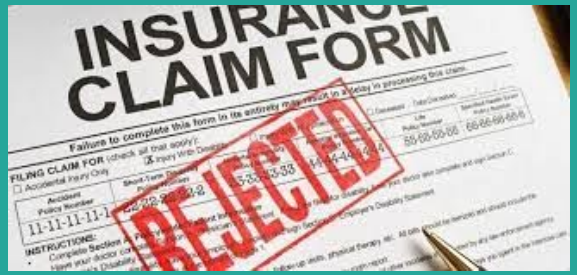
InsureIntel

Insurance Claim Fraud Detection



Guide - Prof. Asim Tewari

Vishvesh Kodihal	203310013
Amit Meena	190100013
Mudit Rathore	190100078
Dushyant Patil	203170020
Neha Singhal	180070038

Project Objective

Aim

The aim of our startup is to provide easy to use and **reliable** solutions to **predict fraud insurance claims** using Machine Learning capabilities and maintaining the privacy of the data of our clients.

The first product predicts the **motor vehicle** fraud insurance claim and later we plan to expand our services to more products.

Brief Description

We wish to develop a **user friendly UI** which uses ML model to predict insurance claim fraud. We train our ML model using the previous insurance claim data provided by our clients.

The model is **trained and updated** regularly for better **accuracy and robustness**. Manual fraud detection requires trained individuals, consumes more time and is not cost effective.

Clients - We provide **B2B** services to **insurance companies** to prevent them from fraud insurance claims in minimum time and affordable prices.

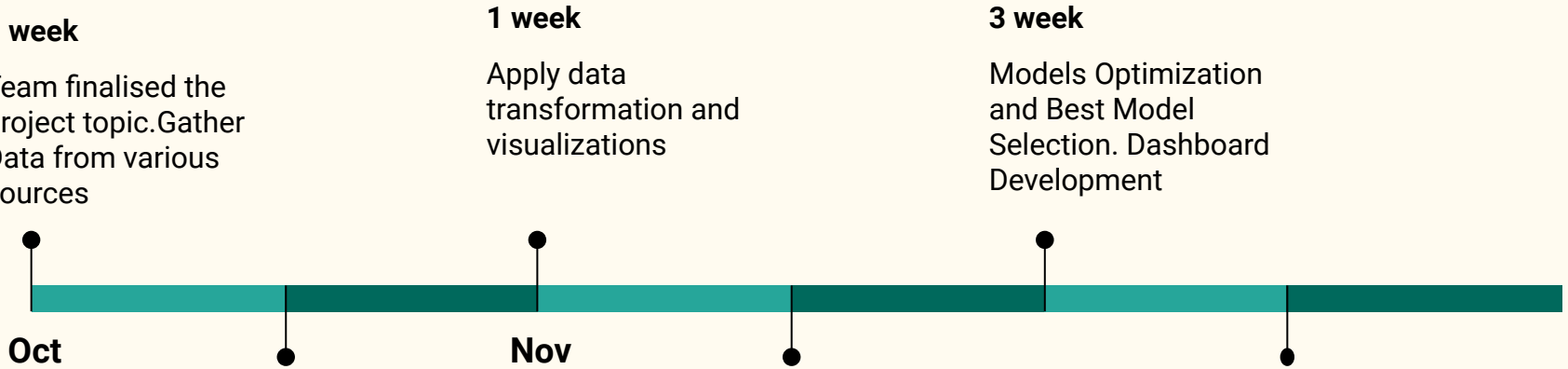
Problem Definition

Customer Requirement	A **cost-effective, accurate**, reliable and easy to use product which provides information whether the insurance claim is fraud or not and uses this pattern for future detections. Manual fraud detection requires **trained inspectors**, consumes more time and cost.
Market Survey	**FRISS** Fraud Detection at Claims, **INSURANALYTICS** (Claims AI Cloud), **BAE SYSTEMS**, Juicy-Score, Bridgei2i, LexisNexis, TransUnion
USP	"**Plug and play**" user-friendly interface, High accuracy (**>90%**) and specificity. Fast and reliable Insurance claim fraud analysis. Complete data privacy and security.
Protection of USP	User **authentication** protected UI with password stored in encrypted form. Patenting and copyrighting of our startup solutions.
Barrier to Entry	Patents and copyrights issue from other companies working in this domain. Data Acquisition
Business Case	• **Feasibility**: Model will be dynamic and mostly suitable for all the Bike, Car or Four wheeler insurance companies • **Impacts**: It will create a release for task intensive scanning of each Claim (as 80% cases are generally fraud) • **Benefits**: Highly scalable model, continuous support, upgradation, Never ending Demand etc

Technology Landscape Assessment

Patents	https://worldwide.espacenet.com/patent/search?q=insurance%20claim%20fraud%20detection&queryLang=en%3Ade%3Afr https://patents.justia.com/search?q=vehicle+insurance+fraud https://patents.google.com/patent/US20160117778/en https://worldwide.espacenet.com/patent/search?q=pn%3DCN106600423A
Published literature	https://www.sciencedirect.com/science/article/pii/S1877050919300079 https://www.sciencedirect.com/science/article/pii/S1574013721000423 https://www.sciencedirect.com/science/article/pii/S0167923617302130 https://ieeexplore.ieee.org/document/8074258 http://www.jcreview.com/fulltext/197-1583405087.pdf
Open libraries	Numpy, matplotlib, keras, Pandas, tensorflow, glob, sklearn
Proprietary libraries	https://github.com/saritmaitra/Fraud-detection--Insurance https://github.com/sharmaroshan/Fraud-Detection-in-Insurace-Claims https://github.com/mehtabhavin10/insurance_fraud_detection https://github.com/ezzaimsoufiane/Auto-Insurance-Claims-Fraud-Detection-with-ML

Project timeline



Strategy

Step 1 | Step 2 | Step 3 | Step 4

Data Collection, Data Cleaning & Filtering, Exploratory Data Analysis (EDA)

Obtain the dataset from the open sources and clean the data by assigning a valid value to not available values or dropping incomplete rows.

Data Transformation, Data Visualization, Feature selection

Analyze the relation between features using correlation matrix and the relevance of a feature to the prediction. Drop out unnecessary features (columns).

Modelling, Evaluation & Optimization, Best Model Selection

Develop ML models such as k-means clustering, logical regression and obtain a model that best fits the data and perform model evaluation using methods as presion, recall, and F1 score.

Dashboard develop -ment, Detailed Reports, Post processing

Design an authenticated user dashboard where the client can enter the claim data for which prediction is to be done and obtain the predicted results on the dashboard.

Problem Description

The dataset contains total 37 columns as given below:

1. **months_as_customer**: It denotes the number of months for which the customer is associated with the insurance company.
2. **age**: continuous. It denotes the age of the person.
3. **policy_number**: The policy number.
4. **policy_bind_date**: Start date of the policy.
5. **policy_state**: The state where the policy is registered.
6. **policy_csl**: How much of the bodily injury will be covered from the total damage.
7. **policy_deductable**: The amount paid out of pocket by the policy-holder before an insurance provider will pay any expenses.
8. **policy_annual_premium**: The yearly premium for the policy.
9. **umbrella_limit**: An umbrella insurance policy is extra liability insurance coverage that goes beyond the limits of the insured's homeowners, auto or watercraft insurance. It provides an additional layer of security to those who are at risk of being sued for damages to other people's property or injuries caused to others in an accident.
10. **insured_zip**: The zip code where the policy is registered.

11. **insured_sex**: It denotes the person's gender.
12. **insured_education_level:** The highest educational qualification of the policy-holder.
13. **insured_occupation:** The occupation of the policy-holder.
14. **insured_hobbies**: The hobbies of the policy-holder.
15. **insured_relationship:** Dependents on the policy-holder.
16. **capital-gain**: It denotes the monitory gains by the person.
17. **capital-loss:** It denotes the monitory loss by the person.
18. **incident_date:** The date when the incident happened.
19. **incident_type:** The type of the incident.
20. **collision_type:** The type of collision that took place.
21. **incident_severity:** The severity of the incident.
22. **authorities_contacted:** Which authority was contacted.
23. **incident_state:** The state in which the incident took place.
24. **incident_city:** The city in which the incident took place.
25. **incident_location:** The street in which the incident took place.
26. **incident_hour_of_the_day:** The time of the day when the incident took place.
27. **property_damage:** If any property damage was done.
28. **bodily_injuries:** Number of bodily injuries.
29. **Witnesses:** Number of witnesses present.

30. **police_report_available:** Is the police report available.
31. **total_claim_amount:** Total amount claimed by the customer.
32. **injury_claim:** Amount claimed for injury
33. **property_claim:** Amount claimed for property damage.
34. **vehicle_claim:** Amount claimed for vehicle damage.
35. **auto_make:** The manufacturer of the vehicle
36. **auto_model:** The model of the vehicle.
37. **auto_year:** The year of manufacture of the vehicle.

Using the dataset we predict whether the claim is fraud or valid, thus we use SVM and XGBoost classification model.
Output : Fraud_reported, Y or N (Y: claim fraud, N: claim valid)

Some columns are not required for classification such as 'policy_number', 'policy_bind_date', 'policy_state', 'insured_zip', 'incident_location', 'incident_date', 'incident_state', 'incident_city', 'insured_hobbies', 'auto_make', 'auto_model', 'auto_year', thus we drop these columns.

In total 26 columns have been used for prediction.