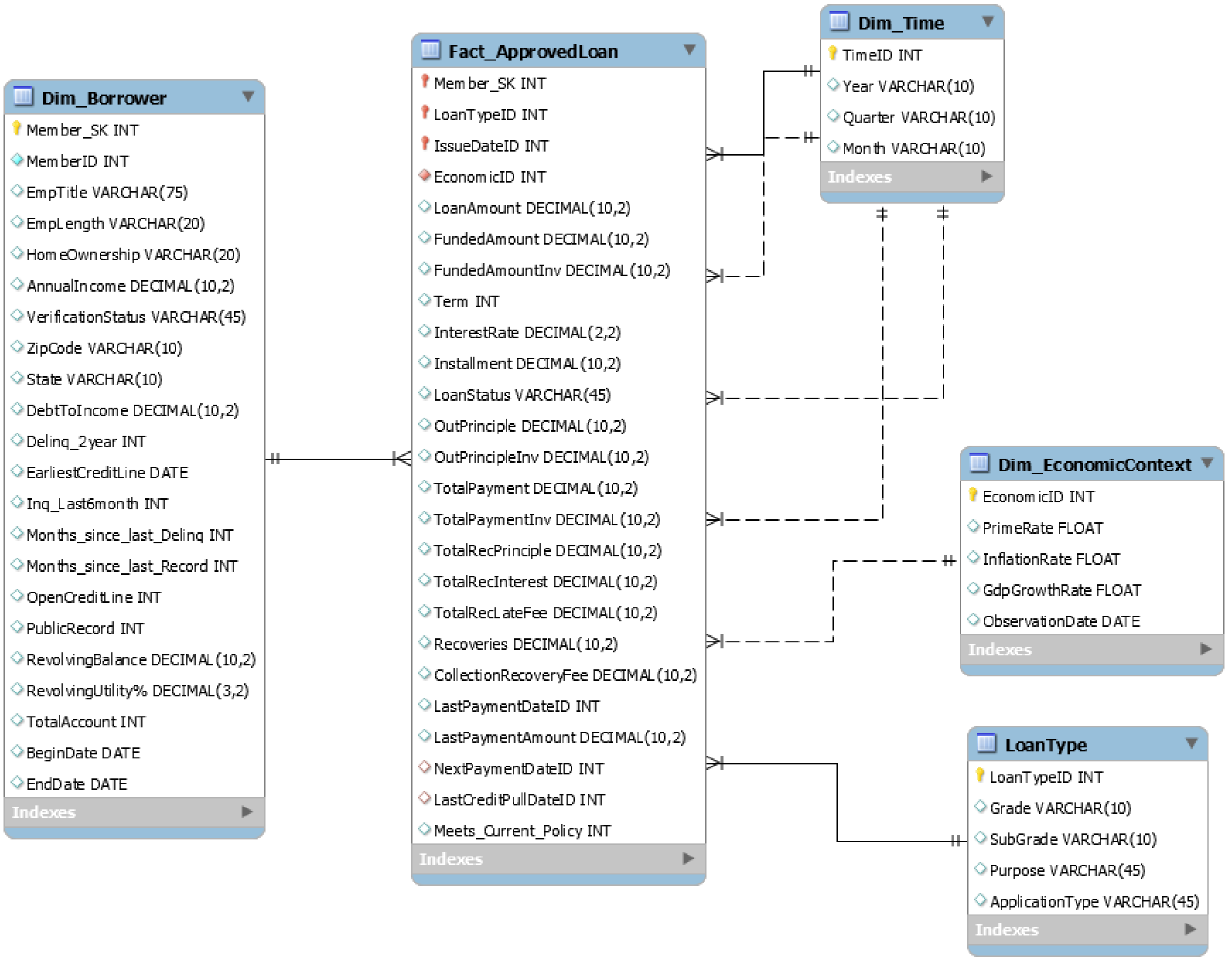
Business Questions

1. Determine what type of customers are qualified for a loan at Lending Club
2. What are the different interest rates offered to borrowers of different grades?
3. How much money (NPV & ROI) does Lending Club make off their customers of each grade?
4. Percentage of loans charged off for customers of each grade

Design

Dimensional Model

With the business questions in mind, the team designed a dimensional model that will help with the implementation of the data warehouse. We will then use the data warehouse to perform analysis on the approved loans data extracted from our competitor. The figure below shows the dimensional model for Lending Club data.



Grain: One row per approved loan application for the borrower (Member_SK, LoanTypeID, IssueDateID)

Fact: Loan Application

(Loan Amount, Funded Amount, Funded Amount by investor, Loan Term, Interest Rate, Installment, Loan Status, Outstanding principle, Outstanding principle by investor, Total payment, Total payment for investor, Total Received principle, Total Received Interest, Total Received Late Fee, Recoveries, Collection Recovery Fee, Last Payment Date, Last Payment Amount, Next Payment Date, Last Credit Pull Date, Meets Current Policy)

It should be noted that the fact table is of type '***accumulating snapshot'*** i.e. multiple snapshots are stacked in a single row using various date attributes. This type of fact table is generally wide and suitable for processes having few major milestones.

Dimensions:

- **Borrower**: This dimension consists of the borrower details such as the income, employment length, delinquency, public record, etc.
- **Loan Type**: This dimension describes the loan depending upon attributes such as grade and sub-grade (assigned by Lending Club), purpose of the loan and application type (Individual/Joint).
- **Date**: This is the time dimension associated with the loan which includes different dates such as Issue Date, Last Payment Date and Next Payment Date.
- **Economic Context**: This dimension provides information about the market interest rates offered in the same time frame and is thus useful for comparison with Lending Club interest rates.

ETL Strategy

1. **Data Extraction Stage**
   During the data extraction stage, data from various sources (as mentioned above) was extracted in the form of excel files.
2. **Data Transformation Stage**
   Multiple transformations were performed to prepare the data before loading it into the data warehouse.
   - The date attributes namely issue date (issue_d), last payment date (last_pymt_d), next payment date (next_pymnt_d), and last credit pull date (last_credit_pull_d) had date values in multiple formats and were transformed into a consistent date format.
   - The missing values in the dataset were analyzed and imputed using suitable values. For example, attributes such as delinquency in the last 2 years (delinq_2yrs) or number of public record (pub_rec) was set to zero when N/As were encountered.
   - Some attributes were recoded as per the requirement of the warehouse design. For instance, the attribute employment length (emp_length) which denotes the length of employment of the borrower in years had values including years 1 to 9, <1 and >10. Employment length less than 1 year were converted to 0.5 years and employment length greater than 10 years were modified to 10 years to obtain all the entries in numerical values.
3. **Data Loading Stage**
   The strategy to load the data was by loading one year of data considering one dimension at a time. The dimensions will be loaded in the below order and then the fact table (Loan Application):
   - Economic Context
   - Loan Type
   - Time
   - Borrower

   Special attention will be given to changing or historical attributes in the Borrower dimension.

Slowly Changing Dimensions (SCD):

- SCD -Type 1: Borrower attributes such as State and Zip Code in his address may change. It is important to maintain the current address of the borrower. Hence, these attributes are type 1 SCD wherein the old values will be overwritten by the new ones.
- SCD Type 2: Borrower attributes such as Annual Income, Debt to Income or Employment Title are declared as type 2 SCD because the historical information is of significance for such attributes.

Data Capacity Planning

Our data warehouse aims to keep track of milestones and payments of each loan. Currently, our data is just a snapshot from 2015-11-01. In order to realize our tracking purpose, we need to store the history of each borrower and each loan. Since data concerning the borrower dimension and fact table will take up most part of the capacity, we make our estimates mainly based on these two parts.

For borrower dimension (most columns are type 2 SCD), we estimate that on average the record of each borrower needs to be updated every month. For effective borrower history tracking, we plan to keep the history of each borrower for the next 5 years. Next, we assume that Lending Club's customer base (now 952496 people) will grow 20% per year. In order to fulfill the demand for the next 5 years, we estimate that our borrowers' capacity in competitive data warehouse capacity will be 6 GB.

$$1.2^5 \times 12 \times 5 \times 40\ mega\ bytes \approx 6\ GB$$

For the fact table, we estimate that each year the approved loans will also increase by 20%. Thus, the fact table capacity in 5 years will by 600 MB.

$$(1.2 + 1.2^2 + 1.2^3 + 1.2^4 + 1.2^5) \times 60\text{mega bytes} \approx 600\ \text{MB}$$

So, in 5 years, the capacity required for the warehouse will be at least 6.6GB, which is very manageable in a regular server or a S3.

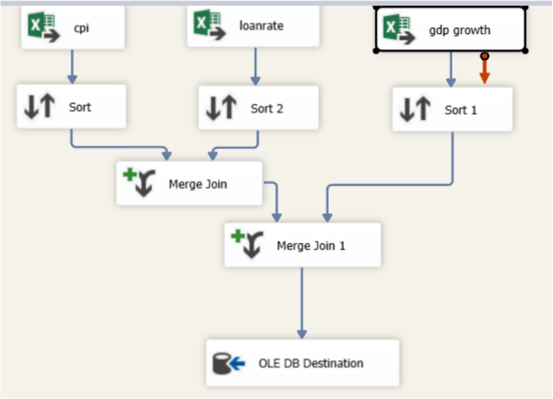
Other Considerations

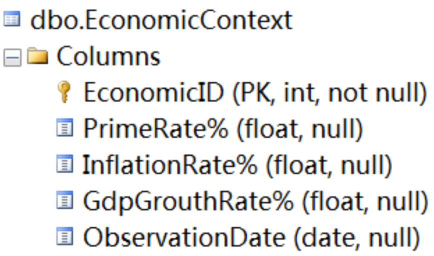
The source of our data warehouse is the public data received from Lending Club website. Normally, they update their data every two months. However, certain changes made by Lending Club, e.g.: Lending Club changes their data field definition or their data disclosing frequency, is beyond our control. We have a dimension LoanType, which includes all the predefined attributes from Lending Club (i.e. loan grade, loan subgrade, purpose, applicationtype). This way, if some definitions change, we only need to update it through the LoanType dimension.

<u>Implementation</u>

Economic Context Dimension Loading

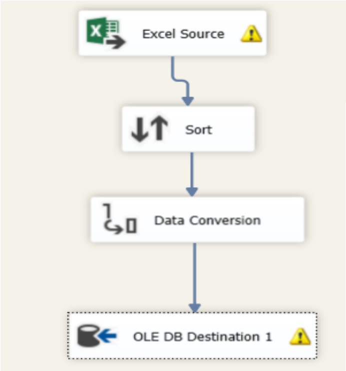
We have three excel files (CPI, bank prime rate, GDP growth) as our data source. In SSIS, we sorted and merged them by "Time", and then loaded it into the table EconomicContext. One issue here was that the time level of GDP growth is quarterly while that of CPI and loan rate is monthly. Our assumption here is that GDP tends to grow smoothly during the quarter and thus we can allocate the GDP growth into each month.

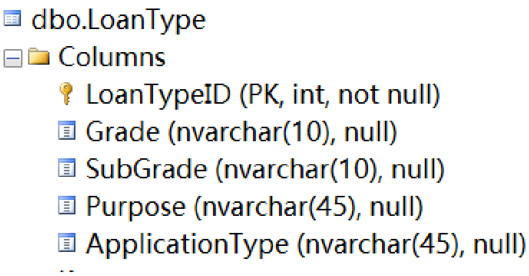



dbo.EconomicContext
Columns
EconomicID (PK, int, not null)
PrimeRate% (float, null)
InflationRate% (float, null)
GdpGrouthRate% (float, null)
ObservationDate (date, null)

Loan Type Dimension Loading

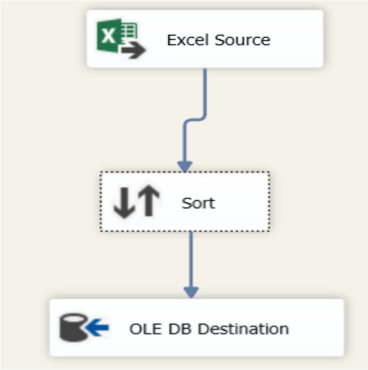
In this LoanType dimension, we have four attributes for each type of loan. We extracted all the combinations of this four attributes that appeared in the data source. One weird issue we faced here was to convert the varchar data type to unicode string, which should not have been an issue if we directly inserted our data into SQL server through SQL script.

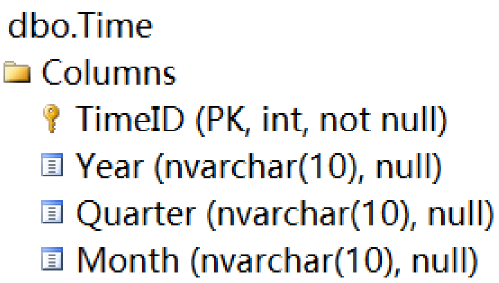



dbo.LoanType
Columns
LoanTypeID (PK, int, not null)
Grade (nvarchar(10), null)
SubGrade (nvarchar(10), null)
Purpose (nvarchar(45), null)
ApplicationType (nvarchar(45), null)

Time Dimension Loading

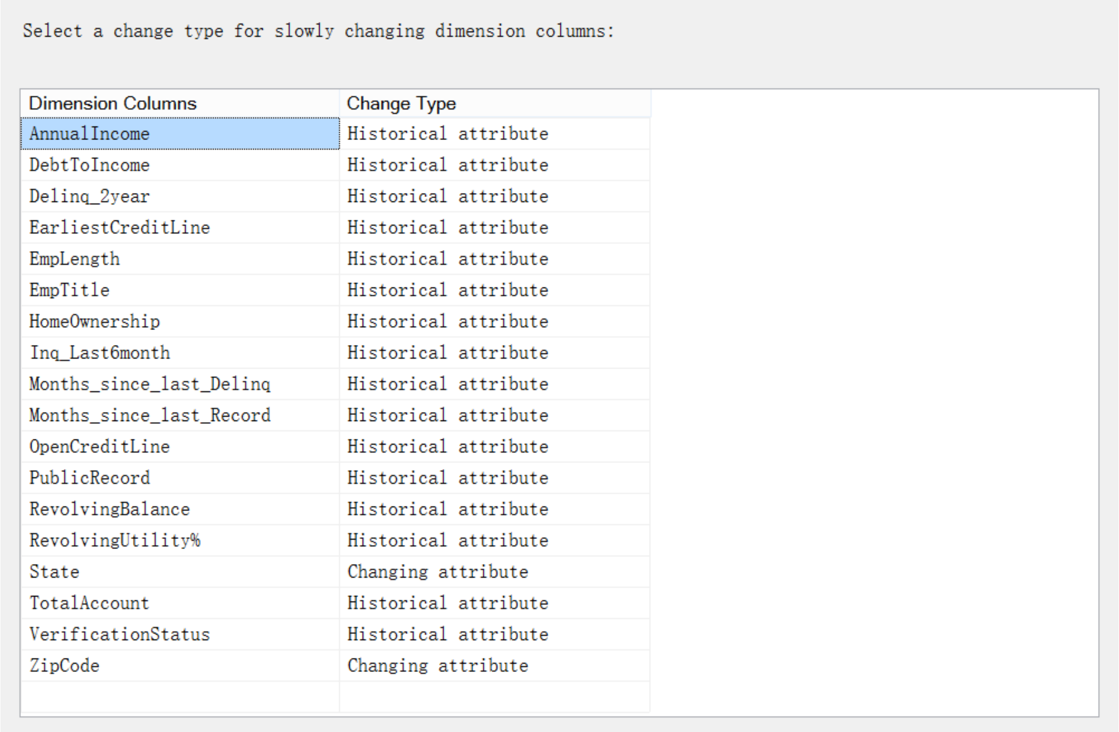
One defect of our data is that we only have monthly data. We selected an early time (1983-1) as our start time of record and 2017-12 as our end time of each record. This way, we don't need to frequently update our time dimension as time goes by. We decided to use a time surrogate key here because we don't want the key to be "smart".




dbo.Time
Columns
TimeID (PK, int, not null)
Year (nvarchar(10), null)
Quarter (nvarchar(10), null)
Month (nvarchar(10), null)

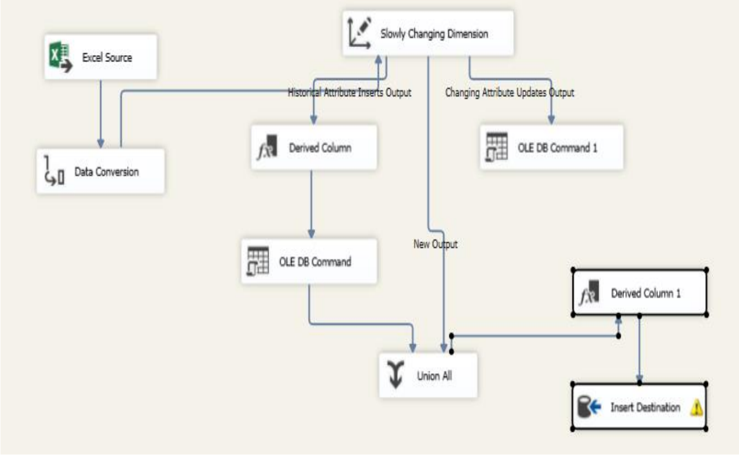
Borrower Dimension Loading

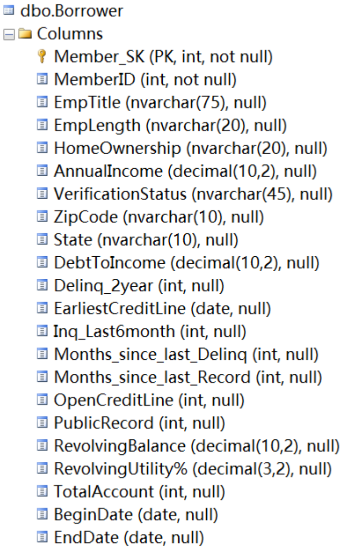
Because of the changing nature of attributes, we are interested in borrower and our business objective to keep track of borrowers' history. Most fields of our borrower dimension need to be "slowly changing". When loading data, we kept two fields ("State" and "ZipCode") as type 1 SCD and kept all other fields as type 2 SCD so that we can keep track of the information from borrowers. In detail, we use "BeginDate" and "EndDate" to record the time effectiveness of each row. We also built a member surrogate key to identify each row.



Select a change type for slowly changing dimension columns:

Dimension Columns	Change Type
AnnualIncome	Historical attribute
DebtToIncome	Historical attribute
Delinq_2year	Historical attribute
EarliestCreditLine	Historical attribute
EmpLength	Historical attribute
EmpTitle	Historical attribute
HomeOwnership	Historical attribute
Inq_Last6month	Historical attribute
Months_since_last_Delinq	Historical attribute
Months_since_last_Record	Historical attribute
OpenCreditLine	Historical attribute
PublicRecord	Historical attribute
RevolvingBalance	Historical attribute
RevolvingUtility%	Historical attribute
State	Changing attribute
TotalAccount	Historical attribute
VerificationStatus	Historical attribute
ZipCode	Changing attribute





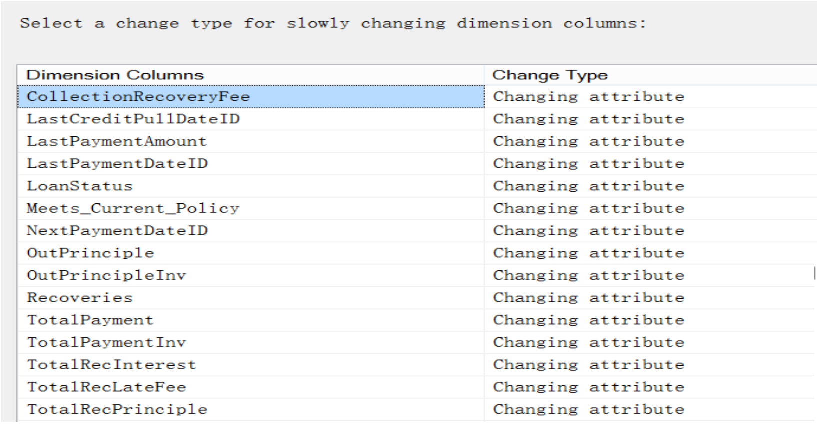
- dbo.Borrower
- Columns
  - Member_SK (PK, int, not null)
  - MemberID (int, not null)
  - EmpTitle (nvarchar(75), null)
  - EmpLength (nvarchar(20), null)
  - HomeOwnership (nvarchar(20), null)
  - AnnualIncome (decimal(10,2), null)
  - VerificationStatus (nvarchar(45), null)
  - ZipCode (nvarchar(10), null)
  - State (nvarchar(10), null)
  - DebtToIncome (decimal(10,2), null)
  - Delinq_2year (int, null)
  - EarliestCreditLine (date, null)
  - Inq_Last6month (int, null)
  - Months_since_last_Delinq (int, null)
  - Months_since_last_Record (int, null)
  - OpenCreditLine (int, null)
  - PublicRecord (int, null)
  - RevolvingBalance (decimal(10,2), null)
  - RevolvingUtility% (decimal(3,2), null)
  - TotalAccount (int, null)
  - BeginDate (date, null)
  - EndDate (date, null)

Fact Table Loading

Loading fact table in SSIS was non-trivial for us, because our fact table is in the form of accumulating snapshots, which means we also need to treat the fact table as SCD in SSIS. Since we used surrogate key in time dimension and we have many milestones in the form of date fields, we have to join time dimension each time we want to find the surrogate key of the date we want.

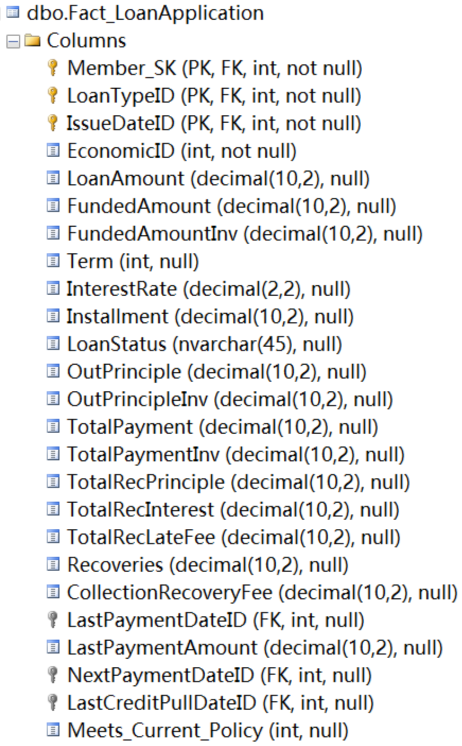
We also ran into a number of other challenges while loading the fact table.

- Data type conflicts:
  - The "decimal" data type is not corresponding to "decimal(m,n) " in SQL server. Later, we found out that "numeric" actually correspondes to "decimal(m,n)" in SQL server.
  - The "int" data type in SQL server corresponds to "four-byte signed integer" in SSIS.
- Time field aggregation:
  - In our time dimension table, "year", "month" are stored separately. Therefore, in order to join time dimension with the date appearing in fact table, we needed to first make a derived attribute combining "year" and "month" and then to convert the data type of that derived attribute in accordance with the date type in fact table.

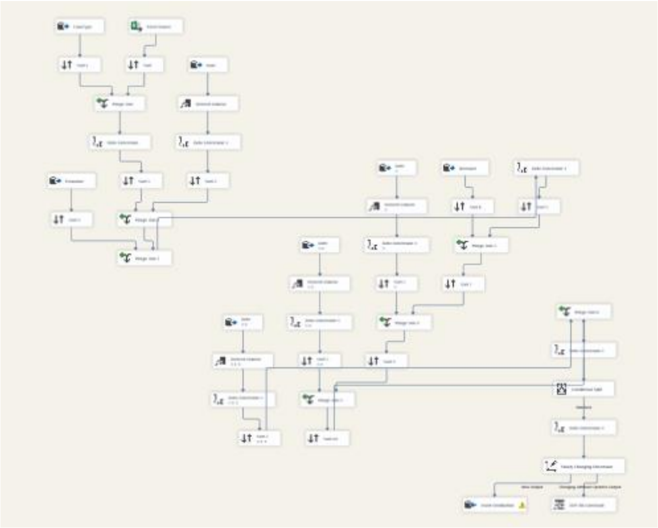


Select a change type for slowly changing dimension columns:

Dimension Columns	Change Type
CollectionRecoveryFee	Changing attribute
LastCreditPullDateID	Changing attribute
LastPaymentAmount	Changing attribute
LastPaymentDateID	Changing attribute
LoanStatus	Changing attribute
Meets_Current_Policy	Changing attribute
NextPaymentDateID	Changing attribute
OutPrinciple	Changing attribute
OutPrincipleInv	Changing attribute
Recoveries	Changing attribute
TotalPayment	Changing attribute
TotalPaymentInv	Changing attribute
TotalRecInterest	Changing attribute
TotalRecLateFee	Changing attribute
TotalRecPrinciple	Changing attribute



dbo.Fact_LoanApplication
- Columns
  - Member_SK (PK, FK, int, not null)
  - LoanTypeID (PK, FK, int, not null)
  - IssueDateID (PK, FK, int, not null)
  - EconomicID (int, not null)
  - LoanAmount (decimal(10,2), null)
  - FundedAmount (decimal(10,2), null)
  - FundedAmountInv (decimal(10,2), null)
  - Term (int, null)
  - InterestRate (decimal(2,2), null)
  - Installment (decimal(10,2), null)
  - LoanStatus (nvarchar(45), null)
  - OutPrinciple (decimal(10,2), null)
  - OutPrincipleInv (decimal(10,2), null)
  - TotalPayment (decimal(10,2), null)
  - TotalPaymentInv (decimal(10,2), null)
  - TotalRecPrinciple (decimal(10,2), null)
  - TotalRecInterest (decimal(10,2), null)
  - TotalRecLateFee (decimal(10,2), null)
  - Recoveries (decimal(10,2), null)
  - CollectionRecoveryFee (decimal(10,2), null)
  - LastPaymentDateID (FK, int, null)
  - LastPaymentAmount (decimal(10,2), null)
  - NextPaymentDateID (FK, int, null)
  - LastCreditPullDateID (FK, int, null)
  - Meets_Current_Policy (int, null)



Analysis & Results

I) Trend Analysis of Delinquency rate over the past years: Lending Club was founded in 2007. From Fig.1, it is clear that there was a lot of fluctuation in delinquency rate in the initial years. However, starting 2009 we see that there is stability in delinquency rate which means that lending club has become cautious in loan approvals.

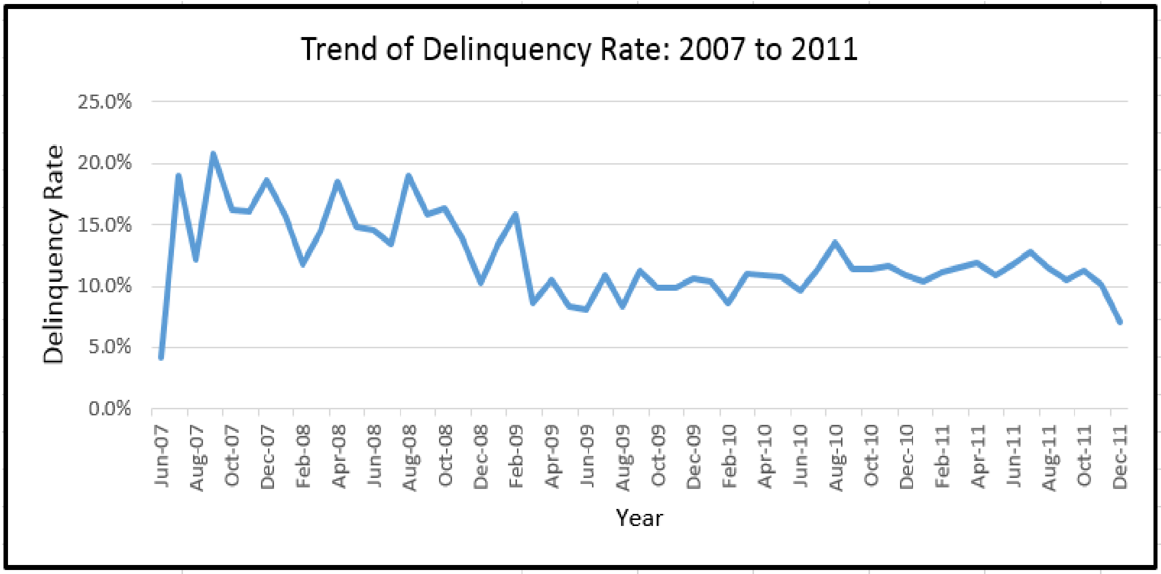


Fig. 1: Delinquency Rate from 2007 to 2011

II) Interest rate vs Grade of the loan: From Fig.2, we see that the interest rate varies across every loan grade. Interest rate increases as grade of the loan varies from A to G. Note that this is due to the fact that A grades are assigned to less risky borrowers, whereas G grades are assign to risky borrowers. Thus, lending club requires a higher rate on the more risky customers.

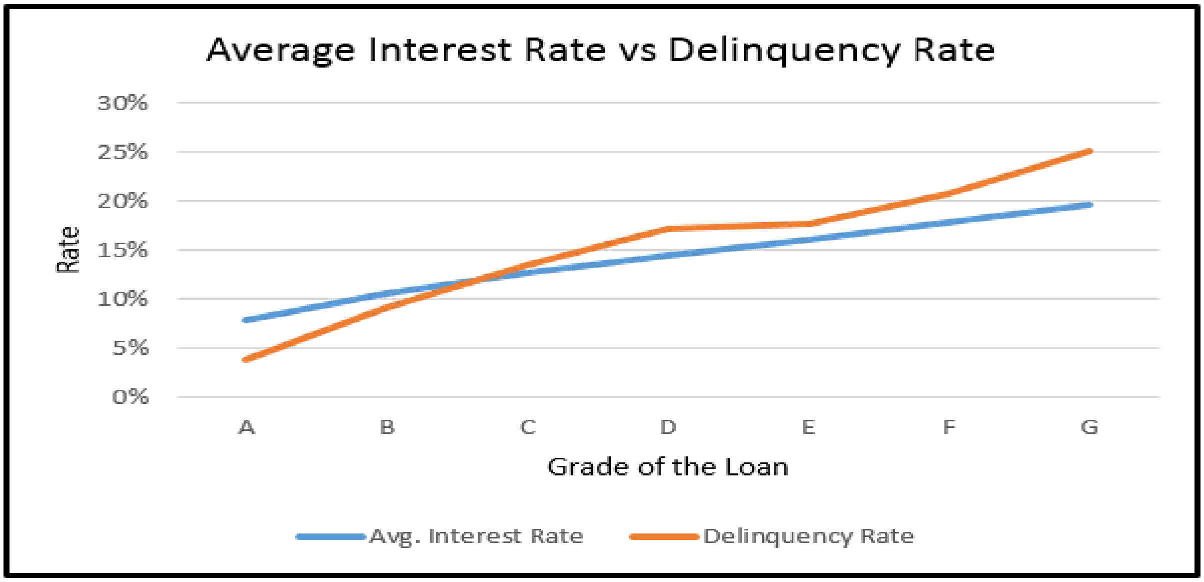


Fig. 2: Average Interest Rate and Delinquency Rate across different Grades

III) Percentage of loans charged off across different grades: From Fig.3, we can get an idea of the number of loans issued and the number of loans charged off (the borrower cannot make any more payments). We can also see that as grades varies from A to G, there is an increase in the percent of loans charged off.

Percentage loans charged off = Number of loans charged off/Number of Loans

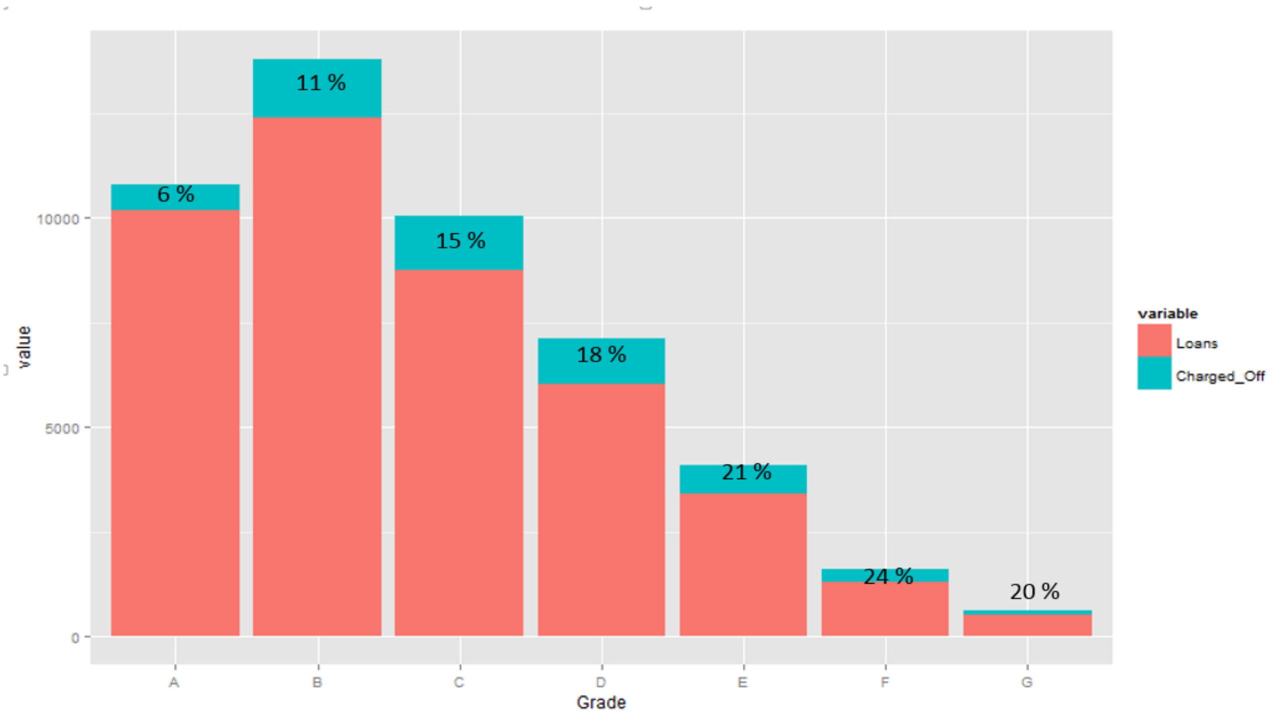
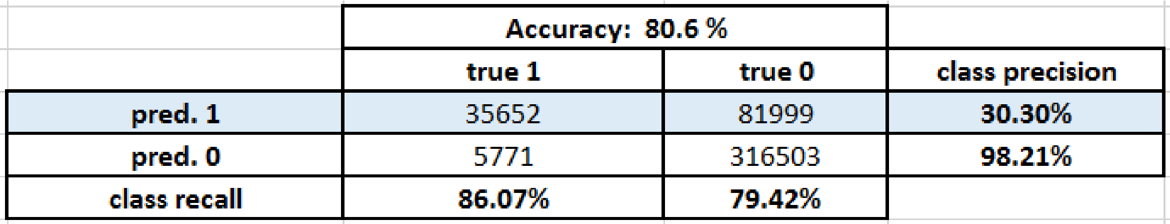


Fig. 3: Proportion of Loans Charged off across different grades

IV) Predicting Approval/Denial of a Loan: To identify the key parameters which contribute to a loan approval or denial, we built a decision tree. From the decision tree, we can interpret that Employment years and DTI (debt to income ratio) are the most crucial factors in loan approvals or denials. The model we built was able to predict loan approvals or denials with an accuracy of 81%. Here is the decision on tree of the best model that we built:

Here is the confusion matrix of our best model:



	Accuracy: 80.6 %		
	true 1	true 0	class precision
pred. 1	35652	81999	30.30%
pred. 0	5771	316503	98.21%
class recall	86.07%	79.42%	

Table. 1: Confusion matrix of our best model

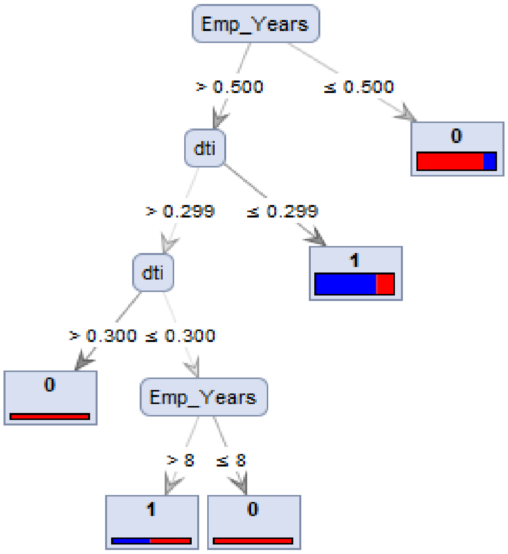
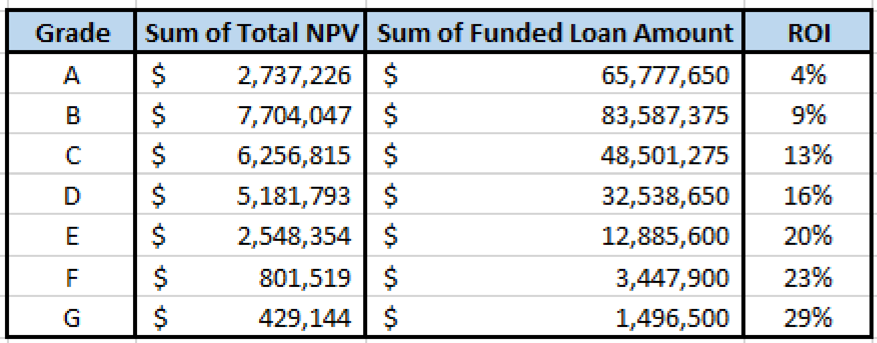


Fig. 4: Decision Tree to predict loan approval or denial

V) Finding the Net Present Value: For a company to be profitable it has to stay ahead of inflation rate. To understand by how much is the company profitable we calculated the net present value for each loan approved at a year level. Based on the interest the borrower (loans are paid over a span of 3 or 5 years) is paying every year, we calculated the NPV in the current year (in this data the current year is 2011).

From the figure, we see that loans with grade G has the highest Return on Investment which is not quite surprising as grade G loans are charged a much higher interest rate.

ROI = Net Present Value / Funded Amount



Grade	Sum of Total NPV	Sum of Funded Loan Amount	ROI
A	$ 2,737,226	$ 65,777,650	4%
B	$ 7,704,047	$ 83,587,375	9%
C	$ 6,256,815	$ 48,501,275	13%
D	$ 5,181,793	$ 32,538,650	16%
E	$ 2,548,354	$ 12,885,600	20%
F	$ 801,519	$ 3,447,900	23%
G	$ 429,144	$ 1,496,500	29%

Table. 2: Net Present Value and ROI across different loan grades

Findings:

We identified the key factors (from the predictive model) which lending club thinks are important in approving loans. Our company Funding club wants to leverage the fact that Lending club has become extra cautious in approving loans (Fig.1). Thereby, we would like to consider the top 5% of the people for whom lending club rejected a loan and charge them a higher interest rate. We would also want to compete for qualified candidates. Through our analysis, we can see that customer across all grades are profitable. Hence, as a company, we are willing to offer a slightly better rate to these borrowers to draw them to Funding Club.

Conclusion

As a new start-up in 2011, Funding Club is looking for opportunities to enter the online lending marketplace. To do so, we extracted Lending Club borrower data from 2007 to 2011 to analyze their customers and lending criteria. With our business purpose in mind, we leveraged dimensional modeling to design out database. We identified the appropriate grain, facts, four dimension tables and one fact table. We then performed data transformation and successfully implemented ETL using SSIS.

With the data available to us, we created a model with 81% accuracy to predict whether customers will be approved or declined at Lending Club. Using this information, we can target the borrowers differentially. For the customer who are most likely not going to be approved by Lending Club, we would consider the top 5% of these customers in terms of qualifications and offer them loans at an annual interest rate excess of 30%. Through our analysis, we saw that qualified customers are quite profitable in terms of NPV at every loan grade and since Lending Club is tightening their policies, it offer us an opportunity to come in and compete with a lower interest rate. We know that might mean lower profitability at the start, but we are willing to sacrifice profitability to build the Funding Club brand and initial customer base.

There are a number of challenges that we ran into during the process. We have highlighted the top four in the section below:

- Data manipulation in SSIS is not as flexible as in Excel. So, we did much of our data manipulation in EXCEL rather than in SSIS, which is not very scalable or automatic when our data volume becomes increasingly larger.(e.g. we have to transform the date type in excel and then loaded it using SSIS)
- When we loaded the fact table, we needed to find a way to treat our accumulative snapshot fact table. We treated the fact table as type 1 SCD and each time the fact table changed, we updated the corresponding row in the fact table.
- When we are tried to evaluate if one loan is profitable or not for Lending Club, we tried to find one standard approach. We extracted economic context data from external data source and used it as a baseline of loan rate and interest rate. Based on that, we calculated the NPV (Net Present Value) and ROI (Return on Investment) of each loan.
- There were many attributes available for approved loans. However, for declined loans many attributes were not available. If we had more information about the loans that were declined, we would be able to predict approved/declined decision for a loan with a better accuracy.
- We tried to look into the impact factors that may be used by Lending Club to grade each loan approved by them. We tried building a predictive model for the same but it did not give us good accuracy.