Northeastern

Case Study 3 - One Dataset Many Users

Part 2 - Summary

INFO6105 37364 Data Science Engineering Methods and Tools

March 22, 2019

Project Group 2

Krina Devang Thakkar

Hardik Bhupendra Soni

Yash Thakur Lekhwani

Yashashri Shiral

Ritesh Manek

Data Preparation – Data Cleansing

In order to have accurate data analysis, we cleaned data by dropping the irrelevant columns and filled out missing values.

We first dropped all of the columns that had:

- More than 50% missing values
- No values
- Columns that had no purpose in predicting the interest rate

Then we checked the null values in the remaining columns to fill them with the median value of the column.

Data Preparation – Data Preprocessing

Once the data has been cleaned, we enhanced our dataset that could be used by our ML model. First, we converted the string values to numerical values. For example, emp_length was cut down to "5" instead of "5 Years". Post this we performed, One Hot Encoding to transform our categorical features that could be understood by the model. A simple example of this is as follows:

ID	Purpose
1	Car
2	Education

ID	purpose_car	purpose_education
1	1	0
2	0	1

Feature Tools vs manual Feature Engineering

In the context of machine learning, a feature can be described as a characteristic or a set of characteristics, that explains the occurrence of a phenomenon. When these characteristics are converted into some measurable form, they are called features.

Featuretools is an open source library for performing automated feature engineering. It is a great tool designed to fast-forward the feature generation process, thereby giving more time to focus on other aspects of building the machine learning model. In other words, it makes your data "machine learning ready".

Before taking Feature tools for a spin, there are three major components of the package that we should be aware of:

- Entities: An Entity can be considered as a representation of a Pandas DataFrame. A collection of multiple entities is called an Entity set.

- Deep Feature Synthesis (DFS): DFS has got nothing to do with deep learning. Don’t worry. DFS is actually a Feature Engineering method and is the backbone of Featuretools. It enables the creation of new features from single, as well as multiple data frames.
- Feature primitives: DFS create features by applying Feature primitives to the Entity-relationships in an EntitySet. These primitives are the often-used methods to generate features manually. For example, the primitive “mean” would find the mean of a variable at an aggregated level.

Manual Feature Engineering, on the contrary, forces the user to manually decide and form a certain feature.

Featuretools can reduce the machine learning development time by 10x compared to manual feature engineering while delivering better modeling performance.

We achieved the features of transforming the features into absolute and percentile that provide us with absolute and percentile values of the features. On the contrary, we had to explicitly write an algorithm to classify weather a loan is deemed as good or bad, for investment. This was a feature in our manual feature engineering.