Predicting Claims from FNOL

By
Satya Venkataswamy

UNDERSTANDING THE PROBLEM

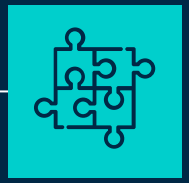
Business Objective

Given the historical data of First Notification of Loss data, predict the claim amounts of each of the insurers.

Problem Statement

Since Incurred values is a continuous value, it is a regression task.

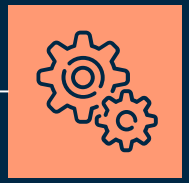
My Approach



01

ISSUES & DATA CLEANING

Otaining the data, issues with the data, clraning techniques identified



02

FEATURE SELECTION & MODEL BUILDING

EDA, Feature Selection and Model Selection through Nested Cross Validation

03

FURTHER STEPS & IMPLEMENTATION

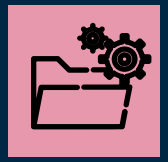
Suggestions on Model Improvements and Practical challenges.

Overview of the Dataset

Dataset Dimensions

7,691 records and

46 features



Data Types

9 categorical features

and 37 numerical ones

Time Frame

Claims range from April

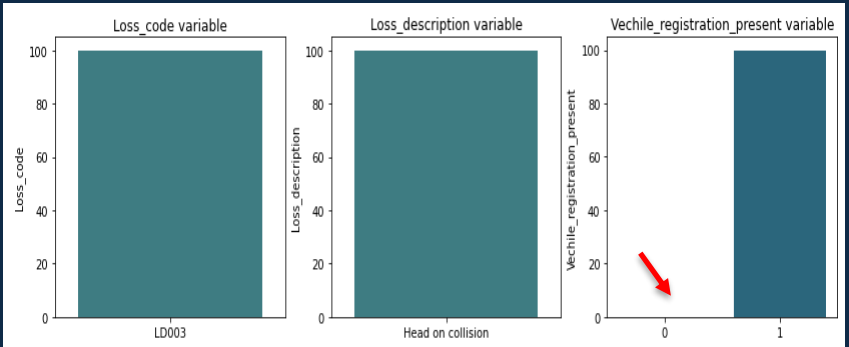
2003 to June 2015

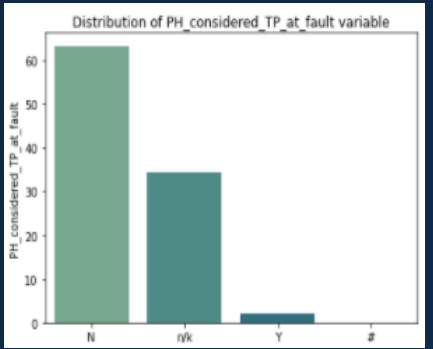
Missing Values

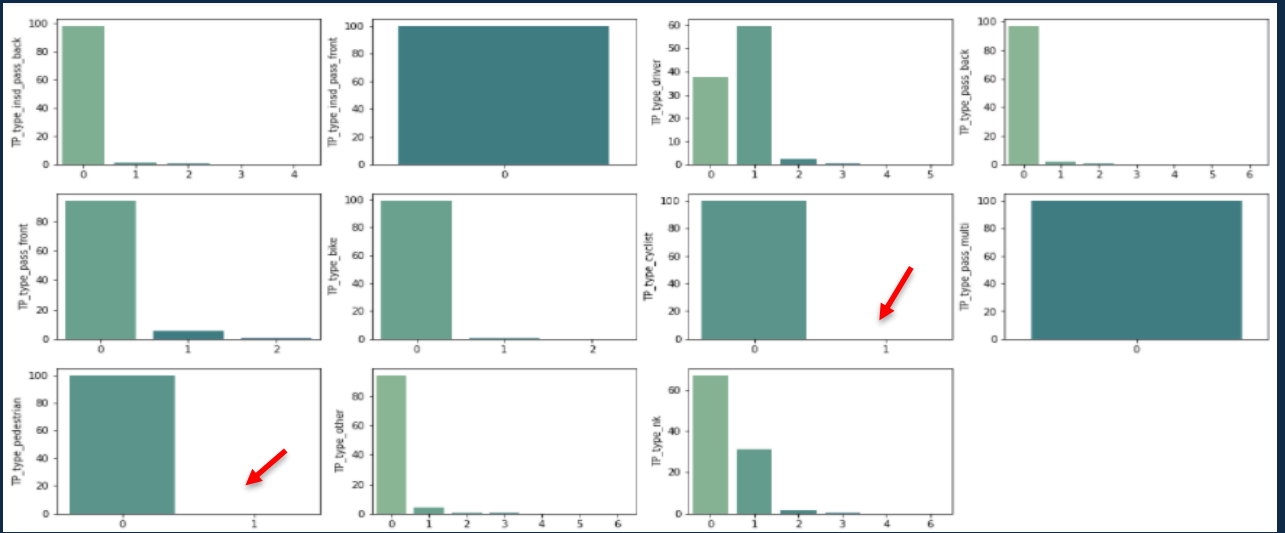
~4.5% of missing values in the weather_conditions column

Issues with the Dataset

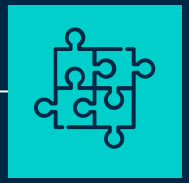
- 3 Columns (**Loss_code, Loss_description** and **Vehicles_registration_present**) with zero variance or near zero variance.
- 4 **Third party type** related variables with zero or near zero variance amongst them.
- Records with **'#'** values in the **PH_considered_TP_at_fault** variable.
- **3 records** with Negative values in the **Notification_Period** variable.
- **Multivariate analysis** and **correlation maps** of TP variables







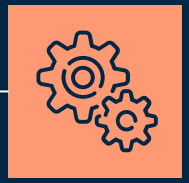
My Approach



01

ISSUES & DATA CLEANING

Otaining the data, issues with the data, clraning techniques identified



02

FEATURE SELECTION & MODEL BUILDING

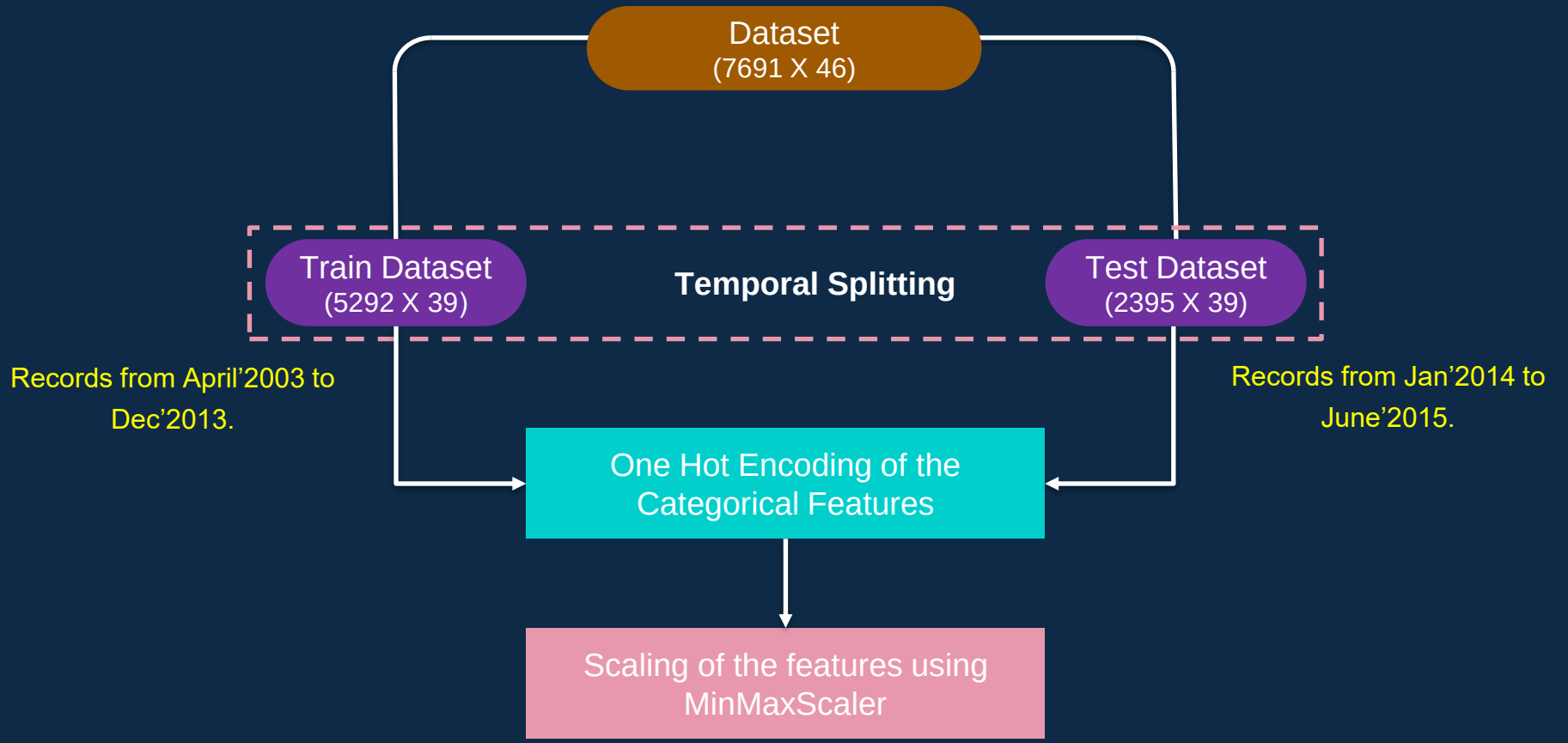
EDA, Feature Selection and Model Selection through Nested Cross Validation

03

FURTHER STEPS & IMPLEMENTATION

Suggestions on Model Improvements and Practical challenges.

Splitting Schema of the Dataset



Feature Selection Approach

Baseline Model (Linear Regression)
MAE Score: 12,019

Feature Selection Methods

Recursive Feature Elimination with CV

22 Features were selected

Base Model (Linear Regression)

Feature Selection Method 1
MAE Score: 11,553.29

Lasso Regression model coefficients with CV

22 Features were selected

Base Model (Linear Regression)

Feature Selection Method 2
MAE Score: 11,539.89

HyperParameter Tuning and Model Evaluation

Experiment Description

EXP1: Features Selected through Recursive Feature Elimination

2x5 Nested Cross Validation

Inner Loop — **Hyperparameter Optimization**

Outer Loop — **Model Evaluation**

Feature Importance on Best Models

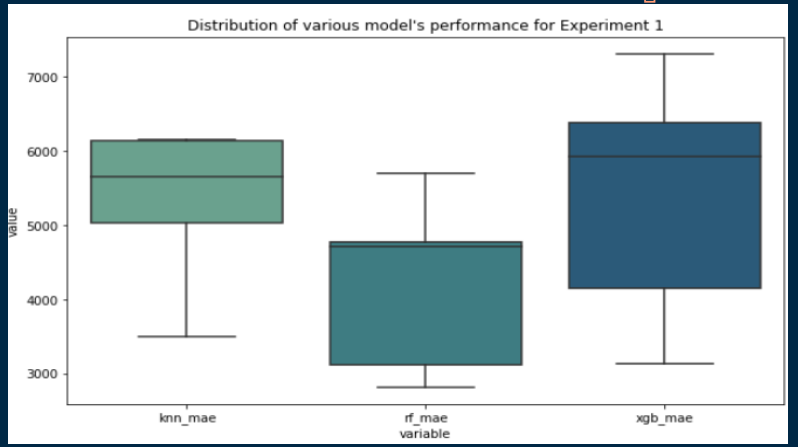
EXP1(A): Features with high feature importance are isolated and Exp1 is repeated

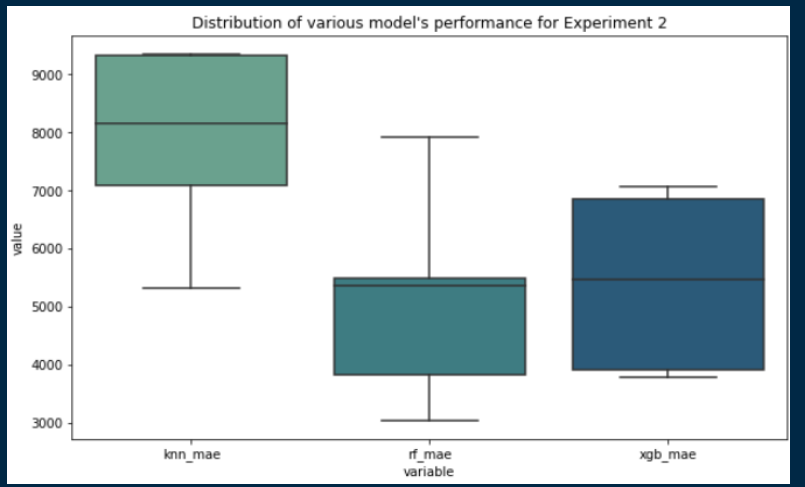
EXP2: Features Selected through Lasso Regression Coefficients

EXP2(A): Features with high feature importance are isolated and Exp1 is repeated

Same Process as above

Distribution of Results



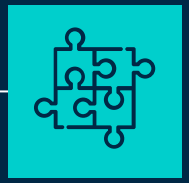


Model Selection

Experiment	Description	Model	Test_MAE_95%CI_Lowerbound	Test_MAE_95%CI_Upperbound
Exp-1	Featrues Selected from RFE feature selction technique	Random_Forest	1932.91	3617.57
Exp-1(a)	Further Feature selection from Exp1	Random_Forest	1877.89	3334.16
Exp-2	Feature Selected from Lasso feature selction technique	Random_Forest	2231.56	3777.54
Exp-2(a)	Further Feature selection from Exp2	Random_Forest	2223.35	3215.76

1. Leveraged Bootstrapping in order to create a distribution of Test dataset performances.
2. As a part of bootstrapping the selected model (best hyperparameters) of each experiment was fit on the train dataset and evaluated on the Test dataset.
3. The above process was repeated for a total of 20 iterations and the Mean Test MAE and Standard Error of the MAE was used to create a 95% confidence interval.
4. (By assuming the test MAE follows a t-distribution as the number of samples are less than 30 and the standard deviation of the population is unknown).

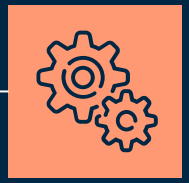
My Approach



01

ISSUES & DATA CLEANING

Otaining the data, issues with the data, clraning techniques identified



02

FEATURE SELECTION & MODEL BUILDING

EDA, Feature Selection and Model Selection through Nested Cross Validation

03

FURTHER STEPS & IMPLEMENTATION

Suggestions on Model Improvements and Practical challenges.

Future Steps and Challenges

Future Steps

1. Explore the performance of other categorical encoding techniques such as Target Encoding etc..
2. Use of Non-linear base models in feature selection techniques.
3. Use Feature Selection techniques such as Mutual Information Gain in SelectKBest Models in order to isolate the important features.
4. Experiment with other implementations of Boosting techniques such as LightGBM etc..
5. Use of RMSE score as an evaluation metric in order to minimize the error on outlier data points.

Challenges

1. Since Nested Cross validation is employed, processing needs would be on the higher side.

Thank You