Lending Club Case Study

Identify patterns among loan defaulters for consumer finance company

Date: 7 March 2022

Presenters:
Mohini Aggarwal
LokSundar Ganthi

Contents

Objective

Business Objectives

- Lending loans to risky applicants is the largest source of financial loss for loan marketplaces. Minimize the risk of losing money while lending to customers by identifying patterns which indicate if a person is likely to default.
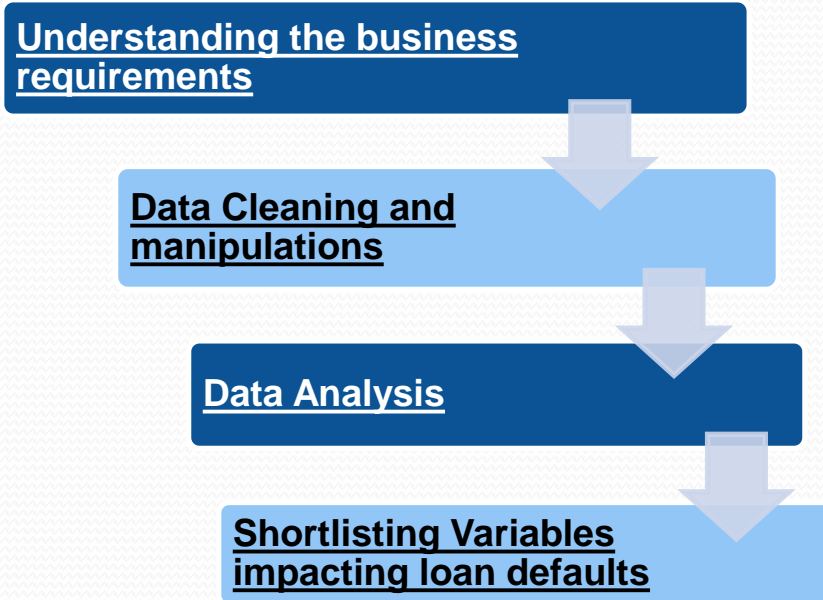- The company can utilize this knowledge for its portfolio and risk assessment.

Data Period and data fields

- **Data Period : 4 years of data from 2007-2011**
- Data contains information about past loan applicants and whether they 'defaulted' or not.
- The analysis will not consider consumer variables since they were not available at the time of loan application.
- Rejected loans will not be a part of analysis since there is no transactional history of those applicants with the company.
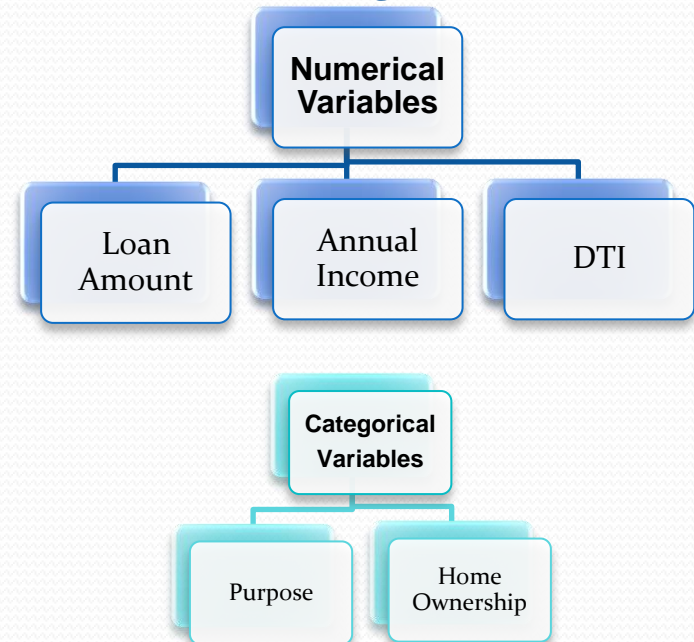
Approach?

| Understanding the requirement | Data cleaning | Data analysis | Recommendation |

Analysis Overview

Process flow

Understanding the business requirements

⬇

Data Cleaning and manipulations

⬇

Data Analysis

⬇

Shortlisting Variables impacting loan defaults

Output

Variables identified as strong indicators of default

Numerical Variables
- Loan Amount
- Annual Income
- DTI

Categorical Variables
- Purpose
- Home Ownership

Input:

Data File Format: **MS Excel file**

Software platform used: **Google Cloud (Colab (Python) & Drive), CSV Files, MS PowerPoint, Adobe PDF**

Importing and Cleaning Data

- Understanding the meaning of variables, their data types and the business requirement.

- Removing rows and columns with less than 85% data.

- Removing rows and columns that are irrelevant to the analysis.

- Data manipulations to impute missing values.

Data Analysis (1/3)…

Univariate/Bivariate Analysis : Exploratory data Analysis to understand the data distribution. A few samples mentioned below:

- Fig 1 and 2 show distributions for 'density' v/s 'instalment' and 'density' v/s 'loan_amnt'
- Fig 3 shows a pivot table for different types of 'home_ownership'
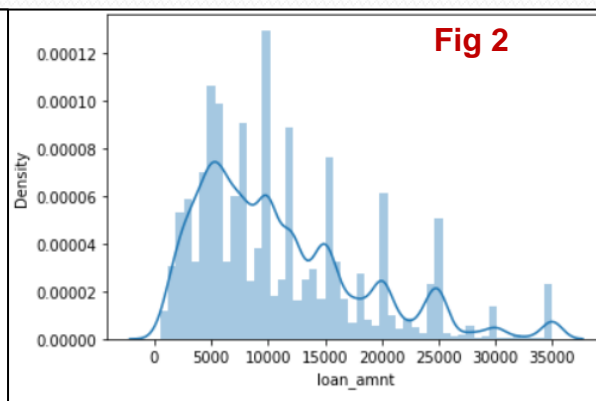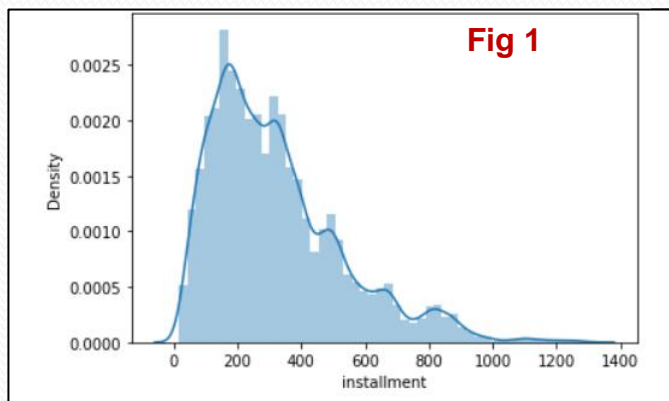- Fig 4 shows 2 way pair plots between 5 selected variables with each other

Fig 1

Fig 2

Fig 4

Fig 3

| home_ownership	member_id	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti
MORTGAGE	855012.431878	12468.973621	12113.559720	11517.578154	354.416023	83021.061885	13.110272
NONE	152646.666667	5600.000000	5600.000000	1592.686667	176.426667	80733.333333	11.746667
OTHER	431096.836735	10663.010204	10148.724490	7354.949676	338.745714	71309.714286	11.251939
OWN	837544.653445	10138.100840	9929.521008	9470.189809	298.636545	58888.182797	13.182171
RENT	833616.640043	9886.584145	9701.300054	9167.347334	296.812659	57236.120159	13.447898
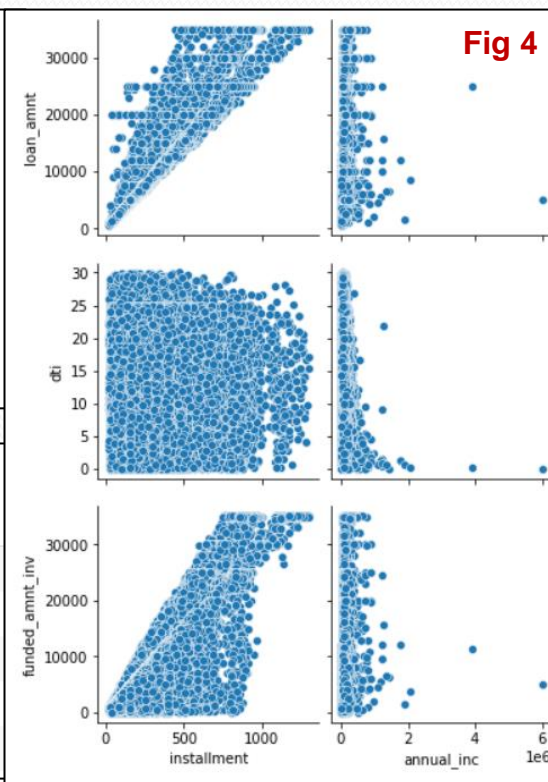
Data Analysis (2/3)…

- Trend of 'purpose' with 'dti' displaying defaulters dominating the non-defaulters for different purpose types

```
[65] fig_dims = (25, 5)
     fig, ax = plt.subplots(figsize=fig_dims)
     sns.lineplot(data=df_credit, x="purpose", y="dti", hue="loan_status")
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f402fa61750>
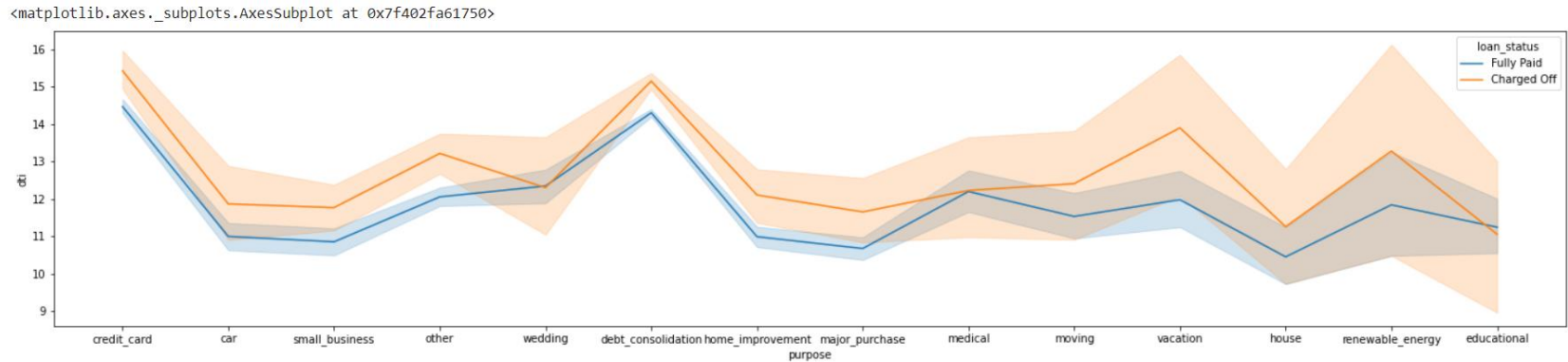
- Trend of 'purpose' with 'instalment' for defaulter and non-defaulters for different purpose types

```
[66] fig_dims = (25, 5)
     fig, ax = plt.subplots(figsize=fig_dims)
     sns.lineplot(data=df_credit, x="purpose", y="installment", hue="loan_status")
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f402fad4250>
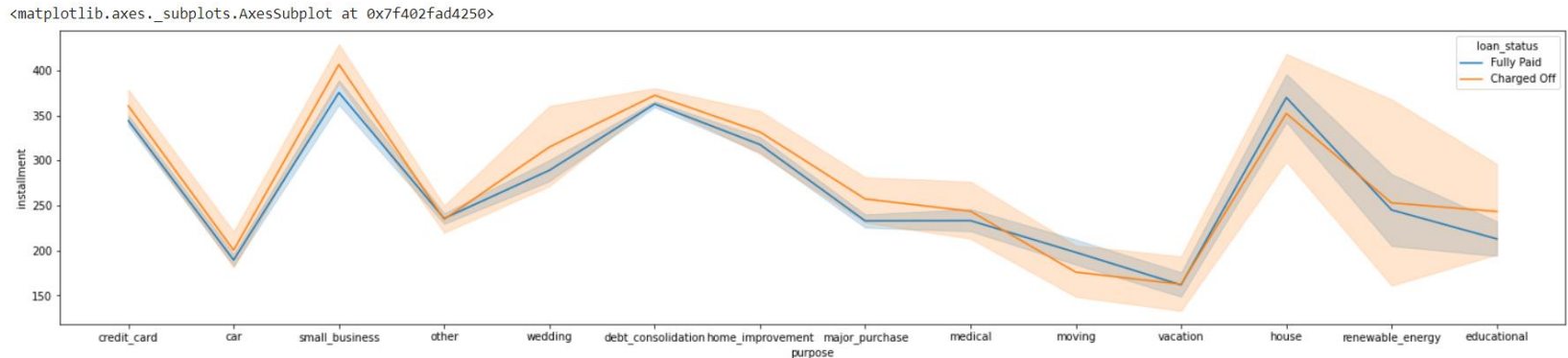
Data Analysis (3/3)…

- Fig 1: Majority of customers are either on rent or have mortgaged property.

- Fig 2: Correlation Comparison: Threshold 30%
  - Displays that 'loan_amnt' and 'instalment' are highly correlated.
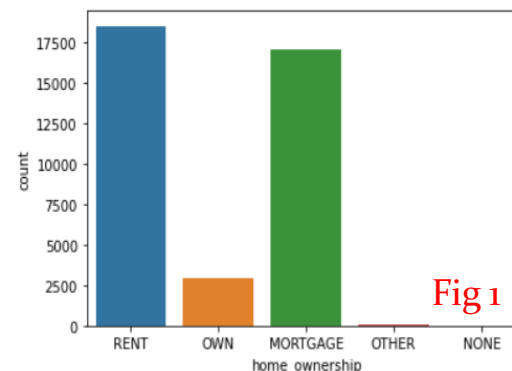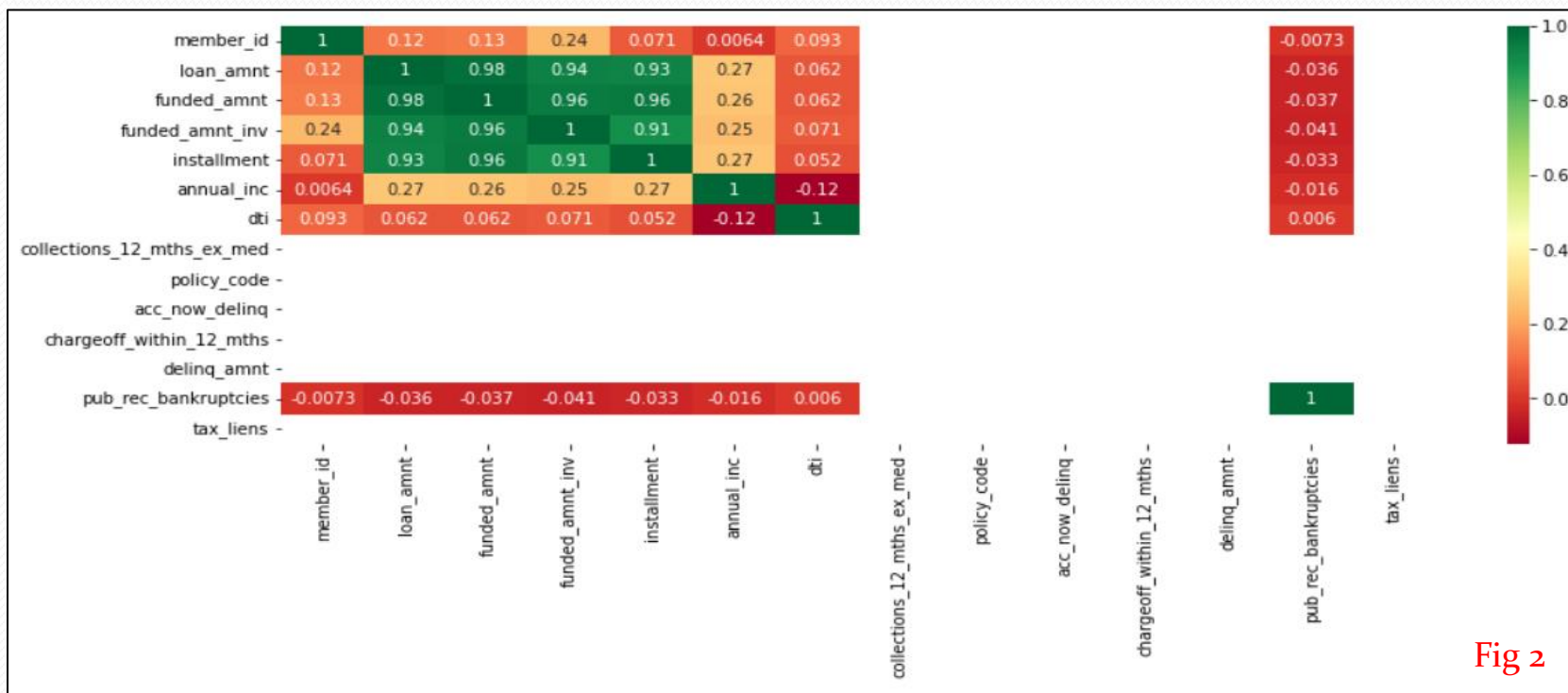 - DTI: High risk customers for loan repayment, not correlating to other variables.

Fig 1

Fig 2

Conclusion

Strong indicators of default:

- **Numerical Variables:**

  - **Loan amount:** Loan amount seems to be one of the important factors as it is highly correlated with 'funded_amnt', 'funded_amnt_inv' and 'installment' which implies that it explains the impact of all these variables.

  - **Annual income:** This variable does not correlate with any other factor considering a threshold of 30% and can impact the default risks.

  - **DTI**: It is not correlated to any other factors and it provides the financial risk appetite of the customer for loan repayment.

- **Categorical Variables:**

- **Home ownership:** Most customers taking the loan are living on rent and have a relatively lower annual income (<100K) which implies risk of loan defaults is high.

- **Purpose:** The trend clearly shows the purpose of the loan where people have been 'charged-off' is more than those who repaid.

Thank you!
Questions?