Electronics & ICT Academy
National Institute of Technology, Warangal

Post Graduate Program in

Machine Learning and Artificial Intelligence

ML Model for Auto Insurance Industry

Capstone Project - III: AI-ML PG Program by NITW E&ICT Academy

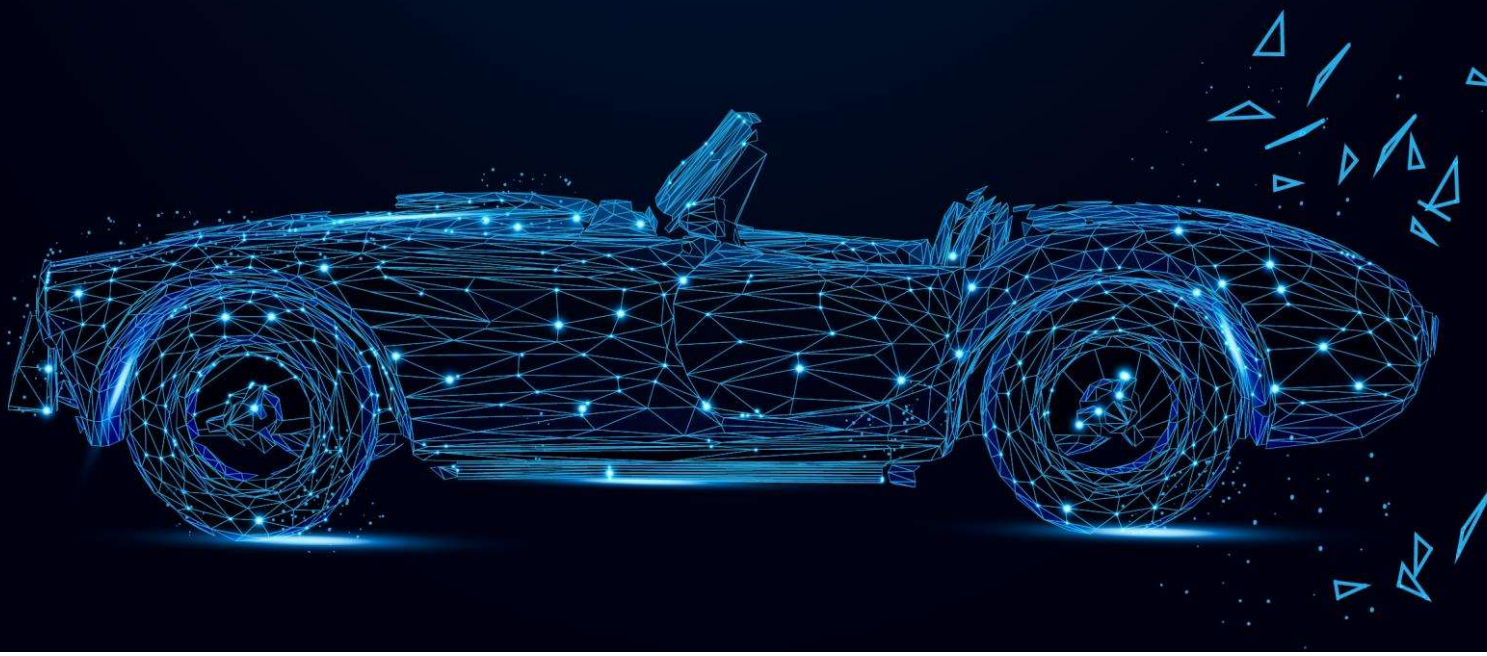


Table of Contents

Aim of the Project

The aim of the project is to build a Machine Learning Model to predict whether an owner will initiate an auto insurance claim in the next year.

Background

The auto insurance industry is witnessing a paradigm shift. Since auto insurance company consists of homogenous good thereby making it difficult to differentiate product A from product B, also companies are fighting a price war (for insurance price). On top of that, the distribution channel is shifting more from traditional insurance brokers to online purchases, which means that the ability for companies to interact through human touchpoints is limited, and customers should be quoted at a reasonable price. A good price quote is one that makes the customer purchase the policy and helps the company to increase the profits.

Also, the insurance premium is calculated based on more than 50+ parameters, which means that traditional business analytics-based algorithms are now limited in their ability to differentiate among customers based on subtle parameters.

Use Cases

The model shall mainly support the following use cases:

1. **Conquering Market Share:** Capture market share by lowering the prices of the premium for the customers, who are least likely to claim.

2. **Risk Management:** Charge the right premium from the customer, who is likely to claim insurance in the coming year

3. **Smooth Processing:** Reduce the complexity of pricing models. Most of the transactions are happening online with larger customer attributes (thanks to the internet and social media). Harness the power of huge data to build complex ML models

4. **Increased Profits:** As per industry estimate 1% reduction in the claim can boost profit by 10%. So, through the ML model, we can identify and deny the insurance to the driver who will make a claim. Thus, ensuring reduced claim outgo and increased profit.

Part of the model development is to identify and prioritize the above use cases.

Process Flow

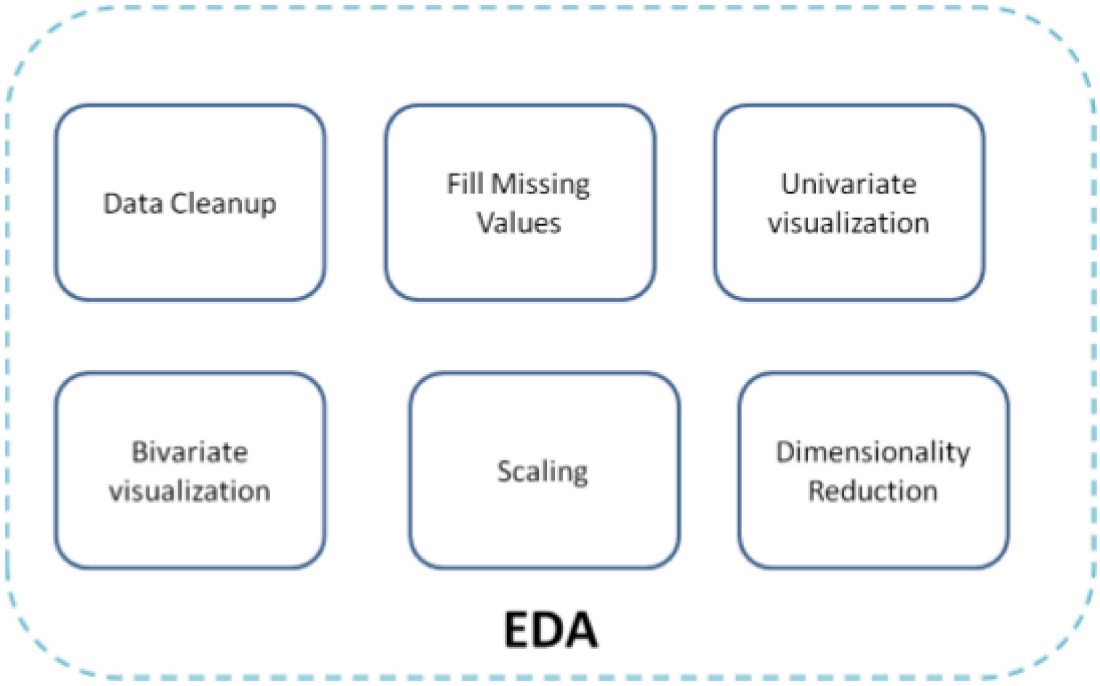
The Machine Learning model mainly consist of two phases:

1. EDA (Exploratory Data Analysis):

Analyze the datasets to summarize their main characteristics (with visual methods). A statistical model can be used, primarily EDA can be used to see what the data can tell us beyond the formal modeling or hypothesis testing task.

Following tasks can be performed as a part of EDA:

- Scaling/Normalization
- Fill the missing values
- Feature selection & engineering

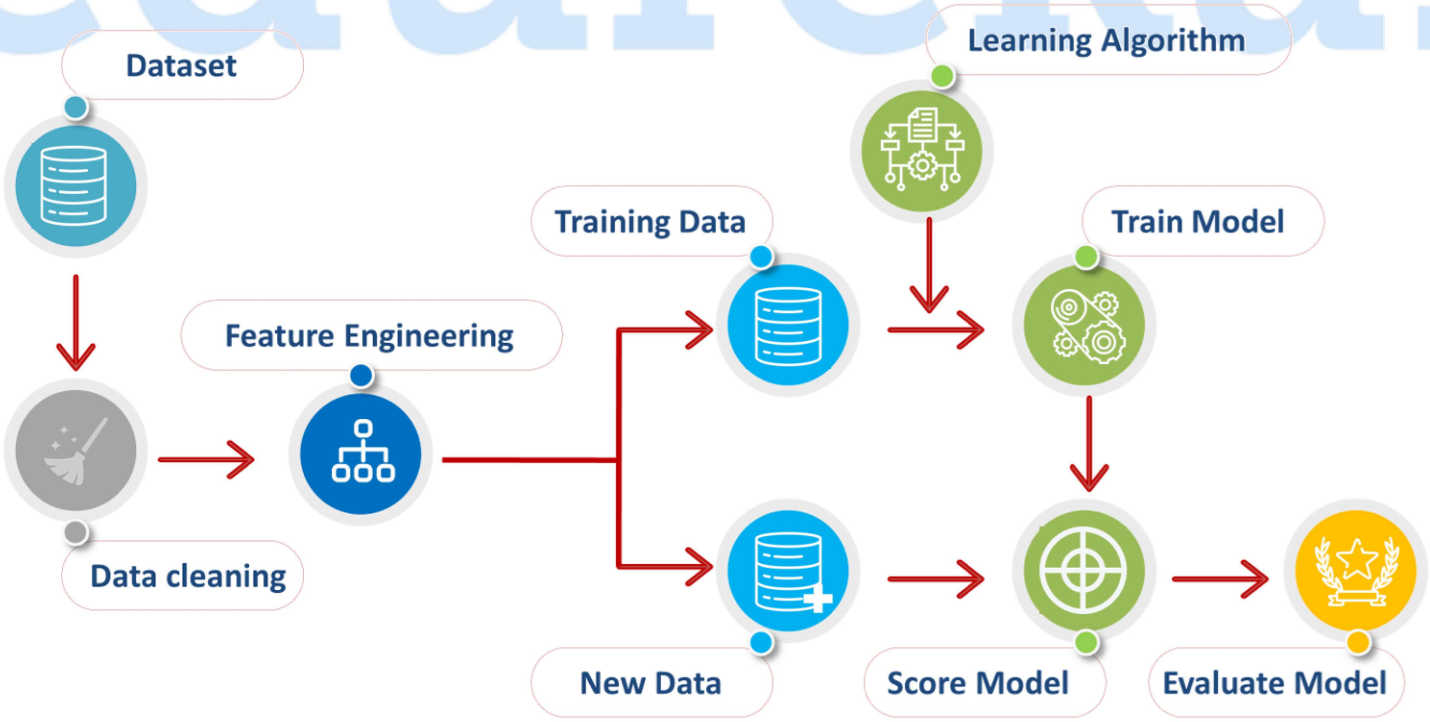


2. Machine Learning Modeling:

After EDA, the modeling comes into the process. The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data. The term "ML model" refers to the model artifact that is created by the training process.

The training data must contain the correct answer (target or target attribute). The learning algorithm finds patterns in the training data that maps the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.

You will use the ML model to get predictions on new data for which you will not know the target.



Following tasks can be performed as a part of Modeling:

- Start with the basic model but eventually move towards ensemble
- Use Deep Learning with sklearn MLPClassifier and check if the Neural Network Model is better than traditional models
- Arrival at a model with best f1-score

Dataset Description

The project involves the use of a dataset with 600k training data and 57 features/data. In the train and test data, features that belong to similar groupings are tagged as such in the feature names (e.g., ind, reg, car, calc). In addition, feature names include the postfix bin to indicate binary features and cat to indicate categorical features. Features without these designations are either continuous or ordinal. Values of -1 indicate that the feature was missing from the observation. The target column signifies whether a claim was filed for that policy holder.

Tasks to be performed

Following are the deliverables (.ipynb files), which needed to be developed with respect to

Exploratory Data Analysis:

1. Write at least 3 important inferences from the data above
2. Is the data balanced? Meaning are targets 0 and 1 in the right proportion?
3. How many categorical features are there?
4. How many binary features are there?
5. Write inferences from data on interval variables.
6. Write inferences from data on ordinal variables.
7. Write inferences from data on binary variables.

8. Check if the target data is proportionate or not. Hint: Below than 30% for binary data is sign of imbalance
9. What should be the preferred way in this case to balance the data?
10. How many training records are there after achieving a balance of 12%?
11. Which are the top two features in terms of missing values?
12. In total, how many features have missing values?
13. What steps should be taken to handle the missing data?
14. Which interval variables have strong correlation?
15. What's the level of correlation among ordinal features?
16. Implement Hot Encoding for categorical features
17. In nominal and interval features, which features are suitable for StandardScaler?
18. Summarize the learnings of ED

Following are the deliverables (.ipynb files), which needed to be developed with respect to Modeling :

1. The Simple Logistic Regression Model seems to have high accuracy. Is that what we need at all? What is the problem with this model?
2. Why do you think f1-score is 0.0?
3. What is the precision and recall score for the model?
4. What is the most important inference you can draw from the result?
5. What is the accuracy score and f1-score for the improved Logistic Regression model?
6. Why do you think f1-score has improved?

7. For model LinearSVC play with parameters – dual, max_iter and see if there is any improvement

8. SVC with Imbalance Check & Feature Optimization & only 100K Records → is there improvement in scores?

9. XGBoost is one the better classifiers -- but still f1-score is very low. What could be the reason?

10. What is the increase in number of features after one-hot encoding of the data?

11. Is there any improvement in scores after encoding?

12. If not missing a positive sample is the priority which model is best so far?

13. If not marking negative sample as positive is top priority, which model is best so far?

14. Do you think using AdaBoost can give any significant improvement over XGBoost?

15. MLPClassifier is the neural network we are trying. But how to choose the right no. of layers and size?

16. At what layer size we get the best f1-score?