Predicting the "Fraud in auto insurance claims" & Pattern extraction

Problem Description

A major general insurance company has a business problem with significant number of claims being reported are fraudulent in nature and it is leading to leakages. So, the Insurer decided to predict the fraudulent ones before even processing the claims to allocate costs appropriately, to keep the thorough investigation process in place and to design proper action plan for the claims etc.

Insurance fraud refers to any claim with the intent to obtain an improper payment from an insurer. Motor and health insurance are the two prominent segments that have seen a spurt in fraud. Frauds can be classified from source and/or nature point of view.

Sources can be policyholder, intermediary and/or internal with the latter two being more critical from internal control framework point of view. Frauds can be classified into nature wise, for example, application, inflation, identity, fabrication, staged/contrived/induced accidents etc.

Fraud affects the lives of innocent people as well as the insurance industry and thus it may be of interest for the health of the Insurance Industry and Society. In fact, Insurers report certain classified cases to Regulator and Law enforcement agencies like Police, Crime Bureaus and others as mandated by the Regulators/Government and required by Law. With the advent of organised gangs and/or collusion, the problem has become more complex and sophisticated and the frauds have been difficult to detect and to prove, if detected.

The framework of prediction of fraud and pattern extraction wil be useful for the insurance companies, regulatory body, intelligence department etc.

Prediction at the time of processing claims will reduce costs and minimize losses.
The intelligence arising out of ever improving prediction algorithms will help retrofitting in terms of improvement of the underwriting process, exercise of good selection of policyholders based on identified profile attributes, strengthening of internal risk management mechanisms and finally, a clear guidance and communication to employees and other stakeholders involved.
At the Industry level, the shared aggregate information helps build appropriate intelligence and resilience while paving the way for collective effort for prevention as well as minimizing losses and to match the efforts of perpetrators.

At the Regulator and Law enforcement level, the intelligence arising out of prediction will help revamp the Regulations/Laws and plan not only enforcement but Industry based initiatives/systems for resilience and to share information for consumption of the Industry and the Society.
Prediction at the time of processing claims will reduce costs and minimize losses for the insurance company. Hence, prediction of fraud plays very important role in auto insurance claims. The company wants to understand the hidden patterns in the data which lead to construction of investigation process as well as claim settlement decision.

About Data: The data consists of:

1. Demographics Data : These files consist of the demographic data of each customer, like CustomerID, Cour

2. Policy Information : These files consist of the customer auto insurance policy information, connected to the

UmbrellaLimit, etc.

3. Data of Claim : These files consist of the details about the insurance claim, that the customer applied for, like Date

4. Data of Vehicle : These files consist of the details about the **Vehicle, connect**ed to the policy.

5. Fraud Data : This Train.csv contains the Fraud information details, like CustomerID, ReportedFraud. ; Target vari

Main Tasks:

1. Exploratory Data Analysis using visualizations in R Notebook or Jupiter notebook format . **(All train**
   a. What kind of insights did you find in the data after feature engineering?
   b. Learning curves : what is your observation based on the learning curves? Is there any bias or
   c. Based on the learning curves observation, which model do you think is suitable for the data and
2. You are expected to build a framework that predicts whether a claim is fraudulent or not ("Y" or "N"
   a. For this purpose you may use traditional approaches and deep learning techniques as well to im
3. **Patterns and suggestions to the company** : You are expected to extract top 20 patterns for fraudule company to reduce their claim processing costs .
4. Viva

Error Metrics

1. Consider “F1 statistic” for “Y” level of Target attribute as error metric and tune the model accordingly.
2. Consider appropriate evaluation metric for deciding the top 20 patterns for fraud on target attribute.

Important Note for the results submission:

Note: While evaluating the predictions submitted, the system will consider “1” as positive level in target attribute and hence please convert the target attribute accordingly and submit the results. It is very important for this problem as the error metic is “F1 statistic” for the target attribute level “Y”. Refer to the samplesubmission.csv file. ((1-”Y” & 0 - “N”))

Visualization Tips
Important Note: **No points for "effort" of putting a plot which is illegible and meaningless.** I for completing visualization exercise. **This will work against you.** Every plot you present should should convey a message. If the message is not evident, write that down clearly under the plot. It with several plots but present only the plots which are telling a story.
General Notes:
Clearly understand difference between following plots 1. Bar plot, 2. Box plot, 3. Histogram, 4. Scatter plot 5. Line plot Decide which one you need before you draw. Most of the technical visualization can be covered b need anything beyond these, think again do you really need it or you are over complicating it?
Readable legends, axis labels, plot titles, tick sizes, are essential of any plot. If your audience cann serves no purpose at all.
Default plotting commands in Python and R generate barebone plot with poor labels and font size would have to write additional lines of codes to make the plot readable with figure and axis option using matplotlib along with seaborn. Some functionalities are stronger in seaborn library while so
Do not blame the tool for giving a bad plot. Tool will give exactly the plot which you ask for and you ask. You just need to know how to ask!
Do not try to present too much information or too many variables in single plot. It is rarely useful for technical visualization are very poor choice as they are very hard to read.

Chose the color (and need for it) wisely. Several powerful plots can be presented just by using bl
is no additional points for making your chart colorful. If you have colors, it better carries a meani
confusing

Some common mistakes seen in INSOFE presentations (other than poor font sizes and axis labels)
1. Box plot for describing frequency or count (Hint: Box plots do not give count. They only give s
2. Boxplot of different columns/variables with different scale on same axis. (Hint: If you cannot
in your plot, then you are probably doing something wrong)
3. Not thinking through if the actual number or percentage of total would be better choice of bar
right answer but think about the story you want to tell. Think which representation would exactly
tell)
4. Overuse of colors when instead a simple B&W scale would have done the job.
5. 3D plots where 2D plot would tell exactly the same story. (3D plot with different colors would

Some great examples of extremely bad visualizations are here: http://livingqlikview.com/the-9-w
ever-created/
Go over these and think why your visualizations are better

Business Case

State the business problem (do not copy paste from the question. Instead write based on your understanding). Explain from business perspective why they are interested in this problem

ML problem statement

Describe how the business problem is converted into a Machine Learning problem and how it will help the business

Data Exploration

Describe what data you have and give a snapshot of rows and columns and describe some important parameters.
If you have merged some datafiles, briefly talk about it but do not exaggerate on merging procedure.
Explain your data and exploratory efforts with visualization (follow the instructions on visualization)

Data Exploration and Visualization

This can largely come from the section above but note that your previous visualizations were without the target variable. You may want to include additional plots now that you have complete visibility on the problem.

Models

Explain choice of your models and objective reasoning behind choosing. Model simplicity and computational time is as important as model accuracy. Hence do not jump on computational intense models first. Make sure increase in computational efforts is coming with proportional benefit in accuracy. If an advanced model is not benefiting in the accuracy, then simpler models are usually better choice. Your modelling effort should logically indicate this trade-off between accuracy and speed. (Both R and Python have ability to compute execution time of chunk of your code i.e. just your model)

Validation and Parameter Tuning

Explain your validation/cross-validation process in brief detail. What data you used for validation and what is your reasoning behind validation effort. How is your validation effort helping the model? What parameters you updated and why? If you did k-fold cross validation, why did you chose so? Did you really benefit from advanced validation methods?

Summary & Conclusion

Overview of your journey (all above sections) and what did you finally conclude. What would be your recommendation for business? What more you could have done?. (Do not mentioned "Could not do due to lack of time". Everyone is on same time scale. We understand your time constraint so there is no need to explicitly mention it)

Appendices

Please attach complete source code, with proper commenting and indenting.