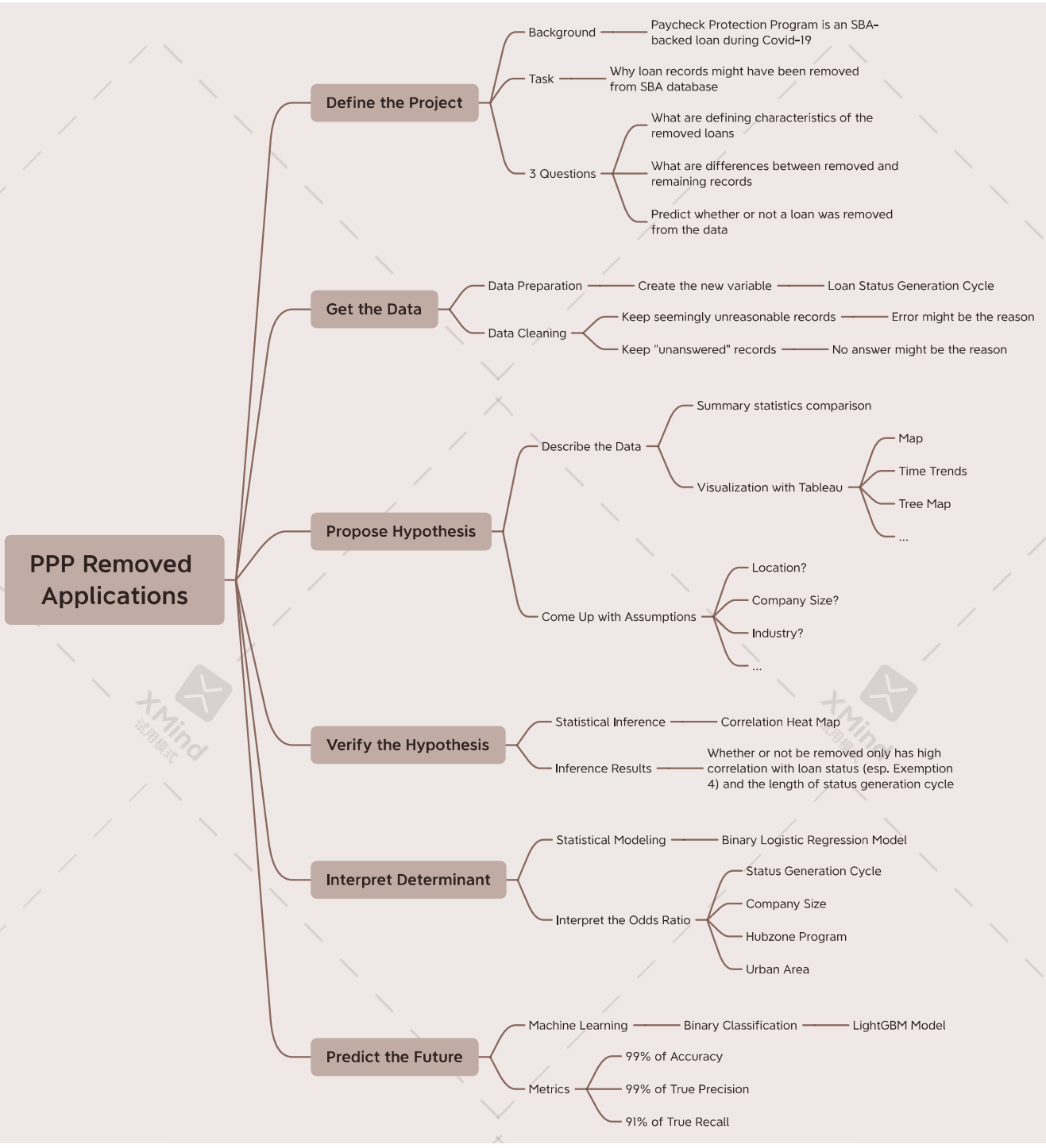
PPP Removed Applications
Define the Project
Background
Paycheck Protection Program is an SBA-backed loan during Covid-19
Task
Why loan records might have been removed from SBA database
3 Questions
What are defining characteristics of the removed loans
What are differences between removed and remaining records
Predict whether or not a loan was removed from the data
Get the Data
Data Preparation
Create the new variable
Loan Status Generation Cycle
Data Cleaning
Keep seemingly unreasonable records
Error might be the reason
Keep "unanswered" records
No answer might be the reason
Propose Hypothesis
Describe the Data
Summary statistics comparison
Visualization with Tableau
Map
Time Trends
Tree Map
...
Come Up with Assumptions
Location?
Company Size?
Industry?
...
Verify the Hypothesis
Statistical Inference
Correlation Heat Map
Inference Results
Whether or not be removed only has high correlation with loan status (esp. Exemption 4) and the length of status generation cycle
Interpret Determinant
Statistical Modeling
Binary Logistic Regression Model
Interpret the Odds Ratio
Status Generation Cycle
Company Size
Hubzone Program
Urban Area
Predict the Future
Machine Learning
Binary Classification
LightGBM Model
Metrics
99% of Accuracy
99% of True Precision
91% of True Recall