Car Insurance Prediction

Machine Learning

Supervised by Dr. Maher Abdulrahman

Our Team

Naim Abdeljawad

Tripoli
52130470

Charbel Saliba

Tripoli
52130306

Contents maps

Introduction

Datasets

Data Cleaning

PreProcessing

Algorithms

Our Conclusion

1

Introduction

The project focuses on leveraging the capabilities of machine learning to revolutionize the way we assess and manage car insurance risks. By predicting accident frequency and tailoring insurance premiums accordingly, we aim to bring precision and fairness to the insurance industry.

2

About the Dataset

Datasets

Policy

Features

- 21 features

Size

Subtitle

- 227469 instances

Target

- 1 numerical label

Claims

Features

- 24 features

Size

- 62521 instances

Target

- 1 numerical label

2

Policy Dataset

Features

1. CUST_ID
2. EXECUTIVE
3. BODY
4. MAKE
5. MODEL
6. USE_OF_VEHICLE
7. MODEL_YEAR
8. CHASSIS_NO
9. REGN
10. POLICY_NO
11. POL_EFF_DATE
12. POL_EXPIRY_DATE
13. SUM INSURED
14. POL_ISSUE_DATE
15. PREMIUM2
16. DRV_DOB
17. DRV_DLI
18. VEH_SEATS
19. PRODUCT
20. POLICYTYPE
21. NATIONALITY

2

Claims Dataset

Features

1. Account Code
2. DATE OF INTIMATION
3. DATE OF ACCIDENT
4. PLACE OF LOSS
5. CLAIM NO
6. AGE
7. TYPE
8. DRIVING LICENSE ISSUE
9. BODY TYPE
10. MAKE
11. MODEL
12. YEAR
13. CHASIS NO
14. REG
15. SUM INSURED
16. POLICY NO
17. POLICY START
18. POLICY END
19. INTIMATEDAMOUNT
20. INTIMATEDSF
21. EXECUTIVE
22. PRODUCT
23. POLICYTYPE
24. NATIONALITY

Data cleaning process

3

1. **Clean 'MAKE' for 'SUM INSURED':**
   - Replace outliers in 'SUM INSURED' with corresponding values in 'MAKE' => 'SUM INSURED' dictionary.
2. **Clean 'Intimated Amount' (Severity Target):**
   - Group by 'MAKE,' calculate mean 'Intimated Amount' for each 'MAKE.'
   - Fill NaN and outliers in 'Intimated Amount' with the mean of each 'MAKE.'
3. **Column Matching for Merge:**
   - Match column names in 'policies' and 'claims' for merge purposes.
4. **Merge 'Claims' and 'Policies':**
   - Perform an outer join to retain the shape of the larger dataset (policies).
   - Fill NaN for every 'POLICY NO' without a claim.
5. **Check Differences in X and Y Columns:**
   - Examine differences between 'X' (policies) and 'Y' (claims) columns.

Data cleaning process

3

6. **Merge X and Y Columns:**
   - Combine 'X' and 'Y' columns into one dataset.
7. **Clean 'MODEL YEAR':**
   - Use a random value between (mean-5) and (mean+5).
8. **Clean Premium:**
   - Address negative values observed in premium graphs.
9. **Vehicle Seats Cleaning:**
   - Clean based on the 'BODY' column.
10. **Clean BODY:**
    - Drop NaN in 'BODY' to facilitate vehicle seats cleaning.
11. **Calculate Percentage of Rows:**
    - Identify rows where the insured's 'DOB' is greater than 'DLI' (Driving License Issue).
    - Swap columns ('DRV_DOB' & 'DRV_DLI').

Data cleaning process

3

12. **Age Cleaning:**
    - First step using 'POLICY START' & 'DOB.'
    - Second step by filling outliers with the mean.
13. **Clean 'REG':**
    - Remove repeated values in 'REG.'
14. **Clean Dates:**
    - Ensure 'Date of Intimation' > 'Date of Accidents.'
    - Clean 'Date of Intimation.'
15. **Clean Driver's License Issue:**
    - Convert 'DLI' to an age column, filling outliers with NaN.
    - Calculate age using 'POLICY START' - 'DLI.'
    - Fill NaN in 'DLI_AGE' with mean value for the corresponding age in the row.

Data cleaning process

3

16. **Age Cleaning:**
    - First step using 'POLICY START' & 'DOB.'
    - Second step by filling outliers with the mean.
17. **Clean 'REG':**
    - Remove repeated values in 'REG.'
18. **Clean Dates:**
    - Ensure 'Date of Intimation' > 'Date of Accidents.'
    - Clean 'Date of Intimation.'
19. **Clean Driver's License Issue:**
    - Convert 'DLI' to an age column, filling outliers with NaN.
    - Calculate age using 'POLICY START' - 'DLI.'
    - Fill NaN in 'DLI_AGE' with mean value for the corresponding age in the row.

PreProcessing

4

1. Target Variable Creation:
   - Created 'freq' as the target variable, representing the number of redundant policies based on different dates of accidents.
2. Column Removal:
   - Dropped 'policy_no' and 'date_of_accident' columns as they were only used in calculating the target variable and are not needed for model training.
3. Label Encoding:
   - Applied label encoding to categorical columns 'REG', 'BODY', and 'MAKE' to convert them into a numerical format suitable for machine learning models.

Machine Learning Models

5

1. Frequency Models
   - Random Forest
   - Poission Regression
   - TensorFlow Kerras

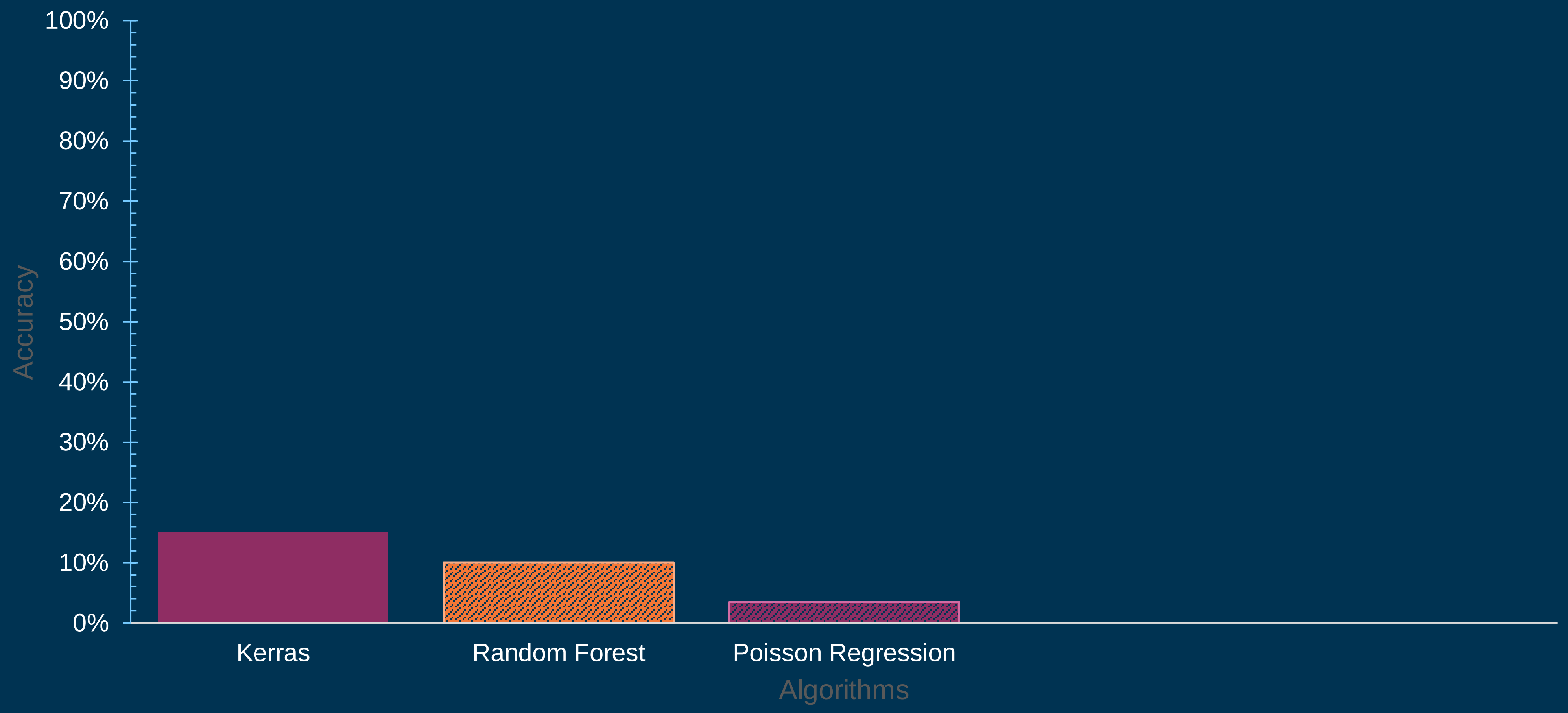
2. Frequency Models
   - Random Forest

Models Metrics

5

MODELS	Algorithms	Mean Squared Error	R-Squared
Frequency	TensorFlow KERRAS	0.8193864309823716	0.15030596762202453
	RANDOM FOREST	0.2726475247068885	0.09877931527912065
	POISSON REGRESSION	0.29169516679095187	0.03581843177294386
Severity	RANDOM FOREST	1129122623.7017262	0.8851468731070606

Frequency Models Accuracy



Severity Models Accuracy

5

Conclusion

6

Frequency

As we scrutinize the performance of our three models, it becomes evident that while all models exhibit lower-than-desired accuracy, one stands out as the most promising candidate. Despite the challenges we faced with overall accuracy, 'TeansowFlow Kerras' demonstrates a comparatively higher accuracy rate, making it the clear choice for our predictive analytics solution. This model, although not achieving our ideal benchmark, outshines its counterparts and holds the potential to significantly elevate the reliability of our predictions in real-world scenarios

Severity

In the exploration of predictive modeling on our two datasets, the Random Forest algorithm has emerged as a standout performer, achieving an impressive R2 score of 85%. This compelling result signifies the robustness and effectiveness of the Random Forest model in capturing complex relationships within our data.

The high R2 score, indicative of 85% variance explained, underscores the model's ability to provide accurate and reliable predictions. As we conclude this analysis, it's clear that Random Forest has demonstrated its prowess in handling the intricacies of our datasets, showcasing its adaptability and versatility.

“

You can have data without information, but you cannot have information without data.”

- Daniel Keys Moran

Thanks