|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  | | | | | | | | |
| **An Intelligent Lending Consulting Tool** | | | | | | | | |
|  | | | | | | | | |
| Applied Project Final Report | | | | | | | | |
|  | | Name: | | ： | Yihuan Chen | |  | |
| Class: | | ： | 4100 Applied Project | |
| Semester: | |  | Spring 2023 | |
| Date: | | ： | 04/29/2023 | |
|  | | ： |  | |
| A paper submitted in partial fulfillment of the requirements for the degree of Master of Science in Management and Systems at the Division of Programs in Business School of Professional Studies New York University | | | | | | | | |

**Table of Contents**

[1. Declaration 1](#_Toc133682125)

[2. Business Background 2](#_Toc133682126)

[3. Project Overview 3](#_Toc133682127)

[4. Project Charter 4](#_Toc133682128)

[5. Project Plan 10](#_Toc133682129)

[6. Literature Review 11](#_Toc133682130)

[6.1 Loan Management 12](#_Toc133682131)

[6.2 Fraud Detection 13](#_Toc133682132)

[6.3 Patent Evaluation 14](#_Toc133682133)

[6.4 Risks Management 15](#_Toc133682134)

[6.5 Alternative Data 16](#_Toc133682135)

[6.6 Conclusion 17](#_Toc133682136)

[7. Lessons learned 19](#_Toc133682137)

[8. Project Chronology and Critique 27](#_Toc133682138)

[8.1 Project Chronology 27](#_Toc133682139)

[8.2 Critique and Improvement: 47](#_Toc133682140)

[9. Sponsor Acceptance Form 56](#_Toc133682141)

[References 57](#_Toc133682142)

# Declaration

I grant powers of discretion to the Department, SPS, and NYU to allow this paper to be copied in part or in whole without further reference to me. This permission covers only copies made for study purposes or inclusion in the Department, SPS, and NYU research publications, subject to customary acknowledgment conditions.

# Business Background

**The company name "the client":**  
IDEAS (International Data Engineering And Science Association)

**Company location**

Los Angeles, California

**Name and role of your proposed project sponsor:**

Jason Geng, Chair of the Board at IDEAS

**Sponsor's location:**

Los Angeles, California

**Your relationship with the client and sponsor:**

Intern and mentor

**Description of the client's business:**  
IDEAS is a non-profit organization. It bridges the gap between academia and industry. IDEAS' vision is to foster the data engineering and data science ecosystems and broaden the adoption of their underlying technologies, thus accelerating the innovations data can bring to society.

# Project Overview

Throughout the LendingClub Risk Assessment Model Project, I completed several essential tasks utilizing various approaches, methods, and technologies. The first thing I did was research LendingClub's background and collect historical loan data. Following this, I explored, validated, and cleaned the data, ensuring it was well-structured for modeling.

I trained the risk assessment model using Python using the XGBoost algorithm and evaluated its performance using ROC-AUC scores. Based on the model's performance, it is clear that it can accurately assess risk for unseen data.

As a final step, I created two Excel interfaces for Ipysheet and Pyxll that allow loan application data to be input and risk assessment results to be received. For LendingClub investors, a comprehensive risk assessment solution was developed through research, data preparation, modeling, and Excel interface development.

# Project Charter

**Business Problem/Opportunity:**

LendingClub is a financial services company. It was the first peer-to-peer lender to register its offerings as [securities](https://en.wikipedia.org/wiki/Security_(finance)) and to offer loan trading on a secondary market. Lending Club would use this money to invest in different loans, such as student loans. Lending Club would offer grade options from A to G, which return from high to low. However, Lending Club didn't show the risk level of each investment. I'm making models to calculate the risk percentage for investors, so they can reduce the potential risks.

**Project Goal:**

1. The project's primary purpose is to produce a model that investors can use to determine the potential risk of a specific loan application.

2. This project is for IDEAS rather than Lending Club, and LendingClub's only role is to provide historical loan data.

3. This project may be used by IDEAS for any of their ongoing entrepreneurial efforts.

**Project Description:**In this project, I will collect data from LendingClub's historical transaction data and build a machine-learning model that helps select a better investment portfolio. I will also use the model to create an Excel sheet that allows clients quickly enter their information and receive the result. The Excel sheet tests the model's functionality to see if the model is working.

Firstly, I would look at the data, organize the data into data frames, and do data exploration and preparation. Secondly, I would build a classification machine learning model to classify each loan. Thirdly, after data analysis, the result would be the risk percentage.

**Project Sponsor:**

* Name and Title: Jason Geng
* Role within the organization: Chair of the Board

**Project Objectives:**

1. Complete data cleaning and analysis of LendingClub's historical data to classify the degree of risk for different categories of applications by the end of Project Status Report A, Mar 9, 2023.

1. Build a binary classification model to determine the criteria for accepting loan applications by the end of Project Status Report B, Apr 6, 2023.

1. Create an Excel sheet connected to the model for users to enter loan application data and receive a suggested determination based upon using the created model by the end of the project, Apr 20, 2023.

**Project Scope**

|  |
| --- |
| **In-Scope Activities:** |
| * I would look at the data, organize the data into data frames, and do data exploration and data preparation * Build a classification machine learning model that will determine the criteria for accepting loan applications * Create an Excel sheet connected to the model for users to enter loan application data and receive a suggested determination based on the use of the created model |
| **Out-of-scope activities:** |
| * Reconsider project idea * Unplanned work and expenses. * Additional tasks or changes to the requirements |

**Risks and Mitigation Strategies**

|  |
| --- |
| * Risk: The sponsor change his mind about the project in the middle of the semester. * Strategy: Discuss with the professor whether we could continue the original project. |
| * Risk: The sponsor took a leave of absence for some reason, and I couldn't communicate with the sponsor. * Strategy: Discuss with the professor and keep following up with the sponsor. |
| * Risk: I delay my work for personal reasons, such as sickness or emergency. * Strategy: Let the professor and sponsor know. Reschedule the project plan if possible. |

**Communication Plan**

1. Frequency: Every two weeks
2. Method: Zoom
3. Content: Keep tracking the project

**Schedule Overview**

|  |
| --- |
| **Project Start Date:**  2/05/2023  **Estimated Project Completion Date:**  05/04/2023 |
| **Major Milestones** |
| Finish data exploration and submit Project Status Report A(3/9) |
| Finish modeling and Submit Project Status Report B(4/6) |
| Create an Excel for input and result (4/17) |
| Submit Final Report (4/27)  Client Presentation (5/4) |
| **External Milestones Affecting the Project, if any** |
|  |
|  |
| **Impact of Late Delivery**  The late delivery results in the delay of the project schedule. |

# 5. Project Plan

Chart, timeline

Description automatically generated

Fig. 5-1

# 6. Literature Review

**Machine Learning in Fintech**

Recently, the financial industry has undergone significant changes, including technological advancements due to the integration of Machine Learning. The utilization of machine learning in the finance industry, also known as Fintech, can revolutionize the delivery of financial services due to improved cost savings, efficiencies, and Management of financial risks. The reason for reviewing this literature is to examine the benefits and risks of integrating machine learning in financial industries. The source selected for the literature review were chosen based on their relevance to the research topic, the author's credibility, publication, and the fact that they are current sources. The literature review focuses on using machine learning in financial services, such as managing loans, evaluating patents, fraud detection, and risk management. The selected sources are analyzed and compared in the literature review to identify themes and conclusions associated with the benefits and risks of applying machine learning in Fintech. By examining the available literature on the topic, this review aims to comprehensively understand how machine learning is used in Fintech and its potential implications in the financial industry.

Machine learning has wide applications in Fintech, from Loan Management, Fraud Detection, Patent Evaluation, Risks Management, and Alternative data. This literature review provides an in-depth analysis of the following areas of applications below.

## 6.1 Loan Management

Loan Management is a crucial component of finance. Integration of machine learning in the process has significantly improved the effectiveness and efficiency of loan management. Arutjothi & Senthamarai (2017) propose using machine learning algorithms to predict the status of loans in commercial banks. The study establishes that these algorithms improve accuracy in loan status prediction since they also aid in risk management. In a similar survey by Jagtiani & Lemieux (2019), the authors focus on using the machine in alternative lending platforms such as LendingClub. The study shows that integrating devices and alternative data sources can result in accurate risk management and loan performance.

Additionally, loan frauds are common in Fintech. According to (Huang et al., 2021), integrating machine learning helps solve the problem of loan fraud by using fraud detection algorithms and other fraudulent activities. The use of machine learning n loan management has significant benefits, which outweigh the risks; the authors have established its usefulness in detecting loans, frauds, loan performance, and accuracy in loan status prediction.

## 6.2 Fraud Detection

There are a significant number of machine learning applications to detect fraudulent transactions in Fintech. According to Stojanović et al. (2021), machine learning algorithms can detect fraudulent transactions in Fintech. The authors highlight that machine learning algorithms help detect fraud since they can detect the trails of financial transactions, analyze behavior patterns and detect anomalies. This would assist the financial industry in identifying and preventing fraudulent activities, decreasing the risks of losing a tremendous amount of finance. In another study by Noor et al. (2019), there is a demonstration of how machine learning is used in a Fintech cyber threat attribution framework that ensures effective identification and attribution of cyber threats to potential actors. Machine learning has several benefits in detecting fraudulent activities since it can automate seeing them and make it efficient and faster. It also helps in accurate detection as the algorithm can analyze vast amounts of data to recognize anomalies and patterns that are not apparent to human beings. This assists financial organizations in detecting fraud earlier, reducing the risks of loss and providing security to the customer's financial information (Stojanović et al. (2021). However, using machine learning in fraud detection also poses some risks where the primary risk is the potential false positives where legitimate transactions can be wrongly identified as fraudulent (Noor et al., 2019). This can make the financial industry reject genuine transactions, leading to customer inconvenience. Another risk is false negatives, where the fraudulent transaction may not be detected, causing the financial institution to suffer losses (Stojanović et al. (2021). The risks call for ensuring proper implementation and monitoring of the machine learning algorithms to ensure accuracy in fraud detection.

## 6.3 Patent Evaluation

Another area in which machine learning is being applied is Patent Evaluation. Using machine learning, Chen & Chang (2021) studied how Fintech Patents influence Taiwan's financial institutions. The authors designed a machine learning model to predict the potential of commercializing Fintech patents. The study's findings show that machine learning modes are more accurate than traditional methods of evaluating how Fintech patents are commercialized. The study also established the essence of machine learning in evaluating patents. The application of machine learning provides an accurate and efficient evaluation process, ensuring that financial organizations can come up with informed decisions concerning the type of patent they can invest in. This ensures the effective use of resources to improve outcomes and returns for the Fintech industries.

However, any technology poses some risks; there are potential risks associated with using machine learning algorithms to evaluate patents; for instance, the machine learning model's accuracy may be affected, leading to potential bias (Chen & Chang, 2021). Therefore, it is crucial to consider different factors before implementing machine learning algorithms in this area of the Fintech industry. Studies indicate that using machine learning in patent evaluation has promising results and potential risks that must be considered to provide accurate results.

## 6.4 Risks Management

In terms of using machine learning for Risks management in Fintech, various studies have been conducted that evaluate both benefits and challenges of machine learning in this application area. For instance, Noor et al. (2019) propose a machine-based Fintech threat detection framework that leverages high-level indicators to compromise the detection of potential threats to the Fintech systems. The authors in the study contend that the machine learning framework can accurately identify the threats to the methods to ensure accurate information on the management risks in the associated information systems in Fintech.

Stojanović et al. (2021) propose the use of machine learning to detect fraud in the systems associated with financial systems. That focuses on the trail of fraudulent activities. The study has found that the algorithms could effectively be applied in identifying the patterns of fraudulent activities in the systems hence preventing the risks of financial loss. A similar study by Xia et al. (2020) evaluates the potential risks in Fintech through machine learning to analyze the Questions and Answers (Q&A) information from a platform that deals with online loan investment. The authors have since established that machine algorithms could be applied in identifying the risks that Fintech systems encounter, including threats of loan default or fraudulent transactions, ensuring a proactive approach or procedure is implemented in mitigating risks (Xia et al., 2020). In general, the studies show how machine learning can be a valuable approach in the Fintech industries, primarily to manage and mitigate the risks of the Fintech systems, such as cyber fraud threats. However, it is crucial to realize that effective use of machine learning in this area also requires critical consideration of potential technical failures.

## 6.5 Alternative Data

In the Fintech industry, alternative data plays a vital role in enabling companies to assess the worthiness of their credits for the sake of potential borrowers. A study by Jagtiani & Lemieux (2018) explores how alternative data and machine learning can impact lending money to borrowers. The study found that using alternative data, such as utility payment and telecommunication data, helps lenders make informed decisions and reduce default risks. This results in a more comprehensive borrower assessment to reduce the lenders' operational costs (Jagtiani & Lemieux, 2018). Alternative data enables the creation of new lending products and improves underwriting accuracy. The authors emphasize ethical considerations in applying alternatives and data and the need to ensure robust privacy protection.

A study by Phat Tien et al. (2021) shows that different factors contribute to the growth of Fintech industries. The study identifies alternative data as a critical driver because it helps the customers better understand the personalized services in the Fintech services. The study establishes the need to ensure data security protection for customers' privacy to ensure no misuse of consumer data (Huang et al., 2021). Using alternative data and machine learning algorithms can also assist in revolutionizing the process of lending to customers. However, companies should be able to trust the use of data in an ethical and legal approach that complies with relevant data protection and regulation to prevent any moral or legal implications.

## 6.6 Conclusion

The literature review covers machine application in Fintech industries specific to loan management, risks Management, patent evaluation, fraud detection, and alternative data (Xia et al., 2020). The articles in review showcase the potential benefits of applying machine learning in Financial industries, for instance, to increase efficiency in fraud detection. However, the studies also address some potential risks that can be posed by machine learning in the industries, such as technical debts and the need to ensure proper data management that leads to the protection of consumer privacy and confidentiality (Xia et al., 2020). The results of the studies align with the general aim of the literature review to evaluate the benefits and risks of machine learning in Fintech. The findings of the literature review also suggest that machine learning has the potential to create significant improvement in different aspects of the Fintech industry.

It is essential to adequately address the potential risks related to the technology (Huang et al., 2021). Machine learning can potentially revolutionize the Fintech industry, but it is essential to implement it in the sectors with significant consideration of caution and its risks. Using alternative data has ensured a considerable expansion in the potential application of machine learning. However, deploying this solution comes with challenges and risks, including privacy concerns from biased algorithms. In general, the benefits of machine learning still outweigh the risks; hence it is vital to use it in Fintech. Further research is required to fully establish the impact of machine learning in Fintech to identify best practices during its integration or implementation.

# Lessons learned

Throughout this project, I have acquired valuable skills and learned to use various tools crucial to the project's success. I used these skills and tools throughout this project to understand and demonstrate what I have learned.

Data cleaning and manipulation using Pandas:

I learned how to clean and manipulate historical data for LendingClub using Pandas. To prepare the dataset for further analysis and model building, missing values were handled, data was filtered, and information was aggregated.

Mathematical operations and data analysis with NumPy:

By using NumPy, I gained experience performing mathematical operations and analyzing data. For efficient data processing, I calculated mean, median, and standard deviation and manipulated arrays.

Data visualization using Matplotlib and Seaborn:

I learned how to create various plots and visualizations using Matplotlib and Seaborn libraries. As a result, I could explore the data, identify patterns, and better understand the relationships between different features, which was crucial for selecting components and building models.

Model building using XGBoost:

I built a binary classification model using the XGBoost library to determine the criteria for accepting loan applications. To improve the model's accuracy and performance, I tuned hyperparameters to obtain optimal results using gradient-boosting techniques.

Creating interactive spreadsheets with Ipysheet:

With the Ipysheet library, I was able to design an interactive spreadsheet that could be used by users to input loan application data and receive a suggested determination based on the model's output.

Building user interfaces with Ipywidgets:

By creating interactive components within Jupyter Notebooks, I gained experience building user interfaces using Ipywidgets.

File handling and directory management using OS module:

Using the OS module to handle files and manage directories, I ensured efficient organization and access to the project's files.

Serialization and deserialization with Pickle:

Using the pickle module's serialization and deserialization functionality, my training classification model can now be stored and loaded for future analysis and evaluation.

This project has greatly enhanced my abilities to clean, analyze, and visualize data and build machine learning models and user-friendly interfaces for various applications. There is no doubt that these skills will be helpful in future projects and professional endeavors.

During the LendingClub Risk Assessment Model project, I acquired technical skills related to data analysis and model building and essential soft skills that contributed to the project's success. Any professional setting will benefit from these skills, and I will continue to utilize them.

Project Management:

I developed my project management skills by effectively planning, executing, and closing the project. As part of this process, the project scope was defined, realistic goals and deadlines were set, resources were allocated, and progress was monitored. The project was completed on time and within the defined parameters by keeping track of milestones and adapting to changing circumstances.

Communication:

Collaboration with my sponsor and professor, sharing progress updates, and presenting results improved my communication skills. This facilitated the smooth flow of information between all parties and ensured everyone was on the same page. Effective communication was crucial to address any issues that arose and keep the project on schedule.

Time management:

By prioritizing tasks, setting deadlines, and breaking complex tasks down into manageable subtasks, I learned to manage my time efficiently throughout the project. In this way, I ensured that each aspect of the project received the attention it deserved and maintained a steady pace of progress.

Organization:

Working on this project helped me develop strong organizational skills. Maintaining a clean and organized workspace included keeping track of project files, resources, and documentation. My organization allowed me to access important information quickly, make informed decisions, and streamline the project process.

Problem-Solving:

I sharpened my problem-solving skills by tackling various challenges that arose during the project. This included resolving technical issues with the user interface, refining the machine learning model, and addressing data discrepancies. I overcame these obstacles by applying critical thinking and leveraging my technical knowledge.

Critical Thinking:

Throughout the project, I developed my critical thinking skills by carefully evaluating the data, making informed decisions, and continuously refining the approach. This process was part of analyzing assumptions, identifying biases, and drawing well-reasoned conclusions. I made better decisions and achieved higher accuracy in the final model by applying critical thinking.

In conclusion, I obtained valuable technical knowledge while developing essential soft skills to succeed in any professional setting due to my work on the LendingClub Risk Assessment Model project. I am better prepared to tackle future projects and contribute to success by combining these newfound skills with my technical knowledge.

As a result of my involvement in the LendingClub Risk Assessment Model project, I gained valuable insights into the financial industry, risk assessment, and machine learning applications. In addition to enriching my knowledge, these insights gave me a broader perspective on the project.

Importance of risk assessment in the financial industry:

Through my work at LendingClub, I have gained a deeper understanding of how risk assessment plays a role in the financial sector. As borrowers can obtain loans at more appropriate interest rates when investors accurately assess the risks associated with different loans, investors can make more informed decisions and minimize potential losses.

The power of data-driven decision-making:

In this project, the importance of data-driven insights was reinforced. I created a model based on historical transaction data and machine learning techniques to objectively evaluate the risk levels of loan applications, resulting in better decisions for investors and borrowers.

Feature selection and preprocessing impact model performance:

I discovered that the choice of features and preprocessing steps could significantly influence the performance of machine learning models. To improve the accuracy and robustness of the risk assessment model, I carefully selected relevant features and appropriately preprocessed the data.

Machine learning potential in the financial industry:

In this project, machine learning applications were demonstrated within the financial sector. A wide range of financial services can benefit from machine learning techniques, from risk assessment and fraud detection to portfolio optimization and algorithmic trading.

The value of user-friendly interfaces for data-driven tools:

As I developed the Excel sheet for the risk assessment model, I realized the importance of creating user-friendly interfaces. Providing an intuitive and accessible platform for users to input data and receive results will make the insights generated by the model easily understood and actionable for experts and non-experts.

The importance of soft skills in technical projects:

Finally, this project demonstrated how vital soft skills such as project management, communication, and critical thinking are essential for technological success. I could manage resources effectively, collaborate with sponsors, and apply critical thinking to solve problems which enabled me to overcome challenges and complete the project on time and on budget.

In conclusion, the LendingClub Risk Assessment Model project provided me with numerous insights that have enhanced my understanding of the financial industry and machine learning applications and emphasized the importance of data-driven decision-making and soft skills in technical projects.

# Project Chronology and Critique

## 8.1 Project Chronology

Week 1:

Day 1-2:

To begin the project, I researched LendingClub's business model, history, and the peer-to-peer lending industry. To achieve the project's objectives, it was essential to understand the company's operations and the loan issuance mechanism.

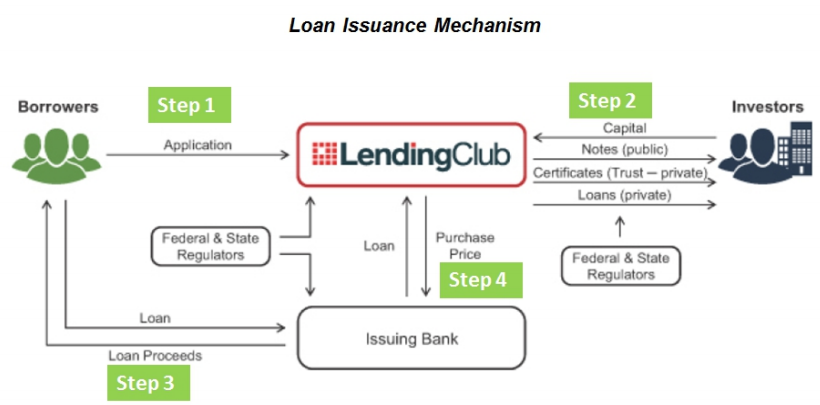
Study loan issuance mechanisms: I explored loan grading, interest rates, and repayment structures as part of my investigation of loan insurance mechanisms. Understanding how loan risk assessment is influenced by these factors is crucial. 

Fig. 8-1

Day 2-3:

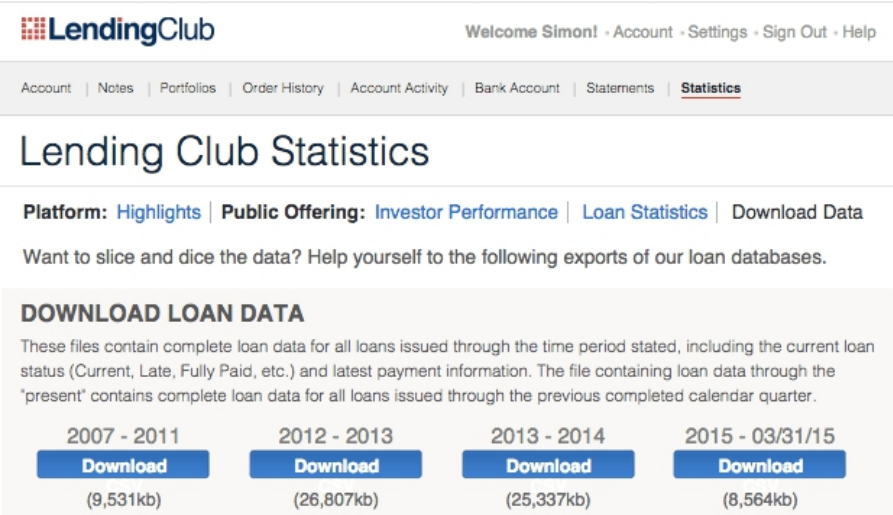
Collect historical transaction data: I sourced historical transaction data from LendingClub to build the risk assessment model. There are comprehensive details about loan applications, borrowers, and loan performance contained in this dataset.

Fig.8-2

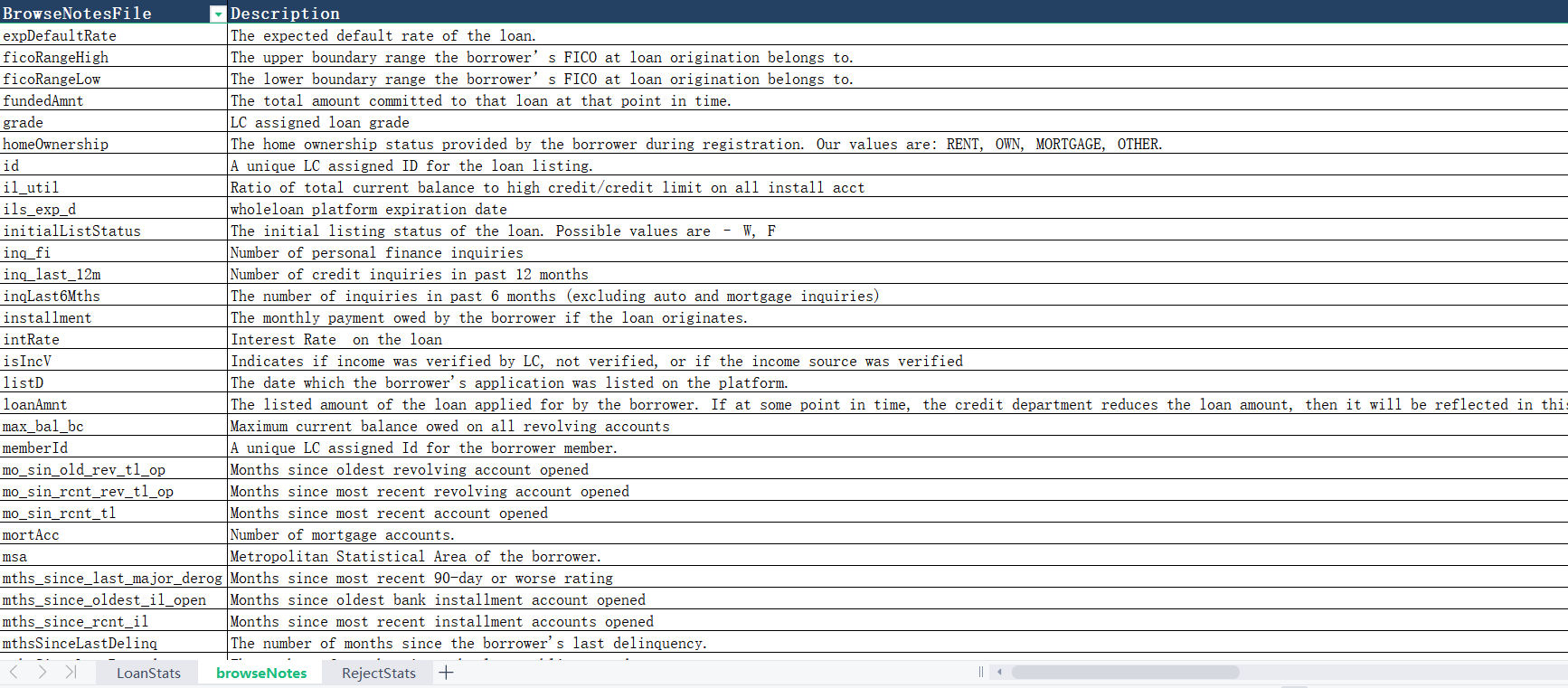
I familiarized myself with LendingClub's data dictionary for practical analysis and utilization of the dataset. This resource detailed the definitions and descriptions of various data fields, ensuring the data could be interpreted and used correctly. 

Fig. 8-3

During the initial stage of the LendingClub Risk Assessment Model Project, I arranged a meeting with the project sponsor to better understand LendingClub's business background and the project objectives and to discuss the process of collecting historical loan data.

In this meeting, we delved into the details of LendingClub's peer-to-peer lending platform, its loan issuance mechanism, and the grading system used for different loan options. We also discussed the project's main objectives, which included developing a risk assessment model for investors to determine the potential risk of specific loan applications.

The project sponsor provided valuable insights and guidance regarding the data collection process, including the best sources for obtaining historical loan data and the data dictionary, which would play a crucial role in the project's data exploration, validation, and cleaning stages.

This initial meeting with the project sponsor was instrumental in establishing a solid foundation for the project, ensuring a clear understanding of LendingClub's business background, project objectives, and data collection requirements. The insights and guidance provided by the sponsor helped shape the project plan and set the stage for a successful project outcome.

Week 2:

Day 1-3:

Analyze the quality of LendingClub's historical transaction data: I dedicated time to analyzing the data quality. During this process, I checked for missing values, duplicate entries, and inconsistencies in the data. I ensured the dataset was reliable and suitable for further analysis by identifying and addressing data quality issues.

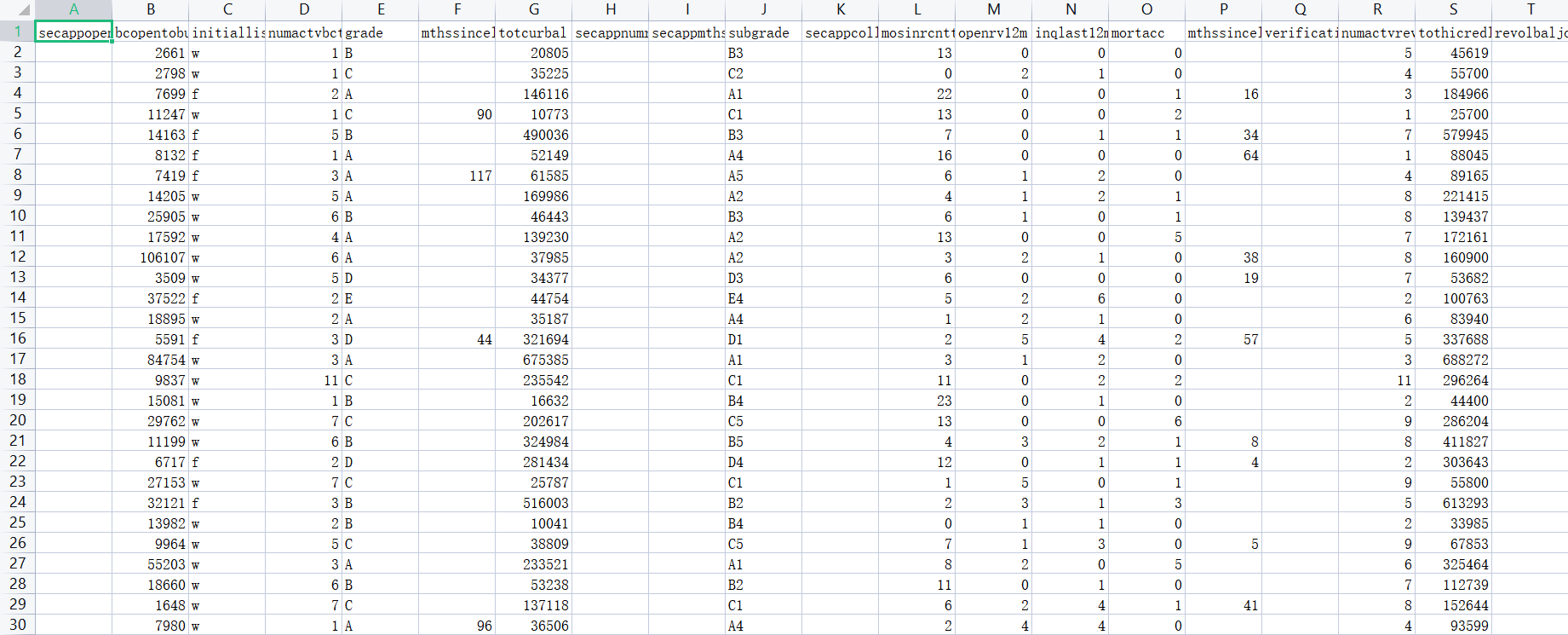
I evaluated the dataset's size and comprehensiveness to build a robust risk assessment model. The dataset provided 443,580 data points, allowing the development of an accurate and reliable model based on a substantial sample.

Fig. 8-4

Day 4-5:

(Data Cleaning) Feature name format conversion: I discovered that features with the same meaning had different names in the dataset. I converted feature names to the standard format to address this issue, ensuring consistency and eliminating potential confusion during analysis and modeling.

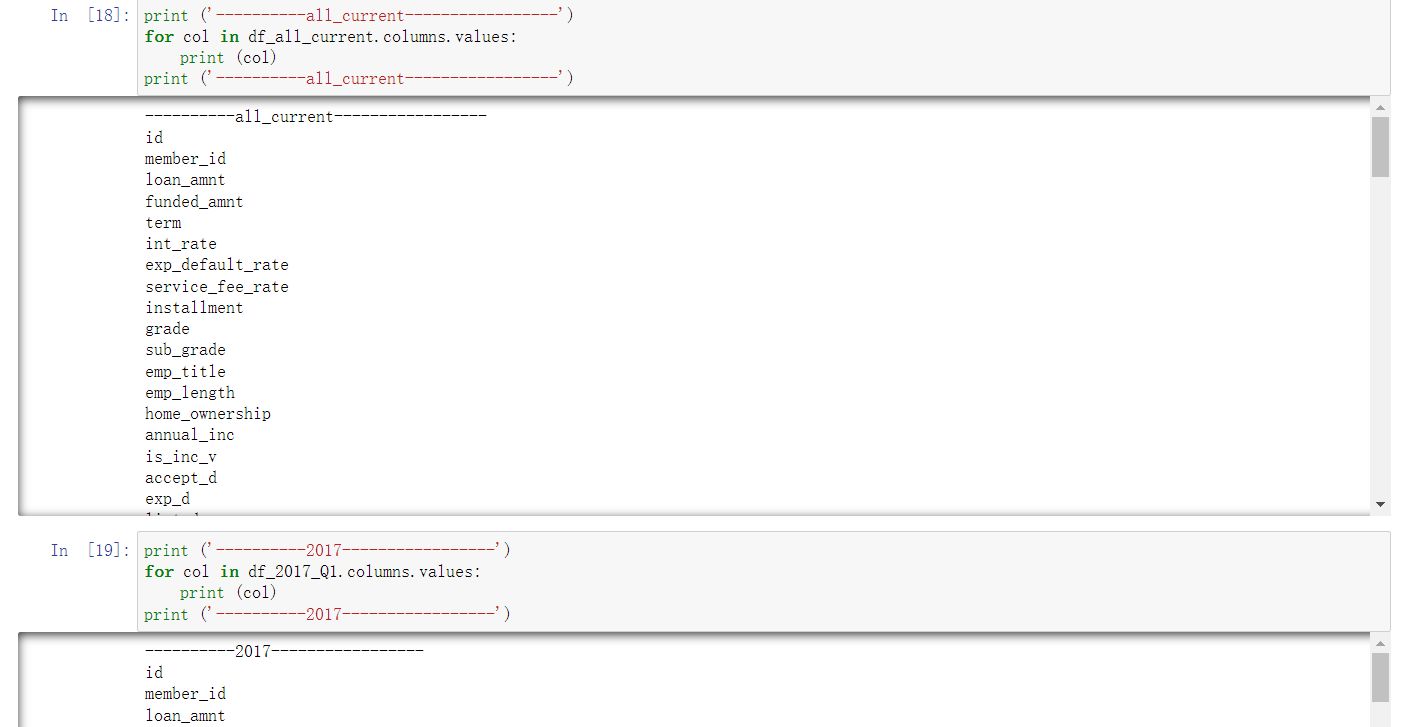


Fig. 8-5



Fig. 8-6

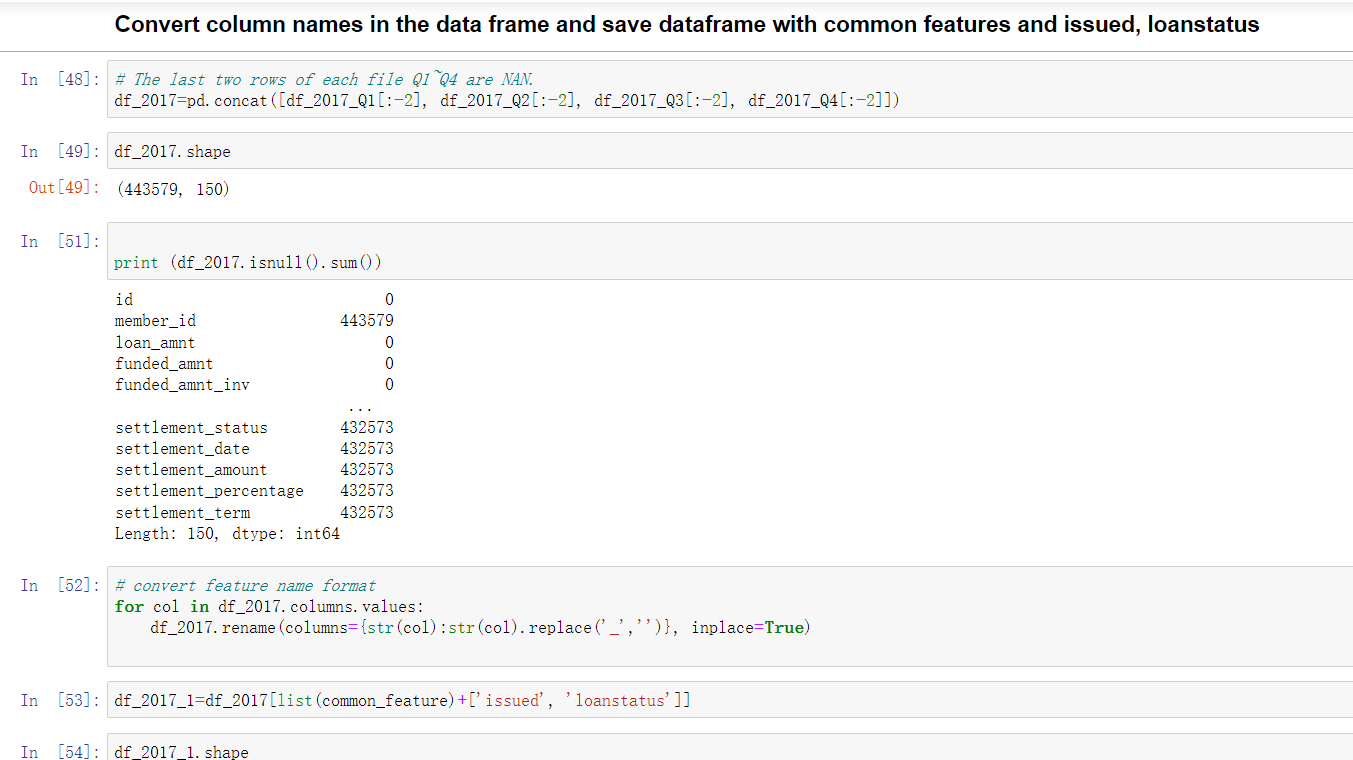
(Data Cleaning) Changing column names and saving the processed data frame: After fixing feature name format inconsistencies, I unified the column names in the data frame. To conduct further analysis and model building, I kept the data frame with standard features, issued loans, and loan status information.

Fig. 8-7

Week 3:

Day 1-3:

Ensure the cleaned and processed dataset is ready for in-depth exploratory data analysis (EDA): I organized and processed the cleaned and processed dataset. The dataset's structure was reviewed to prepare for EDA and key features were identified.

Fig. 8-8

Conducting the EDA - Part 1: During the EDA, I examined employment length, verification status, home ownership, and their relationship to loan status (fully paid off or charged off). I aimed to understand the influence of these factors on borrowers' willingness to repay their loans through visualizations and statistical analysis.

Fig.8-9

Day 4-6:

Conduct EDA - Part 2: I conducted a second EDA to assess the impact of other variables, such as open account numbers, loan grades, and other variables, on loan status. I found no strong correlation between employment length and loan repayment, but the likelihood of loan charge-offs increased as the number of open accounts increased.

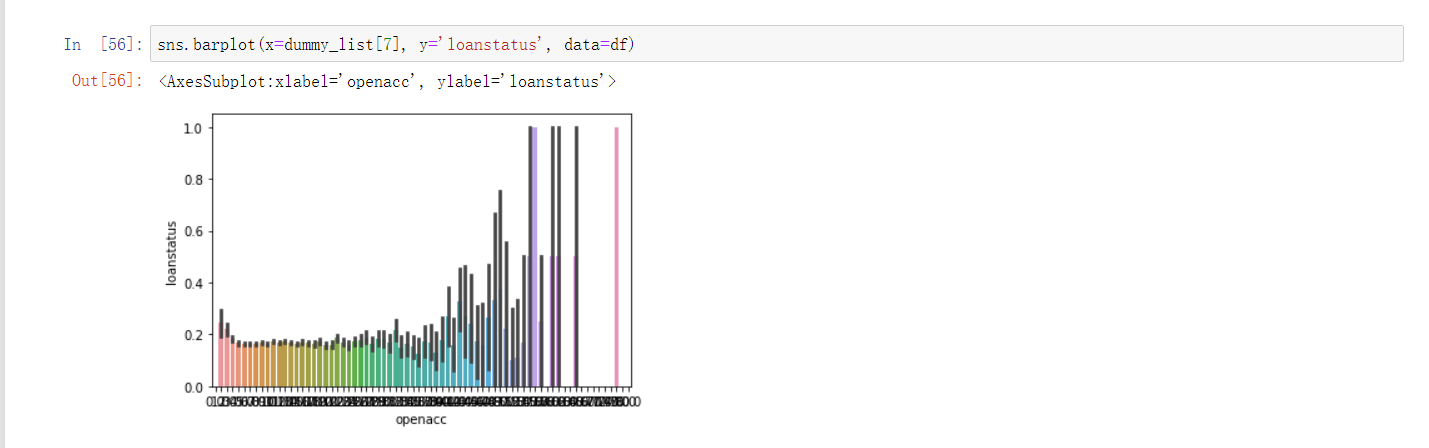


Fig. 8-10

Identify key patterns and relationships between examined features and loan status based on EDA results: I synthesized the findings from the EDA. To build an accurate and reliable risk assessment model, these insights informed feature selection and model development.

Midway through the LendingClub Risk Assessment Model Project, I arranged a meeting with the project sponsor to discuss the exploratory data analysis (EDA) findings and ensure that I was on the right track. Sharing the results of the EDA allowed me to obtain valuable feedback and confirm that my approach aligned with the project objectives.

During the meeting, I presented the graphs and insights generated from the EDA, including relationships between factors such as employment length, verification status, homeownership, open account number, and grade, and their correlation with loan status. The sponsor provided feedback on the findings and offered additional perspectives on the relationships between various factors and loan performance.

Week 4:

Day 1-3:

Review the dataset and identify areas that require further cleaning and feature engineering: I reviewed the dataset and identified areas that require additional cleaning and feature engineering. I checked for missing values, null entries, and outliers and placed categorical columns that needed to convert into dummy variables.

The first step in the data-cleaning process was to address missing values, null entries, and outliers. I carefully examined the data to build a reliable model, removing any identified outliers. I used .isnull() and .value\_count() to determine each feature's missing value. I also found out that the column [fundedamnt] and [loanamnt] have the same values but different column names, so I decided to drop one of them, which is [fundedamnt].

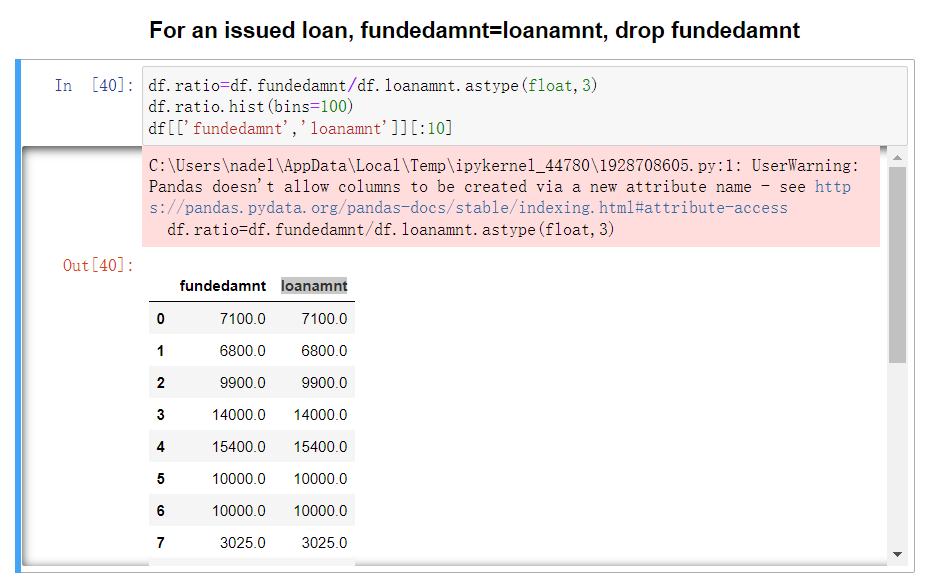


Fig. 8-11

Day 4-7:

I continued the data-cleaning process by transforming categorical columns into dummy variables. For more effective analysis and model building, absolute values convert into binary format. For example, I transformed categorical variables into dummy variables to facilitate modeling. For instance, I flipped the 'Loanstatus' column by mapping "Charged Off" to 1 and "Fully Paid" to 0 using the following code: df['Loanstatus'] = df.LoanStatus.map({"Charged Off": 1, "Fully Paid": 0}). This conversion allowed for a more straightforward interpretation of the loan status in the dataset.

Lastly, I processed the employment length data by dividing it into years, facilitating a more consistent and meaningful representation of this feature. These transformations of categorical variables into more manageable formats played a crucial role in preparing the dataset for effective modeling and analysis.



Fig. 8-12

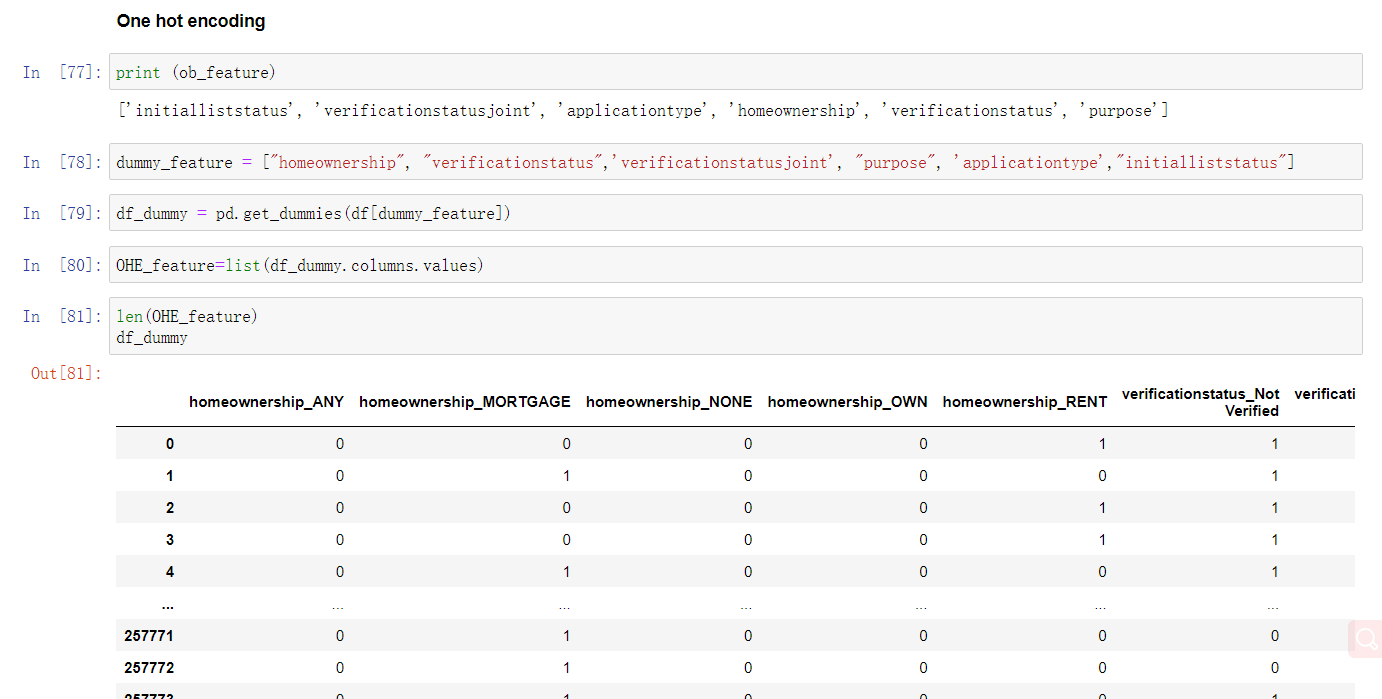


Fig. 8-13

Feature engineering: I incorporated new variables into the dataset and transformed existing ones to enhance their predictive power. As a result of feature engineering, the model would perform better after.

For example, the dataset included the column 'Earliestcrline,' which contained the month and year of the earliest credit time. I split this column into two new columns representing the month and year separately. I then converted the month names, such as January and February, into their corresponding numerical representations (e.g., January as 11, February as 10, and so on).

Next, I utilized the following function to calculate the total number of months elapsed from the earliest credit time to 2017: df['earliestcrline\_month'] = df.Earliestcrline.apply(lambda x: (2017-int(x.split('-')[1]))\*12 + dic\_month[x.split('-')[0]]). This transformation streamlined the handling of date-related data during analysis and modeling, ensuring a more effective interpretation of the temporal information in the dataset.

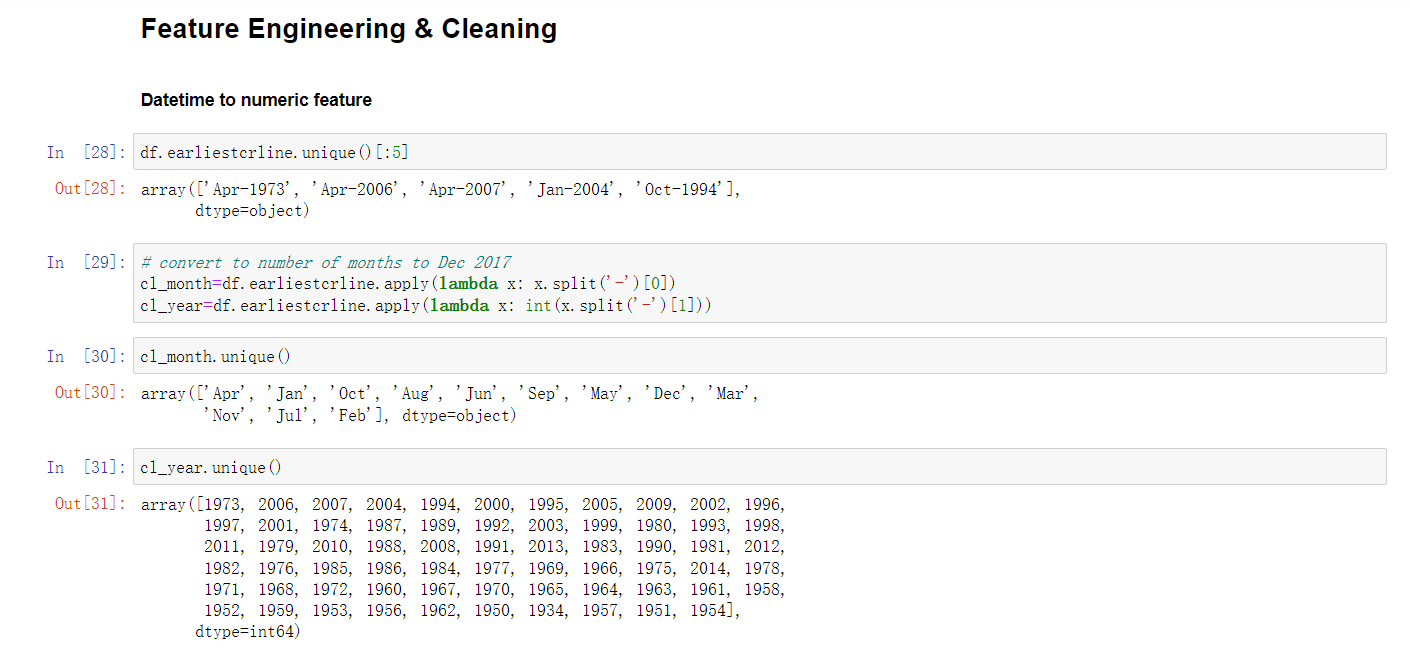


Fig. 8-14

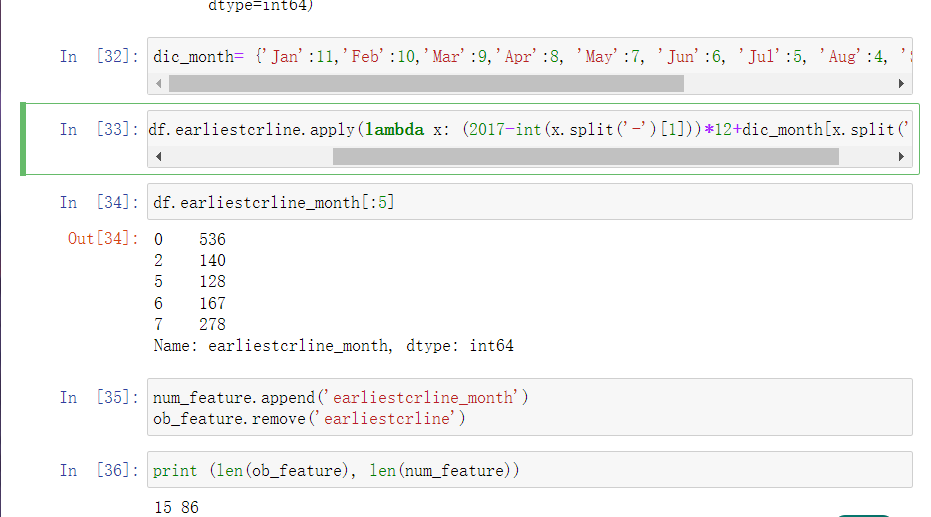


Fig. 8-15

Week 5-6:

Day 1-5:

I completed data cleaning, ensuring that the dataset was accurate, reliable, and well-structured for subsequent model building. I was now ready to move onto the data modeling stage of the project with a clean and properly formatted dataset.

I split the dataset into training and testing sets using the train\_test\_split function to prepare for data modeling. In this way, I trained the model on one subset of the data and evaluated its performance on another unseen subset.

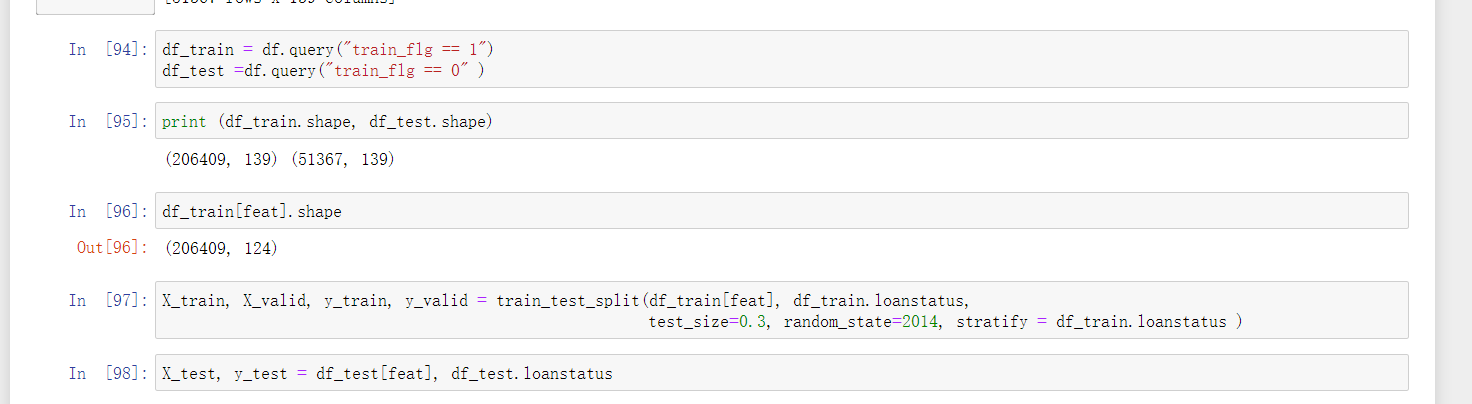


Fig. 8-16

Day 6-13:

I trained the risk assessment model on the training dataset using the XGBoost algorithm. The XGBoost gradient boosting algorithm is well known for its accuracy and speed, making it an excellent choice for this project.

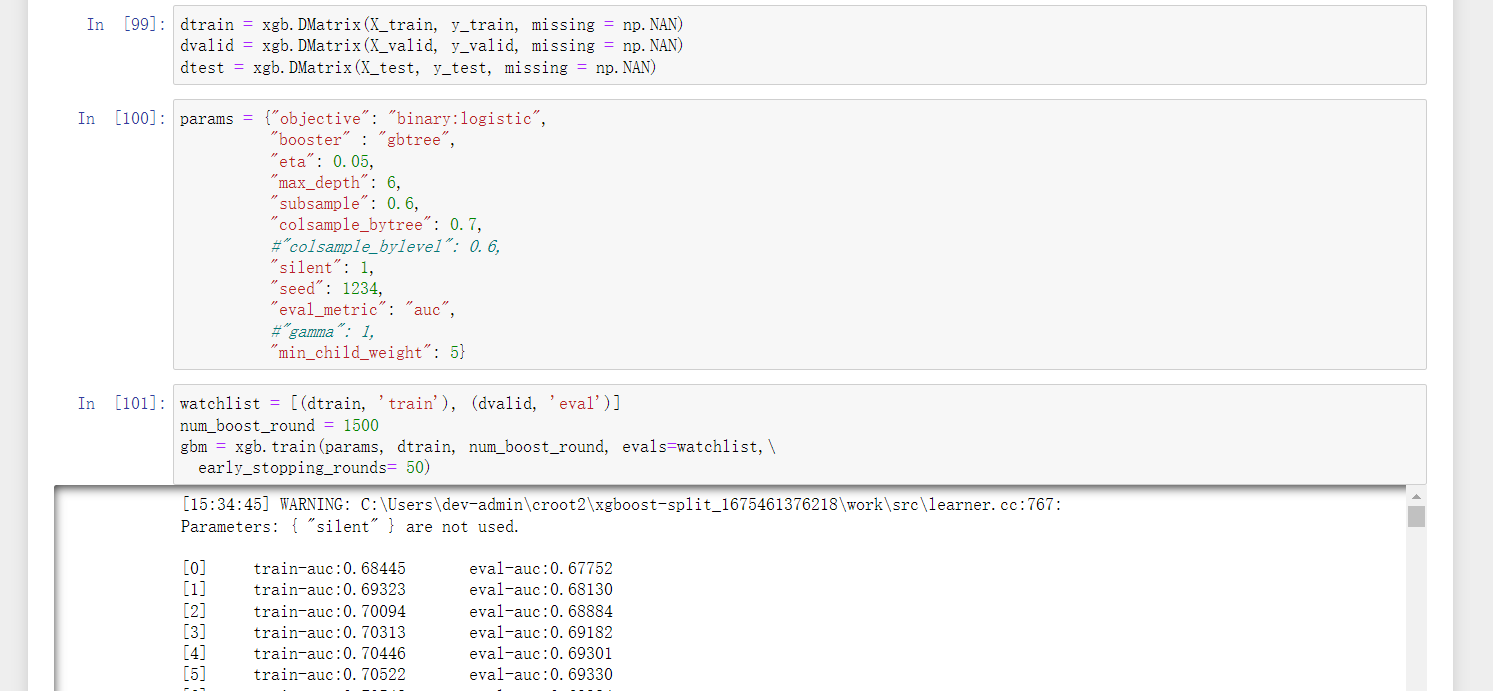


Fig. 8-17

I calculated the ROC-AUC scores for the training, validation, and testing datasets to evaluate the model's performance. For the training set, the ROC-AUC was 0.822893. For the validation set, 0.713800, and for the testing set, 0.716648. The scores were reasonably close, indicating that the model was generalizing well and performing well.

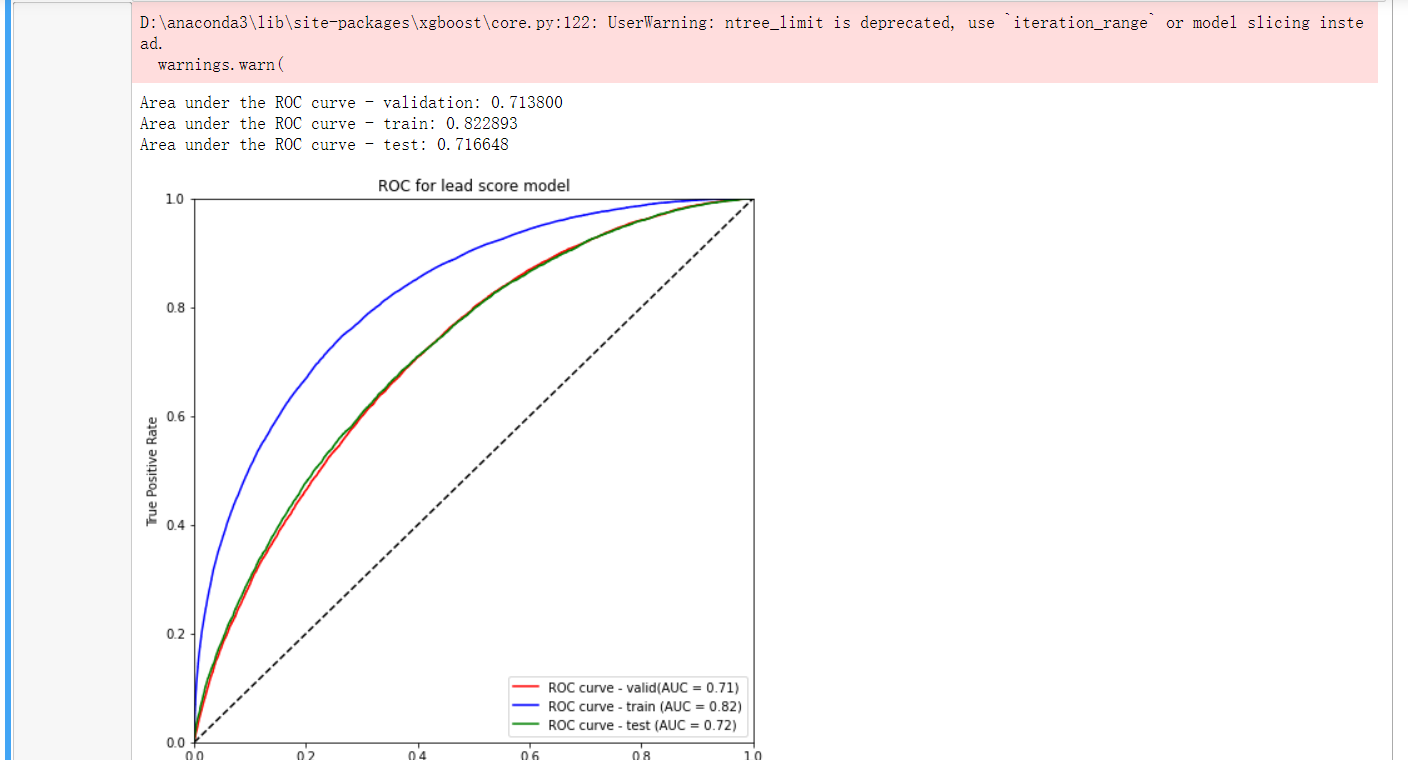


Fig. 8-18

After completing the data preparation and modeling phases, I met with the sponsor again to present the performance of the XGBoost model and discuss the development of the Excel interfaces using Ipysheet and Pyxll. We discussed the pros and cons of each interface and decided to keep both as viable options.

Week 7-8:

Day 1-5:

Create an Excel-based interface for users to submit loan applications and receive risk assessment results: I researched different tools and libraries that could use to create an Excel-based interface for users. Both Ipysheet and Pyxll chose after evaluating the options.

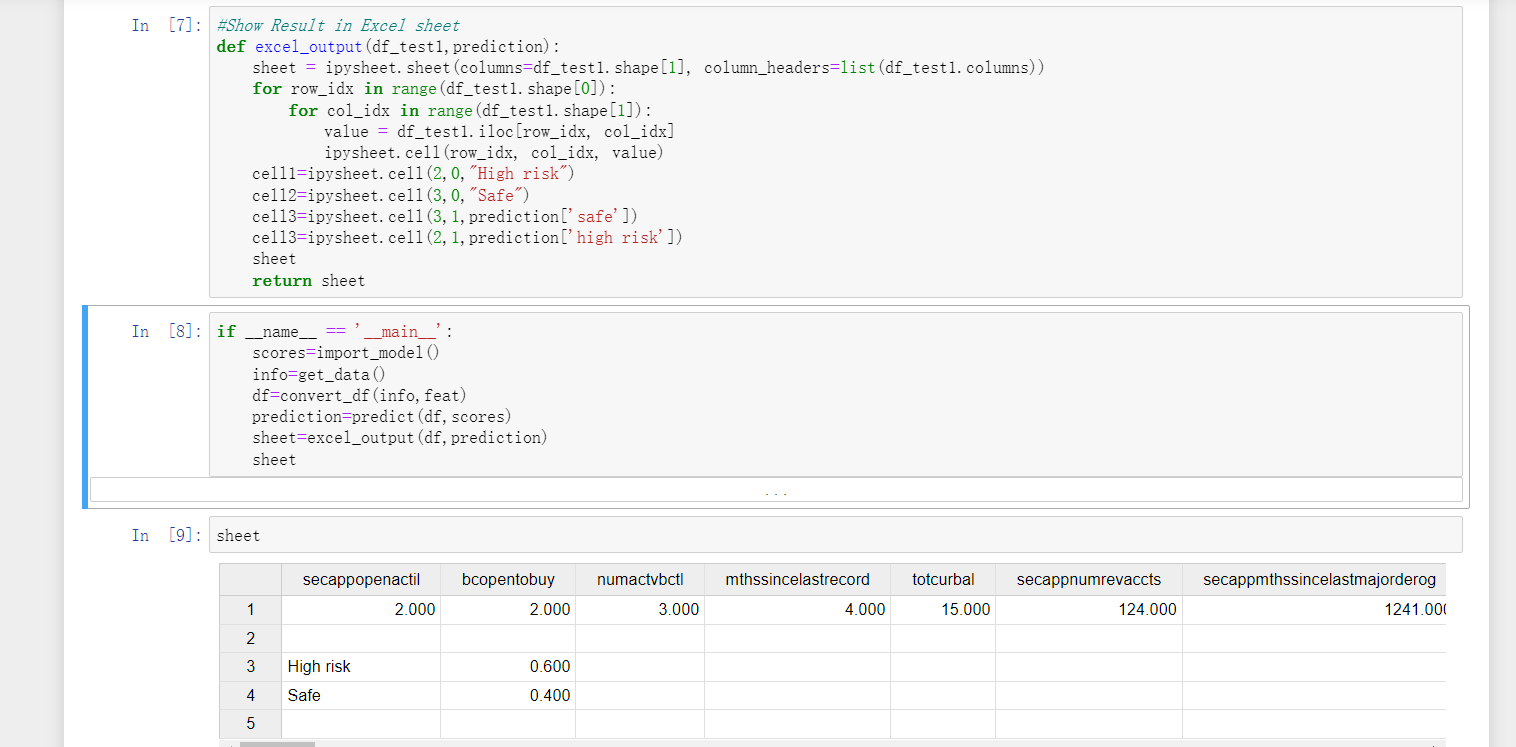
I developed the Excel interface using Ipysheet, a Jupyter Notebook widget for working with spreadsheets. Despite not integrating directly with Excel, I chose Ipysheet as a backup plan due to its ease of use and compatibility with Jupyter Notebooks.

Fig. 8-19

Day 6-13:

My next step was to develop an Excel interface based on Pyxll, a powerful Python library that integrates seamlessly with Excel. Using this solution, users could input loan application data and receive risk assessment results within Excel in a more direct and user-friendly manner.

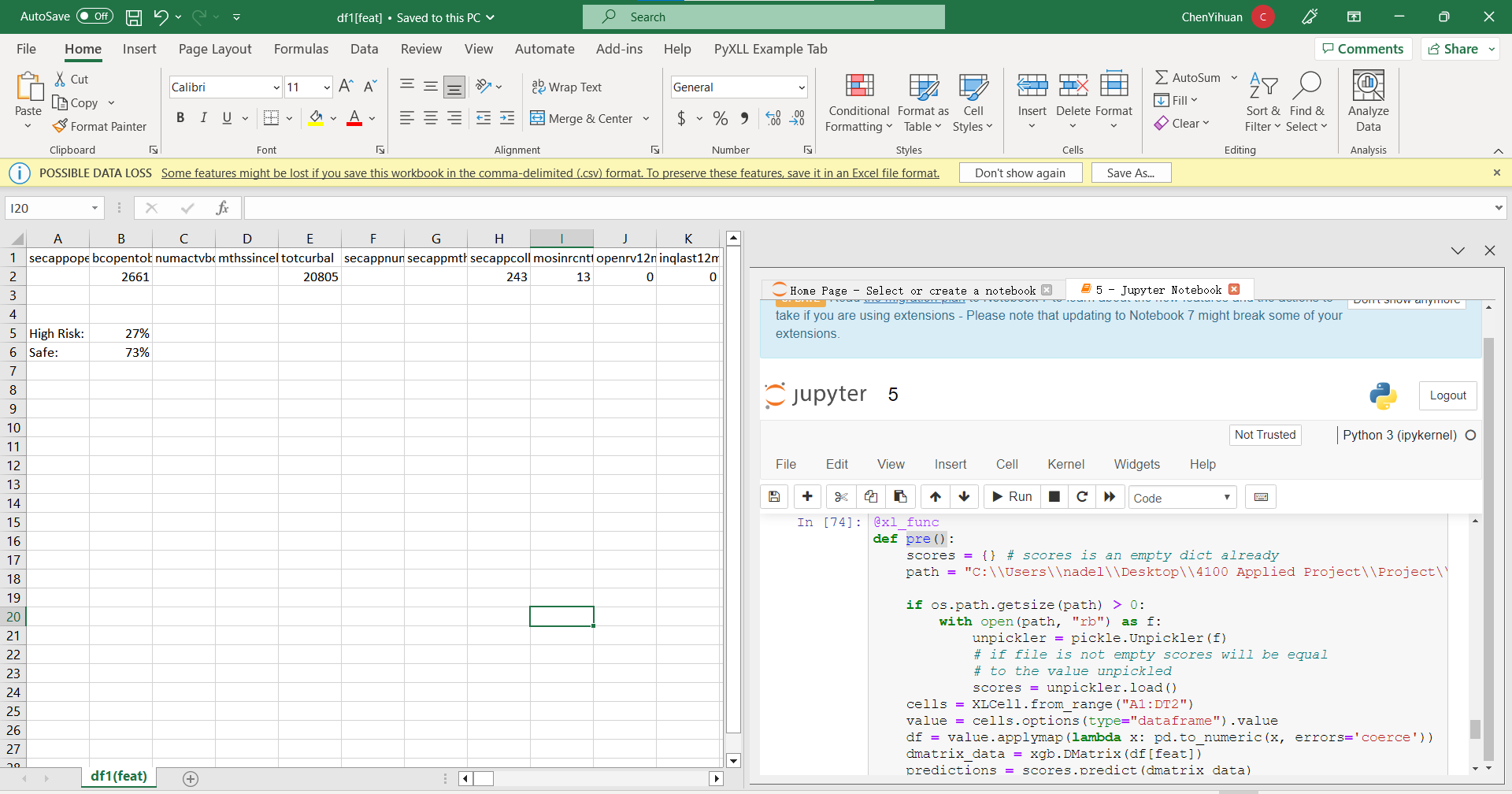


Fig. 8-20

I compared the functionality and user experience of the Ipysheet- and Pyxll-based Excel interfaces. Despite Pyxll's better integration with Excel, I kept the Ipysheet version as a backup solution since Pyxll has a 30-day free trial period.

Toward the end of the project, I organized a final meeting with the sponsor to demonstrate the completed Excel interfaces and provide a comprehensive project summary, including key findings, model performance, and areas for improvement. The sponsor expressed satisfaction with the project outcomes and offered valuable feedback for future iterations or enhancements.

## 8.2 Critique and Improvement:

**Week 1:**

The initial research phase provided a strong foundation for the project by providing a thorough understanding of LendingClub and loan issuance mechanisms. To develop an accurate risk assessment model, that was essential to understand the company's background, the peer-to-peer lending industry, and the loan issuance mechanisms.

Data collection and data dictionary learning were efficient: Historical transaction data were collected efficiently, and the data dictionary was learned effectively. Analyzing the data and extracting meaningful insights was more manageable after I became familiar with the dataset's structure and variables.

Areas for improvement:

In addition to the initial research phase, additional research on risk assessment techniques would have provided additional insights and guided the development of the risk assessment model, even though the initial research phase was comprehensive.

Early identification of potential challenges: It would have been beneficial to conduct an early assessment of potential challenges, such as data quality issues or missing values, during the initial stages of the project. In this way, proactive planning and mitigation strategies would have been able to be implemented, resulting in a smoother project process.

In general, the project's first week was productive, with a strong focus on research, data collection, and data dictionary understanding. In addition to data analysis, model building, and interface development, these efforts laid a solid foundation for the project's subsequent stages.

Week 2:

Critique:

Data validation and sufficiency assessment: A thorough validation process was implemented to ensure the quality and sufficiency of the data. I thoroughly examined the dataset and addressed issues to ensure the data was reliable and suitable for the project's objectives.

Handling feature name format inconsistencies effectively: Feature names were formatted, and the column names were converted efficiently, resulting in a consistent and easily understandable dataset.

Areas for improvement:

It successfully preprocessed the data and converted the feature name format. However, if additional time had been allocated to this stage, potential data quality issues could have been examined in greater detail, and a more thorough cleaning process could have been performed.

A focus could have been placed on identifying key features and performing feature engineering during the data processing phase to improve the dataset's predictive power. As a result, the model-building process could have been streamlined, and performance could have been improved.

This week's work focused on ensuring data quality and sufficiency and cleaning and processing the dataset. As a result of addressing feature name format inconsistencies and converting column names, a clean and well-structured dataset was created, which paved the way for in-depth data analysis, feature selection, and risk assessment.

Week 3:

Critique:

I conducted a thorough and well-executed EDA, examining various features and their relationship to loan status. As a result of exploring multiple factors such as employment length, verification status, homeownership, open account numbers, and loan grades, I gained a deeper understanding of how loan repayment is influenced.

In the EDA process, graphs and visualizations helped illustrate the relationships between features and loan status. As a result, patterns and trends could be identified more efficiently, providing valuable insights during the design phase.

Areas for improvement:

Allocating more time for the EDA could have allowed a deeper exploration of additional features and their interactions, even though the EDA was comprehensive. More nuanced relationships could have been discovered that could have improved the performance of the risk assessment model.

During the third week of the project, I conducted a comprehensive EDA, exploring various features and how they relate to loan status. As a result of this process, the valuable context was provided for the feature selection and model development stages, resulting in an accurate and reliable risk assessment model.

Week 4.

Critique:

Comprehensive data cleaning: The data-cleaning process addressed missing values, null entries, and outliers. To build a reliable model, I eliminated these possible sources of error.

It was possible to improve the dataset's predictive power using feature engineering by turning categorical columns into dummy variables and creating new variables. A better model could develop due to enhancing the dataset's structure.

Areas for improvement:

A significant delay in the project schedule was caused by time management during data cleaning and feature engineering. This stage required careful handling of missing values, null entries, and outliers. I also needed to convert categorical variables into dummy variables differently, which proved more time-consuming than anticipated. Managing time during this stage would have been more efficient to avoid delays and ensure a smoother workflow.

Data cleaning issues could have been identified early: During the initial stages of the project, a preliminary assessment of potential challenges should have been conducted. By doing this, proactive planning and mitigation strategies could have been implemented, ensuring a smoother workflow for the project.

Data cleaning and feature engineering dominated the fourth week of the project. In addition to ensuring that the dataset was accurate, reliable, and well-structured, we also employed practical feature engineering. The efforts made during this stage laid a solid foundation for developing a reliable and precise risk assessment model, despite the delay in the project schedule.

Week 5-6:

Critique:

Modeling data preparation: The data cleaning and preparation for modeling are well executed. A reliable risk assessment model could be developed using the clean and well-structured dataset.

Model training and evaluation: The training and testing datasets produced satisfactory ROC-AUC scores using the robust XGBoost algorithm. The model could evaluate unseen data accurately and generalize well based on its performance.

Areas for Improvement:

A more comprehensive approach would have involved comparing the performance of multiple algorithms in addition to the XGBoost algorithm, which is used to train the risk assessment model. A better-performing model for this problem could have been identified by exploring alternative algorithms, such as random forests, support vector machines, or neural networks. This aspect of the project was not analyzed due to time constraints.

Week 7-8.

Critique:

Choosing Ipysheet and Pyxll as tools to build the Excel input and result interface demonstrates a thoughtful approach to finding suitable solutions for the project's requirements.

Two Excel interfaces successfully developed: Developing both Ipysheet-based and Pyxll-based Excel interfaces provided flexibility and redundancy, ensuring that a functional solution was available for users to input loan application data and receive risk assessment results.

Areas for improvement:

A potential challenge with Pyxll is the 30-day free trial limitation, despite Pyxll providing better integration with Excel. The possibility of exploring alternative tools or negotiating a long-term license for Pyxll could be explored in the future.

A streamlined development process for the Excel interface could have streamlined the process of developing two Excel interfaces, providing redundancy while saving time and effort. The tool should offer seamless Excel integration and long-term, cost-effective licensing.

Throughout weeks 7 and 8, I researched, selected, and developed two Excel interfaces for users to input loan application data and receive risk assessment results. Ipysheet-based and Pyxll-based interfaces were developed successfully, ensuring a functional and user-friendly solution while providing redundancy in case one option wasn't viable in the long run. The future development process for Excel interfaces could streamline, and licensing considerations for tools such as Pyxll could be addressed in the medium and long term.

Regular meetings with the project sponsor and professor throughout the LendingClub Risk Assessment Model Project were instrumental in ensuring alignment with expectations, addressing concerns, and incorporating valuable feedback to deliver a comprehensive risk assessment solution for LendingClub investors.

# Sponsor Acceptance Form

Text

Description automatically generated

# References

Arutjothi, G., & Senthamarai, C. (2017). Prediction of loan status in the commercial bank using machine learning classifier. *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. <https://doi.org/10.1109/iss1.2017.8389442>

Chen, T.-H., & Chang, R.-C. (2021). Using machine learning to evaluate the influence of FinTech patents: The case of Taiwan's financial industry. *Journal of Computational and Applied Mathematics*, *390*, 113215. <https://doi.org/10.1016/j.cam.2020.113215>

‌Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management*, *48*(4), 1009–1029. <https://doi.org/10.1111/fima.12295>

Huang, C., Nourian, A., & Griest, K. (2021). Hidden Technical Debts for Fair Machine Learning in Financial Services. *ArXiv.org*. <https://doi.org/10.48550/arXiv.2103.10510>

Noor, U., Anwar, Z., Amjad, T., & Choo, K.-K. R. (2019). A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise. *Future Generation Computer Systems*, *96*, 227–242. <https://doi.org/10.1016/j.future.2019.02.013>

Pham Phat Tien, Quddus, A., Asad, A., & Sarfraz, H. (2021). *THE FACTORS OF FINTECH: A LITERATURE REVIEW*. ResearchGate; unknown. <http://dx.doi.org/10.7441/dokbat.2021.34>

Stojanović, B., Božić, J., Hofer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., Sundaram, M., Jordan, E., & Runevic, J. (2021). Follow the Trail: Machine Learning for Fraud Detection in Fintech Applications. *Sensors*, *21*(5), 1594. <https://doi.org/10.3390/s21051594>

Xia, H., Liu, J., & Zhang, Z. J. (2020). Identifying Fintech risk through machine learning: analyzing the Q&A text of an online loan investment platform. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-020-03842-y>