Week 1 Report

Kristina Schiffhauer, Sagar Dubey, Hewitt Trinh, Jinwoo Oh, Stefanie Montgomery

## 1. Data Acquisition

Data attributes can be categorized into three broad types, as seen below.

1. Borrower characteristics: This category includes attributes such as FICO score, employment status, and annual income. They are characteristics that relate to the borrower and are known at the time a loan will be funded. These attributes will be key when constructing models to predict some unknown KPI for loans Jasmin might want to invest in.
   * Key attributes: annual\_inc, delinq\_2yrs, dti, earliest\_cr\_line, emp\_length, fico\_range\_high, fico\_range\_low, home\_ownership, loan\_amnt, open\_acc, pub\_rec, purpose, revol\_bal, revol\_util, verification\_status, term, funded\_amnt\*, issue\_d\* (\*These are not technically borrower characteristics, but they are known at the time of loan issuance, so we are characterizing them as such).
2. Platform decisions: These are attributes such as loan grade and interest rate, and are determined by the borrowing platform. As these will also be known to investors before a loan is issued, these attributes have significant potential predictive power. It is worth noting that the initial interest rate assigned by LendingClub is subject to change throughout the life of the loan. We will discuss this more later.
   * Key attributes: int\_rate, grade
3. Loan performance: This category includes attributes related to the performance of the loan post issuance, such as status and total payment amount. These attributes are not known at the time investors would be considering loans, and thus should not be used in predictive models. However, these attributes will be critical to study using descriptive analytics in order to form hypotheses about “good vs. bad” investments. It is worth noting here that the dataset does not contain a variable for returns on completed loans. Instead, we will need to define and calculate return variables in order to quantify the “goodness” of each loan.
   * Key attributes: installment, last\_pymnt\_d, loan\_status, recoveries, total\_pymnt

As an investor, the above categories of attributes are important to Jasmin for different reasons. Borrower characteristics and platform characteristics (those which remain static throughout the life of the loan) will serve as the features for any predictive models that Jasmin builds. Though loan performance attributes are not known at investment time, these will be key for building derived return variables used to train models.

## 2. Business Understanding

There are two key decisions that Jasmin will need to make. The first is the amount of money she wants to invest in LendingClub. Once she has decided how much is right for her, she will need to decide which loans to invest in and how much to invest in each. This is the main decision we will be focusing on for our analysis. She will need to determine which combination of borrower characteristics, platform decisions, and loan performance will help her achieve her objective - to maximize the amount of money returned. This objective will be subject to a number of constraints, including the amount of risk she is willing to take on in her investments. We will dive deeper into these constraints in a future report.

In order to be confident in her model’s ability to determine the optimal investments, Jasmin will need to be able to evaluate its performance before making any real-life decisions. She can handle this by splitting her data into two parts - one for training the model and one for evaluating its performance. While this sounds very straightforward, there are a number of considerations that will need to go into deciding how to split the data in half. Since the data is spread across so much time and has a lot of nuance, she will need to consider factors such as the length of the loan tenure and how that impacts the quality of investment, the time to default of loans that do, and whether early loan repayment has a significantly negative impact on a loan’s appeal. We will also dive into the impact of these considerations on the data splitting quality in a future report.

When it comes time to actually evaluate performance, Jasmin will need to have KPIs for determining the model’s success. These KPIs will have varying levels of straightforwardness and complexity, but all can be important to her final decisions. The first and most straightforward is the accuracy in predicting whether a loan will default. If it does default, she will also want to be able to predict when it will happen. On the other side, if a loan is predicted to be repaid, she will want to be able to have accuracy around when and whether it will be before the end of the loan term. Finally, she will want to be able to predict the rate of return. Evaluating these KPIs should give Jasmin a good indication of her model’s ability to achieve her objective.

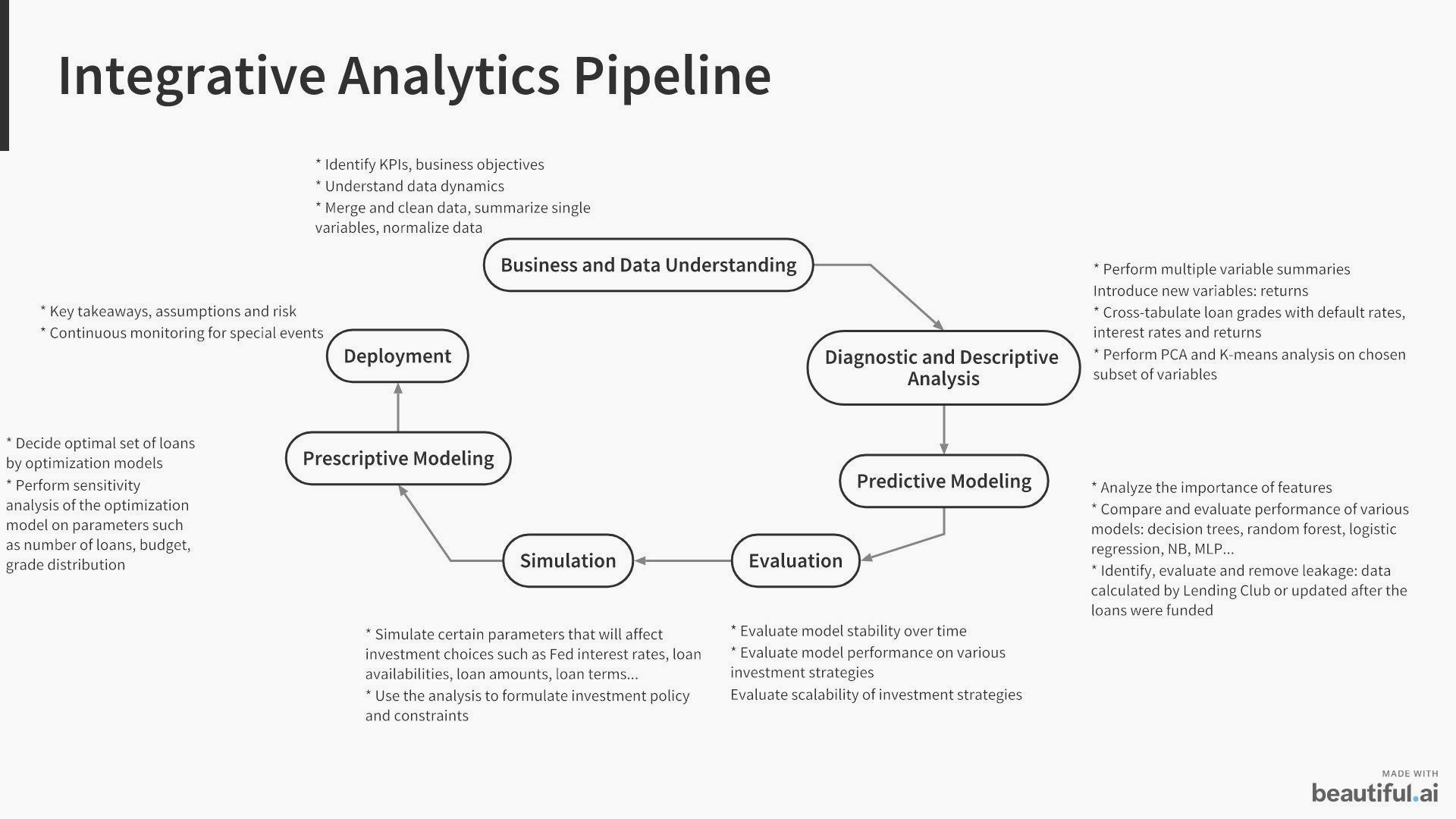
## 3. Sources of Value in the Data

As discussed previously, we believe Jasmin’s objective is to maximize the returns out of the loan portfolio she will invest in, subject to some risk constraints. Past data provides us with the ability to build models to learn the characteristics of “good” loans and then build models to predict “goodness” of loans we might invest in.

The below table maps various types of models we plan to explore in this case to specific business use cases and expected data to be used in each.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Description** | **Analytics Techniques** | **Use Cases + Data Used** |
| **Descriptive** | **Data Exploration** | General data analysis including data cuts, time trends, geographic plots, word clouds, and dashboards | Sample min/max, mean, variance, skew | Preparation for predictive modeling and KPI choice (exploration of all three attribute types). |
|  | **Inferential Statistics** | Drawing conclusions about populations based on observed sample | Hypothesis testing  Anomaly detection | Understanding of variable distribution and skew in preparation for variable transformation for use in further analysis (log or inverse transformations, etc.). |
|  | **Pattern Mining** | Discovering patterns in the data, e.g. ▪ similarity among records ▪ relations among variables | Clustering (K-Means)  Dim reduction (PCA) | Understanding of loan grade structure - What borrower characteristics are associated with each grade? Attributes used will be borrower characteristics.  Understanding of defaulted and early payback loans - what borrower and platform characteristics lead to these outcomes? This information will be used to inform feature selection for predictive models. |
| **Predictive** | **Categories** | Predicting binary or ordinal outcomes, e.g. product choice, churn, fraud, purchase events, etc. | Logistic regression  Decision trees  Discriminants, SVM  Non-parametric (kNN) | Predict whether a loan will default or be paid back early, using borrower and platform characteristics as predictive features. Test against derived return variable(s) created using loan performance attributes. |
|  | **Values** | Predicting continuous outcomes, e.g. revenues, time, growth rates, etc | Regression: linear, ridge, lasso, panel, etc.  Time series analysis  Neural networks  SVM | Predict probability of default and probability of early payback, using borrower and platform characteristics as predictive features. Test against derived return variable(s) created using loan performance attributes. |
| **Prescriptive** | **Optimization** | Identify set of parameters that optimize system performance under given constraints | Continuous: Linear Programming  Discrete: Integer optimization | Use previously created predictive models to assign default probability to each loan. Use this probability along with borrower and platform characteristics as input to a knapsack optimization problem to maximize returns. |
|  | **Simulation** | Simulate how a complex system will behave under wide range on scenarios | Monte Carlo | Simulate how various levels of risk tolerance affect our expected return. |

## 4. Integrative Analytics Pipeline

****

The workflow above is our tentative analysis pipeline for the project. It is worth noting here that we plan to use simulation techniques in tandem with optimization models to arrive at the optimal set of loans for Jasmine to invest in. The ideas behind simulation tasks in this pipeline are to incorporate the uncertainties around parameters that ultimately affect optimal investment choices, thus definitely affecting the final business objective, return on investment. Those parameters are, tentatively, general economic indicators such as Fed interest rate, inflation rate and other loan availability indicators such as availability of loans amount and loan terms in each grade. Understanding the extent to which the fluctuations of those parameters to final outcomes will help us come up with sound investment policies that can weather different scenarios. Of course, we proceed under the assumption that no Black Swan events such as Covid-19 outbreak or 2008 financial crisis will occur during the investment timeframe.

## 5. Data Ingestion

We limited data to 2015 and 2016 due to file size. We will explore adding additional years of data later.

## 6. Data Dynamism

Assuming that the loans are independent of one another, we will look at each loan individually rather than groups of loans. We have to be wary of the effect of *leakage*, guarding against using attributes that are not available at the time of loan application, yet are strong indicators of the loan performance. The total payment made on a loan is a case in point. As such, all the loan performance indicators can cause leakage. However, it is safe to assume that borrower characteristics will be available at the time of loan application and will remain static. Certain platform decisions such as interest rate are subject to change over the life of the loan, and may introduce leakage into predictive models.

## 7. Data Cleaning

We did the data exploration using both Tableau and Python, and performed data cleaning in three steps:

a) Identify and remove outliers such as high values of annual income, revolving balance, dti, and revol\_util. The figure below shows these variables before the outliers were removed. We also removed rows having null value in any of the numeric columns.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

b) Log-transform numeric variables such as annual\_inc, revol\_bal , open\_acc, and total\_pymn that showed high negative skewness. The effect of log-transforming annual income is shown here.

|  |  |
| --- | --- |
| Annual Income (Before log-transform) | After log transform |

c) Normalize all the numeric variables using min-max scaling. It will ensure no column has an outsized impact on the prediction outcome because of its potential high value.

We also saw that some features such as delinquencies and public records do not show much variability, so we may drop them in a later feature evaluation exercise. Additionally, we may want to bin features such as recovery fees, FICO scores, annual income to handle outliers in a more graceful way.