FICO Analytic Challenge Week 3 Handout

Feature Engineering I

Summary

Feature extraction is crucial in fraud detection neural network modeling because it transforms raw data into informative, non-redundant features that enhance the model's ability to distinguish between fraudulent and legitimate activities. By highlighting patterns, relationships, and anomalies in the data, effective feature extraction improves the model's accuracy, efficiency, and ability to generalize to new, unseen fraud scenarios.

Variable Importance in Modeling

Variables or features are essential components of machine learning models. They represent the data used to learn patterns and make predictions. Proper handling and selection of these variables are critical for developing models that are robust, accurate, and interpretable.

Feature Engineering

Feature engineering involves creating or modifying features to enhance model performance. It applies domain knowledge to transform raw data into meaningful representations, including transforming features, creating new ones, and aggregating data.

Binary Features

Binary features are attributed with two possible values, often representing distinct categories or states, such as 0 and 1. They simplify decision-making in models and can be crucial for tasks like fraud detection, where they help improve model performance by providing clearer distinctions.

Key Takeaways

- **Variable Importance Matters**
    - **Model Performance**: Relevant variables enhance accuracy and efficiency.
    - **Interpretability**: Key variables make models more understandable.

    - **Data Collection**: Guides future data gathering to include valuable information.
- **Techniques for Determining Variable Importance**
    - Feature importance ranking
    - Correlation analysis
    - Domain expertise
- **Feature Engineering Steps**
    - **Transforming Features**: Enhance predictive power through mathematical transformations.
    - **Creating New Features**: Develop new variables to better represent the problem.
    - **Aggregating Data**: Combine or summarize data for improved analysis.
- **Binary Features**
    - Simplify decision-making by representing two states or categories.
    - Examples include distinguishing between high/low dollar transactions, time of transaction, and transaction types.

Overall Learning Goal

To understand the importance of variables in modeling, including how to determine and apply variable importance techniques, and to effectively utilize feature engineering and binary features to build more accurate, interpretable, and efficient machine learning models.

1. Variable Importance in Modeling

Variables, or features, are the building blocks of machine learning models. They represent the data used for learning patterns and making predictions. Proper handling and selection of variables are crucial for building robust, accurate, and interpretable models.

Why Variable Importance Matters:

- **Model Performance:** Focusing on relevant variables enhances model accuracy and efficiency.
- **Interpretability:** Identifying key variables makes models more understandable.
- **Data Collection:** Informs future data collection efforts to gather valuable information.

Techniques to determine variable importance include feature importance ranking, correlation analysis, and domain expertise.

2. Feature Engineering

Feature engineering involves creating or modifying features to improve model performance. This step applies domain knowledge to transform raw data into meaningful representations.

Feature Engineering Steps:

- **Transforming Features:** Applying mathematical transformations to enhance predictive power.
- **Creating New Features:** Generating new variables that better represent the problem.
- **Aggregating Data:** Combining multiple variables or summarizing data for better analysis.

3. Binary Features

A binary feature is an attribute that can take on only two possible values. These values often represent two distinct categories or states, such as 0 and 1. Binary features are commonly used in neural networks because they simplify the decision-making process. For example, in a dataset predicting whether a transaction is fraud or not, it can be helpful to know whether it was an online purchase or not. Another example is if the transaction happened during the night or in daytime. There is a sea of binary features we can create from the information we have in our dataset, but it is important that we choose smartly in the sense that it helps prediction with better performance.

Main Binary Features:

- **High/Low Dollar**
- **IS_0_TO_5_AM**
- **CP/CNP**

4. Exercised & Challenges

1. **Review:** Deep dive into data analysis again. Try to understand fraud and non-fraud behaviors by looking at data from different aspects.
2. **CP/CNP:** Remember the analysis on card present/not present? In cases where the card was not present, the fraud to non-fraud ratio was higher. What does this piece

of information tell us? Probably it can help us better predict whether a transaction is fraud. What we can do is to create a feature called "is_CNP" based on the analysis in week 2, where it is 1 if the card is not present and 0, otherwise.

3. **Explore:** Look through the data and try to visualize different features:
    a. See if you can find patterns between features and the output.
    b. Find patterns and correlations between different features.
    c. Search and think about possible new features that you can extract from existing ones.
    d. Use analysis on data to prove/reject hypotheses about the new features you have created.

5. References

1. https://medium.com/pythons-gurus/mastering-feature-engineering-c8ec714d9656
2. https://towardsdatascience.com/feature-engineering-for-machine-learning-eb2e0cff7a30
3. https://medium.com/dataman-in-ai/how-to-create-good-features-in-fraud-detection-de6562f249ef