华南理工大学

South China University of Technology

本科毕业设计(论文)开题报告

Graduation Project (Thesis) Proposal

论文题目 **Thesis Title**	Bankruptcy prediction using machine learning techniques

学　院:	**School of Computer Science & Engineering**
专　业:	**Computer Science and Technology**
学生姓名 :	**Student's Passport Name** **TREJO VENEGAS SEBASTIAN ANDRES**
学生学号 :	**Student ID** **201969990328**
指导教师:	**Supervisor** **Dong Min** 董敏

1、 课题背景及意义(含国内外研究现状综述)

1. Subject background and significance (Contains reviews of current research status at home and abroad)

Subject background:

Bankruptcy is a legal process that allows individuals or organizations to discharge their debts and start fresh. In the case of banks, bankruptcy can occur when a bank is unable to meet its financial obligations and is no longer able to operate effectively.

There are several reasons why a bank may go bankrupt. One common reason is poor management, where the bank makes risky investments or loans that result in significant losses. Other factors that can contribute to bank bankruptcies include economic downturns, changes in regulations, and unforeseen events such as natural disasters.

When a bank declares bankruptcy, its assets are sold off to repay its creditors, and its customers may face significant losses. Depositors may lose their savings, and borrowers may have their loans called in or be forced to find new lenders. The government may also step in to bail out the bank, using taxpayer money to stabilize the financial system.

Bank bankruptcies can have significant ripple effects throughout the economy. They can erode public trust in the banking system, lead to decreased investment and economic activity, and create instability in the financial markets. As such, preventing bank bankruptcies is a priority for regulators and policymakers, who work to maintain stable and healthy financial systems through regulation and oversight.

Significance:

The banking industry generates vast amounts of data on a daily basis, including customer transactions, account balances, and credit scores. This data can be analyzed and leveraged to gain insights into customer behavior, risk management, and overall business performance. Data science techniques can be used to analyze this data, uncover hidden patterns and relationships, and develop predictive models to inform business decisions.

2、 课题研究主要内容及研究基础

2. **The main content and basis of the research**

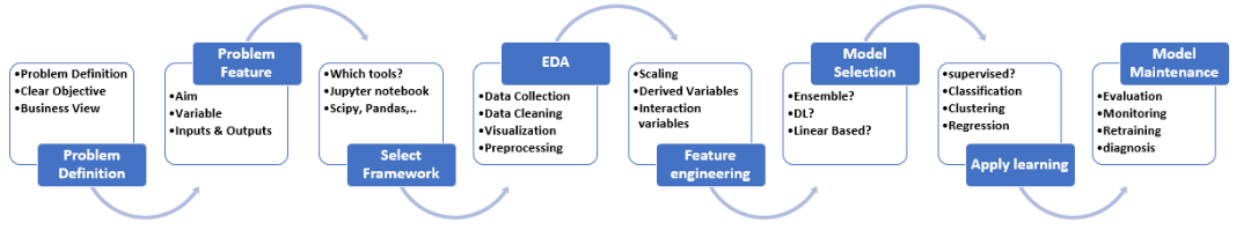
Main content

Data science insights can provide significant value to the banking industry. For example, data analysis can be used to identify customers who are at risk of defaulting on their loans, allowing banks to take proactive measures to mitigate this risk. Data science can also be used to personalize customer experiences and improve customer satisfaction, by analyzing customer behavior and preferences to make targeted product and service recommendations. Additionally, data science techniques can be used to optimize business operations, such as identifying areas for cost savings or improving efficiency. Overall, data science has the potential to drive significant improvements in customer satisfaction, risk management, and business performance in the banking industry.

3、 研究(或调研)方案和思路(技术路线)

3. **Research method and research route**

for general framework I will use this workflow guide to follow this project needs



Steps:

1. Data Cleaning and Preprocessing: Jupyter notebooks provide an interactive environment for cleaning and preprocessing data. Python libraries like Pandas and NumPy are used to manipulate and transform data, making it easier to analyze and extract insights. In this case using Visual Studio Code to develop this project, testing our data extracted from Kaggle website; database from Santander Bank.

2. Data Visualization: create interactive data visualizations that help to identify patterns and trends in the data. Python libraries like Matplotlib and Seaborn will be used to create a wide range of visualizations, from scatter plots and histograms to heatmaps and network graphs depending on the answers we are looking for.

To graph predictions for bank bankruptcy, I will need to create visualizations that effectively communicate the model's predictions and the underlying data. For example:

Receiver Operating Characteristic (ROC) Curve: ROC curve is a common visualization tool for binary classification models. It shows the relationship between the true positive rate and the false positive rate at different threshold values. The area under the ROC curve (AUC) can also be calculated, which is a measure of the model's overall performance.

Confusion Matrix: A confusion matrix is a table that summarizes the performance of a binary classification model. It shows the number of true positives, true negatives, false positives, and false negatives. This can help to understand the model's strengths and weaknesses.

Line Plot: Line plots can be used to visualize the predicted probabilities of bankruptcy over time. This can help to identify trends or patterns in the data.

Heatmap: A heatmap can be used to visualize the correlation between different variables. This can help to identify which variables are most strongly associated with the likelihood of bankruptcy.

Scatter Plot: Scatter plots can be used to visualize the relationship between two variables. This can help to identify patterns or relationships that may be useful for predicting bankruptcy.

Overall, the choice of visualization depends on the specific problem and the data available. It is important to carefully select visualizations that effectively communicate the insights from the data and the model's predictions.

3. Machine Learning: apply machine learning algorithms to extract insights from data. Python libraries like Scikit-Learn and TensorFlow will be used to train models, evaluate their performance, and make predictions on customers' activities. There are several machine learning algorithms that can be used to predict bank bankruptcy. Here are some commonly used algorithms:

Logistic Regression: This algorithm is used for binary classification problems, where the output variable is either 0 or 1. In the case of bank bankruptcy prediction, the output variable is whether or not the bank will go bankrupt. Logistic regression models can be used to model the relationship between various financial and non-financial variables and the likelihood of bankruptcy.

Decision Trees: Decision trees are a type of algorithm that uses a tree-like structure to model decisions and their possible consequences. In the case of bank bankruptcy prediction, decision trees can be used to identify the most important variables that contribute to the likelihood of bankruptcy.

Random Forests: Random forests are an ensemble learning technique that combines multiple decision trees to improve the accuracy of predictions. In the case of bank bankruptcy prediction, random forests can be used to identify the most important variables and their interactions.

Support Vector Machines (SVMs): SVMs are a powerful machine learning algorithm that can be used for binary classification problems. In the case of bank bankruptcy prediction, SVMs can be used to identify the most important variables that contribute to the likelihood of bankruptcy.

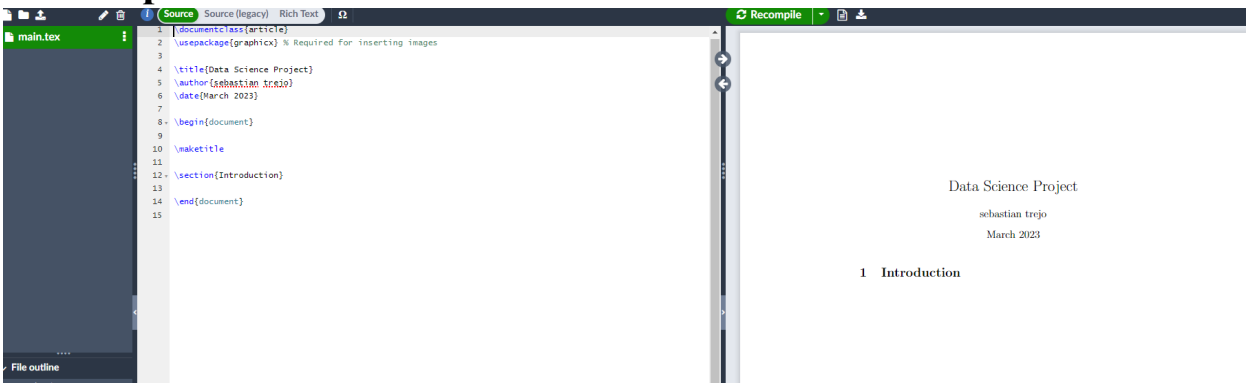
Neural Networks: Neural networks are a powerful machine learning algorithm that can be used for a variety of problems, including classification and regression. In the case of bank bankruptcy prediction, neural networks can be used to model the complex relationships between financial and non-financial variables and the likelihood of bankruptcy.

4. Exploratory Data Analysis (EDA): Exploratory data analysis. Python libraries like Pandas and NumPy are used to summarize the data, compute descriptive statistics, and generate visualizations that help to understand the underlying patterns and relationships in the data.

Overall, applied data science with Python in Jupyter notebooks is a powerful approach for extracting insights from data. apply machine learning algorithms and perform exploratory data analysis

4、 论文框架结构

4. **Paper framework**



LaTeX Introduction (Image example)

LaTex is a document preparation system that is widely used in academia, especially in mathematics, computer science, physics, engineering, and other technical fields. It is a typesetting system that allows you to create professional-looking documents with complex mathematical equations, figures, and tables.

LaTeX works by separating the content of a document from its formatting, allowing you to focus on the content while the system handles the layout and formatting. This makes it easier to create

high-quality documents, especially those with a lot of technical content.

Researchers often use LaTeX to write their papers because it offers several advantages over other typesetting systems:

It produces high-quality output: LaTeX is designed to produce high-quality documents, even when they contain complex equations or technical diagrams. It is easy to create complex mathematical equations: LaTeX has a built-in typesetting system that makes it easy to create complex equations and symbols. It is highly customizable: LaTeX allows you to customize the formatting and layout of your document to meet your specific needs. LaTeX is a powerful tool that can help researchers produce high-quality documents with ease.

Paper Framework

Introduction: Begin by introducing the background information on bank bankruptcy. Explain the importance of predicting bankruptcy in the banking industry and the potential impact on stakeholders. Also provide an overview of the approach to predicting bankruptcy using data science methods.

Literature Review: Conduct a thorough literature review to identify previous research on predicting bank bankruptcy. Summarize the methods and techniques used in previous tests, as well as their strengths and limitations (from each model). Identify gaps and aim to cover them up.

Research Questions and Hypotheses: “Can machine learning algorithms accurately predict bank

bankruptcy based on financial data?" or which data will be more useful from prediction analysis.

Methodology: Describe step by step the process of data extraction, data sources, data preprocessing techniques, and the specific machine learning algorithms used to predict bankruptcy. Supply these models with visualizations to better understanding.

Results: Present the results of the analysis, including visualizations or charts that help to explain each result. Compare the performance of the different machine learning models, including their accuracy, precision, recall, and other relevant metrics.

Discussion: Interpret the results of the analysis and relate them back to the research questions and hypotheses. Identify the strengths and limitations of each model, and suggest future research directions.

Conclusion: Summarize the main practice of the study and restate the importance of predicting bank bankruptcy in the banking industry.

References: List all the sources to cite in my paper in the appropriate citation format using LaTex.

5、 参考文献

5. **References**

1. https://serokell.io/blog/deep-learning-for-computer-vision
2. databases: https://www.kaggle.com/search?q=computer+vision
3. Deep Residual Learning for Image Recognition, Kaiming He, Xiangyu Zhang, Shaoqing Ren ,Jian Sun, Microsoft Research
4. VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE

RECOGNITION, Karen Simonyan * & Andrew Zisserman + Visual Geometry Group, Department of Engineering Science, University of Oxford {karen,az}@robots.ox.ac.uk
5. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications ,Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, Google Inc.
6. Densely Connected Convolutional Networks, Gao Huang Cornell University, Zhuang Liu*Tsinghua University, Laurens van der Maaten Facebook AI Research, Kilian Q. Weinberge Cornell University
7. MobileNetV2: Inverted Residuals and Linear Bottlenecks, Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, Google Inc.
8. CSPNET: A NEW BACKBONE THAT CAN ENHANCE LEARNING CAPABILITY OF CNN, Chien-Yao Wang, Hong-Yuan Mark Liao, I-Hau Yeh, Yueh-Hua Wu*, Ping-Yang Chen, Jun-Wei Hsieh
9. Grad-CAM: Visual Explanations from Deep Networks, via Gradient-based Localization, Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna, Vedanta · Devi Parikh · Dhruv Batra
10. AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMA RECOGNITION AT SCALE, Alexey Dosovitskiy*, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
11. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, Mingxing Tan 1 Quoc V. Le 1
12. Rethinking the Inception Architecture for Computer Vision, Christian Szegedy Google Inc, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens.Zbigniew Wojna University College London
13. Machine learning and sentiment analysis: Projecting bank insolvency risk, Diego Pitta de Jesus, Cássio da Nóbrega Besarria.
14. https://www.kaggle.com/code/mjbahmani/santander-ml-explainability/notebook
15. https://www.researchgate.net/publication/338289416_Review_of_bankruptcy_prediction_using_machine_learning_and_deep_learning_techniques

6、 工作进度安排**Work schedule**

序号 **No.**	设计(论文)各阶段任务 **Design (Thesis) tasks in each stage**	时间安排 **Schedule**
1	proposal	24/02/23
2		
3		

指导教师意见**Supervisor's comment**:

指导教师签名:
Supervisor's signature

年/**Year** 月/**Month** 日/**Day**

说明:

1. 学生应通过调研和资料搜集(要有10篇以上相关文章的阅读量),主动与指导教师讨论,在指导教师指导下完成开题报告。

2. 理、工科开题报告撰写不少于2500字,人文社科开题报告不少于3500字。

3. 开题报告字体为小四号,宋体(英文需用Times New Roman字体),行距为固定值20磅。

Directions:

1. Students should actively discuss with your supervisor after investigation and data collection (Read more than 10 related articles), and complete the proposal under the guidance of the supervisor.

2. Writing no less than 2500 words for the proposal.

3. The font of the proposal is 12pt, Times New Roman, and inter-line space should be 20 pounds.