1. **Baseline**

Currently, the firm aims to provide insurance to 5% of new homes, and it can expect to earn a profit of 30% of the insurance premium (AMTI) multiplied by 5%, plus the avoidable loss on market share. In the base case, the estimated yearly profit is $177,281.76. Based on the analysis, it is recommended to choose the Extra Trees Classifier model by TPOT with a prediction threshold of 0.31. The setting of threshold is motivated by the point where the profit curve reaches its maximum. This is expected to yield a yearly profit of $432,055, which is significantly higher than the base case.

2.**Action Intervention**

To optimize profit, we evaluated DataRobot and Python predictions on a holdout set and identified the threshold that generated maximum profit using the Extra Trees Classifier model. However, this threshold may lead to missed customers and increased false negatives, so Rockafella should balance profit maximization with risk mitigation. Setting a policy to limit false positives and targeting customers with high premium potential are possible approaches. Also, Rockafella could validate customer information before issuing insurance policies.

3. Data management procedures

On data exploration, data quality and preprocessing

Overall data quality is worrisome, and requires a high quality feature engineering. Using Pandas Profiling one can tell, there are inconsistencies and many features missing data or having negative values. Also the amount of redundant features are concerning and needs to be meticulously investigated. The target (BUYI) is extremely imbalanced, further exploration of sampling routines is recommended. Class 1: (49814 / 51808) = 96.15% Class 0: (1994 / 51808) = 3.85%

Outlier removal and numerical transformations

Amount of negative values, replaced by the median for the following features: UNITSF, LOT. LOT also contains irrational extremes at the higher ranges, these also replaced with the median, however for LOT the zeros could represent that the house is a flat and therefore they are kept in this case. In the HHAGE feature, there were some extreme values observed, such as the age of 93 with a high frequency of 0.9%. While these cases were not numerous, they suggest that the surveyors may have opted for the highest age rather than specifying an accurate

value. Additionally, values below 18 were deemed inappropriate. As a solution, the extreme values were replaced with the median age. Originally the dataset had 1=yes, 2=No, in some cases 3=No categorical variables. To make it logical we replaced them everywhere with dummies yes=1, no=0. The dataset also captured survey responses which were negative values (-6, -7, -8, -9) since these inputs referring to “Blank / Not reported / Refused / Don’t know, they were replaced by zeros.

(i)1 Feature creation

To capture the potential impact of the quality of the house, features, as EVROD, EROACH, CRACKS, HOLES were recoded 0=No, 1=yes, dummies. The motivation behind this was to create a feature that could indicate the bad quality of the house. BAD_SHAPE is calculated as the average of other negative features. *df[‘BAD_SHAPE'] = df[['EVROD', ‘EROACH', ‘CRACKS’,'HOLES']].mean(axis=1)*

AVG_TOTAL_SQR feature was created to represent the average size related to the property, and is calculated on the basis of LOT and UNITSF. ISFLAT =1 if LOT =0, otherwise 0. The idea of this feature was to capture whether the house is flat or not, which could potentially hold value.

(i)2 Features were excluded due to being redundant, followed by reason:

ZINC2 Household income is 0.97 correlated with family income, considered to be redundant. CELLAR and MOBILITY are also redundant because they lack predictive value. MOBILTYP has many missing values and may not align with Rockafella's insurance policy.. Thus, it is considered to be redundant. METRO3 only has two categories instead of five when comparing with the AHS codebook, and WINTERNONE is related to whether a home has warming, seems no significance. FRSTOC - One can be 3rd or 5th owner and still buy insurance, considered to be redundant. CLIMB - It is difficult to extract any information for this feature and therefore it is considered to be redundant.

These features show insignificance and therefore some of them had many missing values therefore removed from the table.

Hot encoding

REGION was one-hot encoded using Python to determine if the location of the house has any predictive value.

Dealing with excess zeros, missing values

To handle missing values and excess zeros, we either removed the feature entirely or replaced outliers with either the median or zeros, depending on the specific feature and the number of missing values.

NOTE

BUILT year is important to be mentioned since it is time series. Although we are not predicting house price, therefore it could be left in the table. Otherwise one needs to be careful about this before creating features. The feature VALUE shows some correlation with AMTI. We Assume that the value of the house is available at the time of the prediction. Hence, we can keep it in the table.

Leakage

AMTI - We excluded from the training set to prevent data leakage since we aim to predict insurance purchase. However, we stored it separately for calculating profits. Negative values in AMTI were present due to missing or unanswered responses in the survey. To avoid any impact on our profit calculations, we transformed negative values to zero.

c. Employed data partitioning:

To handle the imbalanced dataset, we split the data into 50% for training, and 25% each for validation and testing. This approach allows for more robust model tuning through hyper parameter optimization, and a larger dataset for testing the model's performance

Qualitative ProfitComponents

		Predicted	
		Not Purchased	Purchased
Actual	Not Purchased	Avoidable labor cost of insurance offer + Avoidable reduction in market share	Avoidable reduction in market share
	Purchased		Preventable loss of profit margin + Avoidable reduction in market share

Quantitatively:

Avoidable_labor_cost = |BUYI - 1| * 500

Preventable_Loss_of_profit_margin = BUYI * AMTI * 30% * 5%

Avoidable_market_share_reduction = BUYI * 200

NB : |abs| stands for absolute values.

Justification

In case of (TP) we manage to get the preventable amount of insurance loss on the 30% profit margin of AMTI. Although, we have to keep in mind that we can only target the subset of 5%. Additionally we also get the avoidable reduction in market share since we made correct predictions (we don’t predict it falsely).

However, if we falsely predict that the customer would buy but wouldn't, (FP) then we still get the avoidable reduction in market share since our agents prepared the offers, but we loose the avoidable labor cost and the preventable loss on profit margin. If we predict correctly that insurance is not purchased (TN) we avoid the labor cost that it takes to prepare the offer, at the same time we don't get the preventable loss on profit margin. Although, assume we do still get the avoidable reduction in market share, since we did not fail to prepare the offer.

4.a Model Selection

Two methods was carried out for model selection, TPOT in Python and DataRobot. LGBT with Early Stopping and from TPOT Extra Trees Classifer. Given the high imbalance in the dataset, downsampling the majority class was a practical option, and a sampling strategy of 0.85 was used, stratified=y (random state=77). To get a more throughout overview, for DR we made synthetic observation with SMOTE (50-50) and the models were trained on the resampled data. Predictions were made on the holdout data, exported from datarobot and compared with TPOT. To make the calculations we added AMTI back along with the calculations "Avoidable labor costs", "Preventable loss of profit margin" and "Avoidable reduction in market share" and proceeded with the valuations for both models. NB these features must be excluded before splitting the data otherwise would have made the predictions contaminated.The expected value of someone buys insurance, depends on whether they actually buy insurance. Thus, we based our calculation on this and first calculated the value if we predict that someone buys insurance, and also if not. Correspondingly calculated the value for DR and for TPOT based on their prediction labels. DR_VALUE: $323,685.66

TPOT_VALUE: $339,689.43

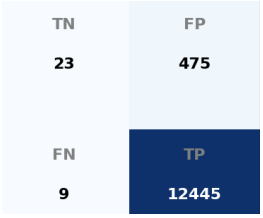
Value_Predic_Buy = IF(BUYI=1,(AMTI*0.3*0.05)+200, 200)

Value_Predict_Not_Buy = IF(BUYI= 0, 500+200, 0)

On decision thresholds

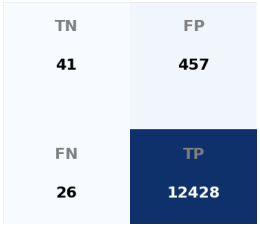
We found that a probability threshold of 0.31 for TPOT would maximize revenue, resulting in a maximum value of $432,055.98. However one must keep in mind that we would miss out potential customers who would have purchased insurance.

On hyperparameter tuning

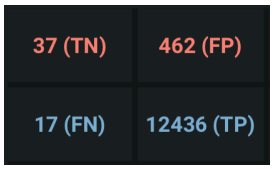


Random forest classifier given by TPOT returned the following results on the validation set.

It seemed quite high to have 26 false negatives, and I assume that in reality it would be most costly for the firm by not offering insurance. Thus, we used grid search with cross validation to fine tune the hyper parameters for the pipeline.



Max_debth=7, cv=5, Best Validations score:0.96, Threshold:0.45 We retrained the model and evaluated on the test set, given a better outcome on the reduction of the FN from 26 to 9 but increased the FP to 475. While it is plausible to reduce the FN the moderate amount of FP is concerning. Meanwhile for Data Robot used 50-50 SMOTE and specified cross validation with 5 folds.



DR identified LGBT with Early Stopping to be the best model. Testing the data on the test set, with decision threshold 0.1012 the model performed similarly compared to Python. DR has an F1 score of 0.9811 which indicates that it performs relatively well on unseen data. Looking at the accuracy would not be optimal due to the imbalance of the dataset.

Potential Actions

Given that the FP rate is a quite high mainly due to the data is extremely imbalanced a potential action could be for Rockafella to consider prioritizing its customer base and offering insurance where the premium is highest. Also assuming that the company has some limited resources to prepare insurance offers. It could be a potential target to limit the number of false positives as well, however that would increase the FN rate. An important intervention could also be to let the agent contacting with the customer before preparing the insurance. This would help Rockafella to mitigate the risks related to FP.

Significant Features: VALUE, BUILT, BAD_SHAPE, ZINC Property value, house age, overall quality, and family income are all important factors in determining whether someone buys insurance. Higher property values and older houses with risks might lead to a higher need for insurance. Inadequate house condition may also lead to increased seek after insurance.Whereas higher family incomes could also lead to more insurance purchases.