**Northeastern University**

**DS5500 33766 Capstone: Appl in Data Science SEC 01**

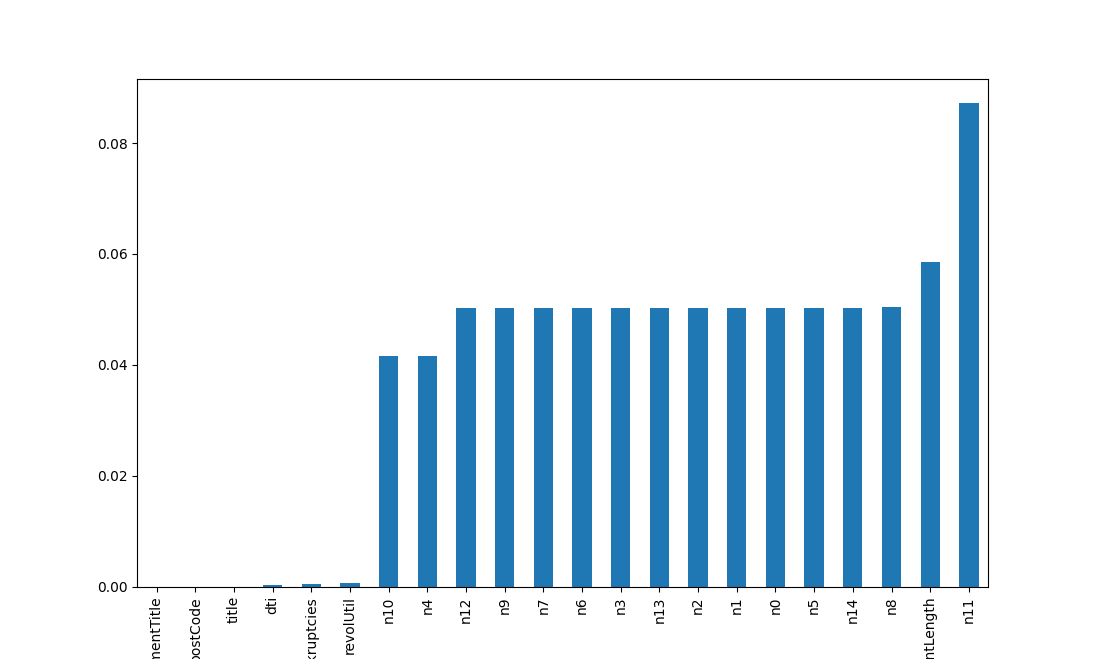
**Spring 2024**

**Data Cleaning and Preprocessing Methodologies**

In the capstone project, a meticulous approach to data cleaning and preprocessing was employed. The initial phase involved data correction and cleaning, where negative values in the "dti" column were set to zero, acknowledging the impossibility of negative debt-to-income ratios. The "employmentLength" field underwent a transformation, stripping text elements and converting it into a numeric format for enhanced analysis. Additionally, the "regionCode" was altered to a categorical object, and date-related fields such as "issueDate" and "earliesCreditLine" were parsed and transformed into more analyzable numeric formats.

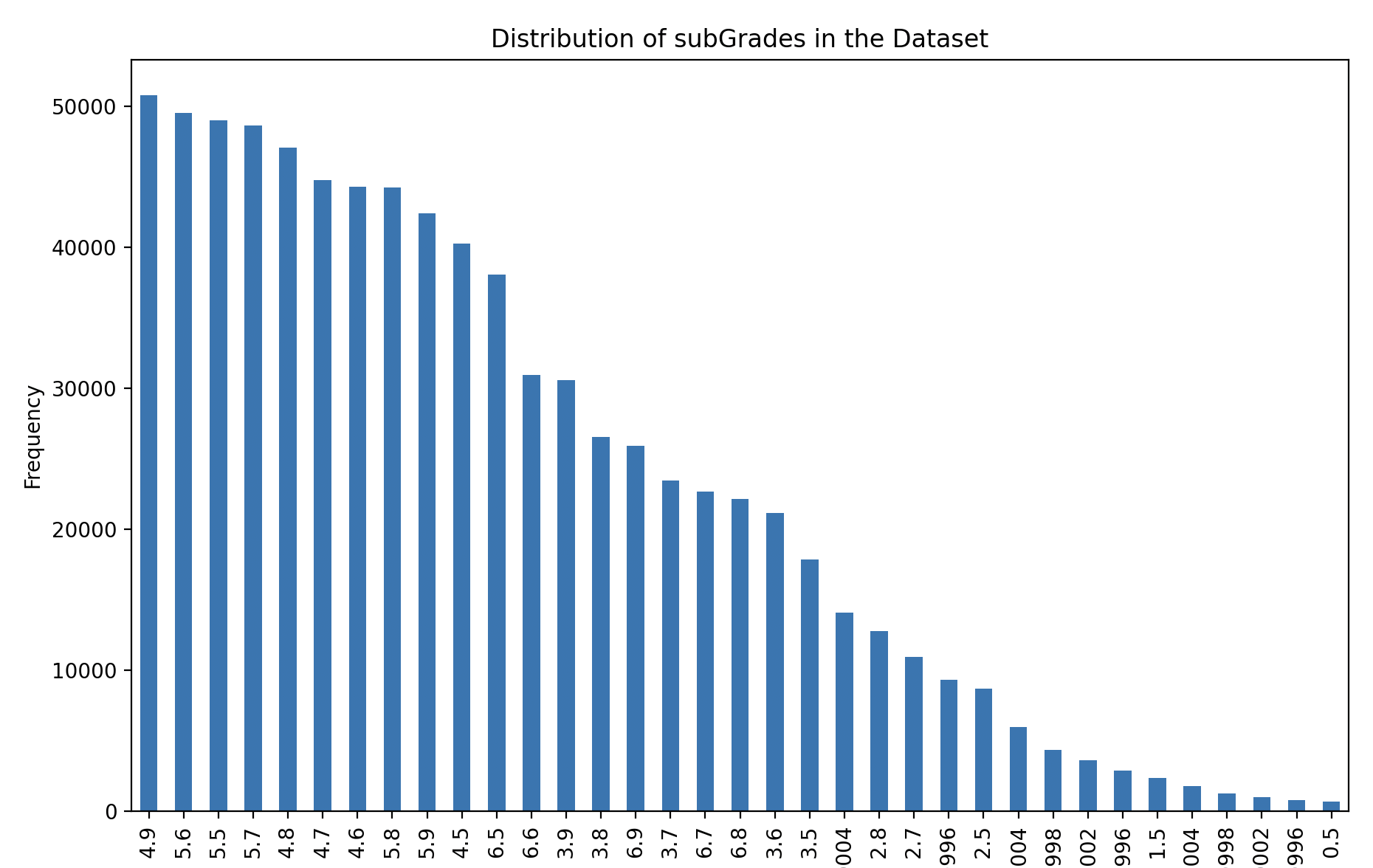
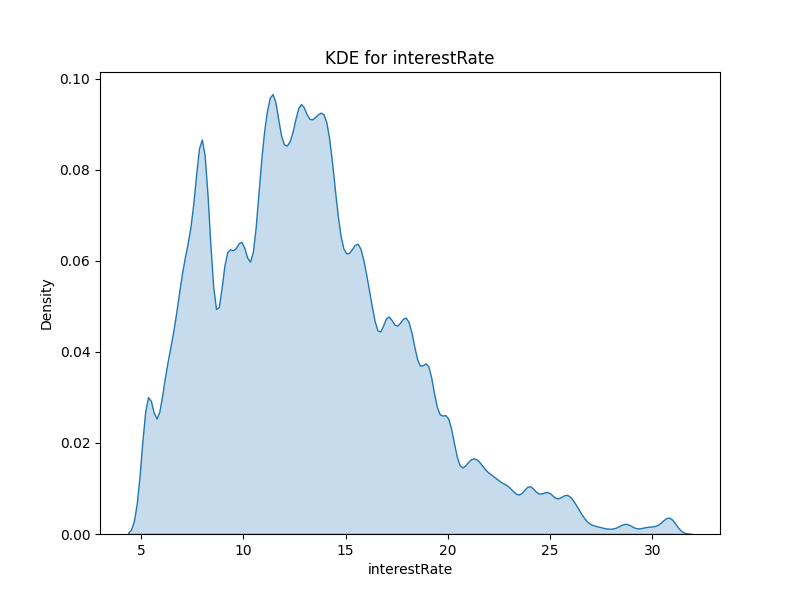
The project also included feature engineering and encoding. Here, ordinal encoding was applied to the "subGrade" feature, replacing it with numerical values, accompanied by the removal of the "grade" column. The month and year were extracted from "earliesCreditLine," converting it into a float type after basic arithmetic operations.

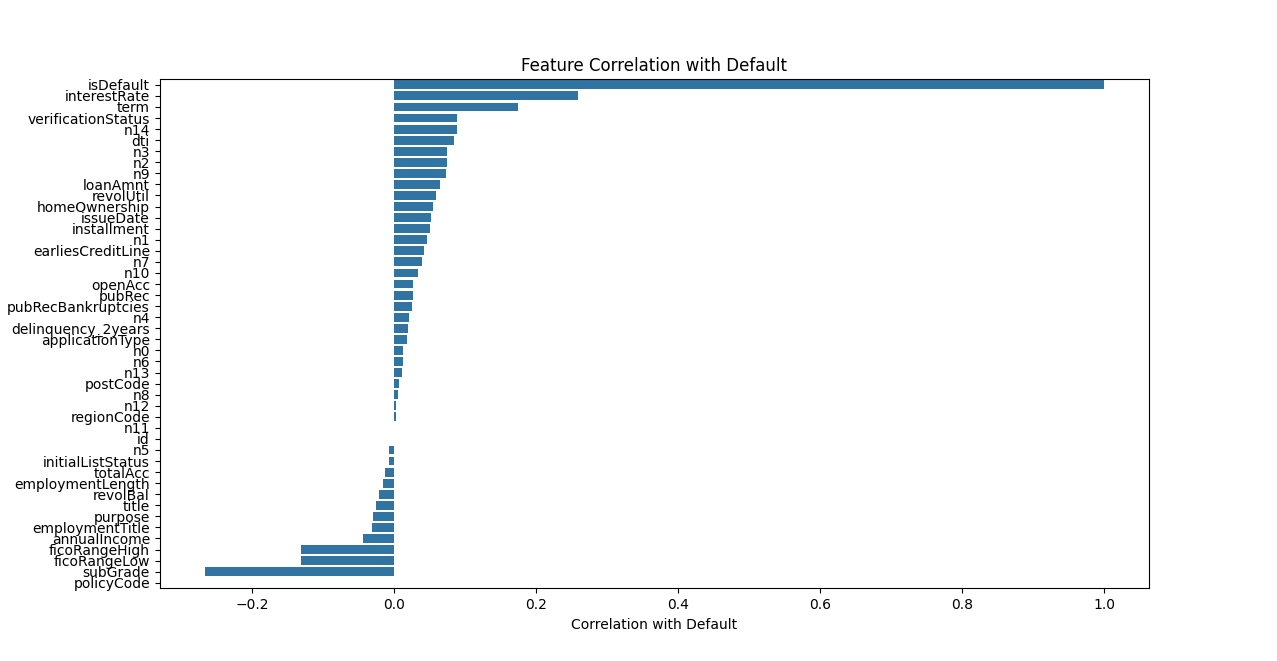
Handling missing values was another crucial step. Missing data in a range of features were filled with zeros, and a RandomForestRegressor was utilized to impute missing values in "employmentLength". The same approach was applied to the "dti", "pubRecBankruptcies", and "revolUtil" fields.



Normalization was executed using the StandardScaler to normalize features with extreme values. This was followed by feature selection and reduction, dropping less relevant features based on correlation analysis with the target variable and implementing one-hot encoding for categorical variables.

The exploratory data analysis included generating a collinearity matrix heatmap, density plots, and bar plots, providing visual insights into feature correlations.





**Techniques to Enhance Dataset Quality and Reliability**

The methodologies encompassed string manipulation and conversion, enhancing the dataset's consistency and usability. Predictive imputation using RandomForestRegressor ensured robustness and informativeness of the dataset, particularly where missing data could introduce bias. Standardizing features via normalization addressed the challenges posed by different feature scales, enhancing the performance of various machine learning algorithms.

**Transformation or Normalization Processes**

Date strings were parsed and converted into formats that provided more analytical value, such as extracting years or converting to DateTime objects. Feature encoding through ordinal and one-hot encoding was essential to convert categorical variables into a suitable format for modeling. Scaling, specifically through StandardScaler, was applied to normalize features with large or small values.

**Challenges Encountered and Addressed**

Addressing missing values posed a significant challenge, tackled through zero-filling and predictive modeling via a Random Forest. Feature engineering, particularly for the "earliesCreditLine," required complex and potentially error-prone string manipulations and mappings. Additionally, the calculate\_average\_metrics function in the second script indicated inefficiencies due to iterative row-wise operations, a common challenge in handling large datasets.

This comprehensive approach to data cleaning and preprocessing underpinned the reliability and robustness of the capstone project's dataset, setting a solid foundation for subsequent analytical processes.