greatlearning
Power Ahead

Capstone Presentation

Tushar Babbar
PGP DSBA DEC'20

Business Problem Understanding

Problem Statement: Life Insurance Data

- The dataset belongs to a leading life insurance company.
- The company wants to predict the bonus for its agents so that it may design appropriate engagement activity for their high performing agents and upskill programs for low performing agents.

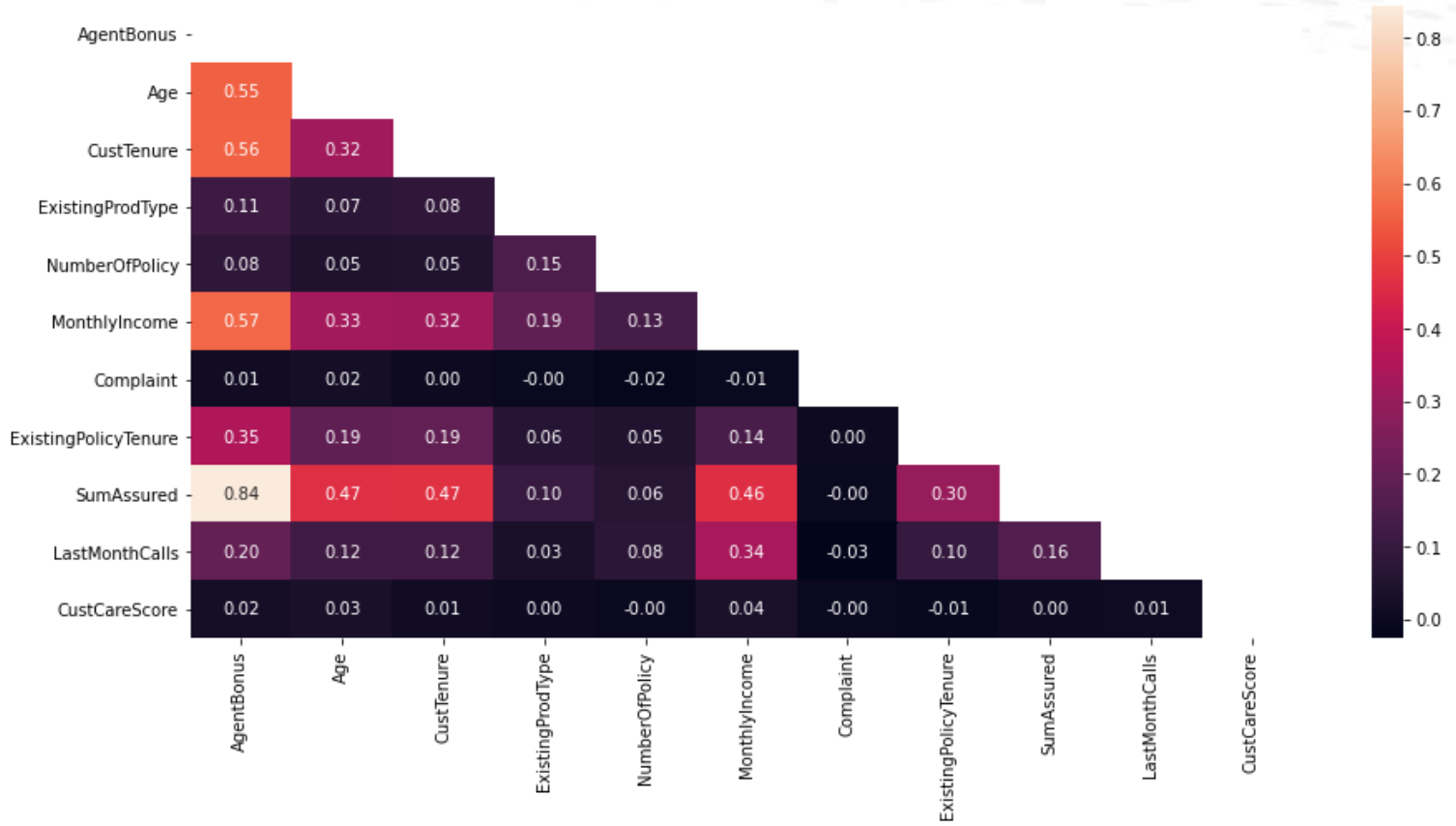
Need for this Study/Project

- With this problem we want to better understand how the insurance company agents are performing, it's not to underpay or overpay, as the payment is regulated by IRDA.
- With the predictions it's better for the company to understand where they need to focus more as for agents selling less policies the company needs some booster training performs. As the policies are as good as the agents portray it to be to the potential customer.
- While the agents performing good i.e. selling more policies there needs to be a way to reward them, to make their contribution known so that they perform the same and even better in future.

Why is this (agent bonus) important for the business/company?

- A company is as good as their employers.
- For a Life Insurance Company, their agents are the best way to make the companies policies, aims, and perks known to the customer. Once the customer is intrigued by the policy delivery by the agent, its easier to convince the customer hence improving the sales and thereby motivating the agent as well.
- With this, the market share of the company will gain more ground dominating the potential opponents.
- Moreover, the agents can be classified into categories giving the company better insight where the need to put more effort.
- The customer feedback can help the company develop improved and updated policies/products. Meeting customer needs.
- Hereby, the easiest way to retain their agents.
- Overall, multiplying and adding to company's profit.

Understanding Variable Correlation



- Complaint and CustCareScore have almost no correlation with any other parameter, hence dropping these columns will not make a difference.
- AgentBonus and SumAssured have high correlation with each other of 0.84.
- Here the lighter colors depict high correlation and darker colors depict low correlation

Modelling Approach Used & Why

- Modelling approach used here is Linear Regression, which is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

- The other models tested and compared alongside Linear Regression were Decision Tree Regressor, Random Forest Regressor and Artificial Neural Network (ANN) Regressor.

- Model Outputs (Without Model Tuning):

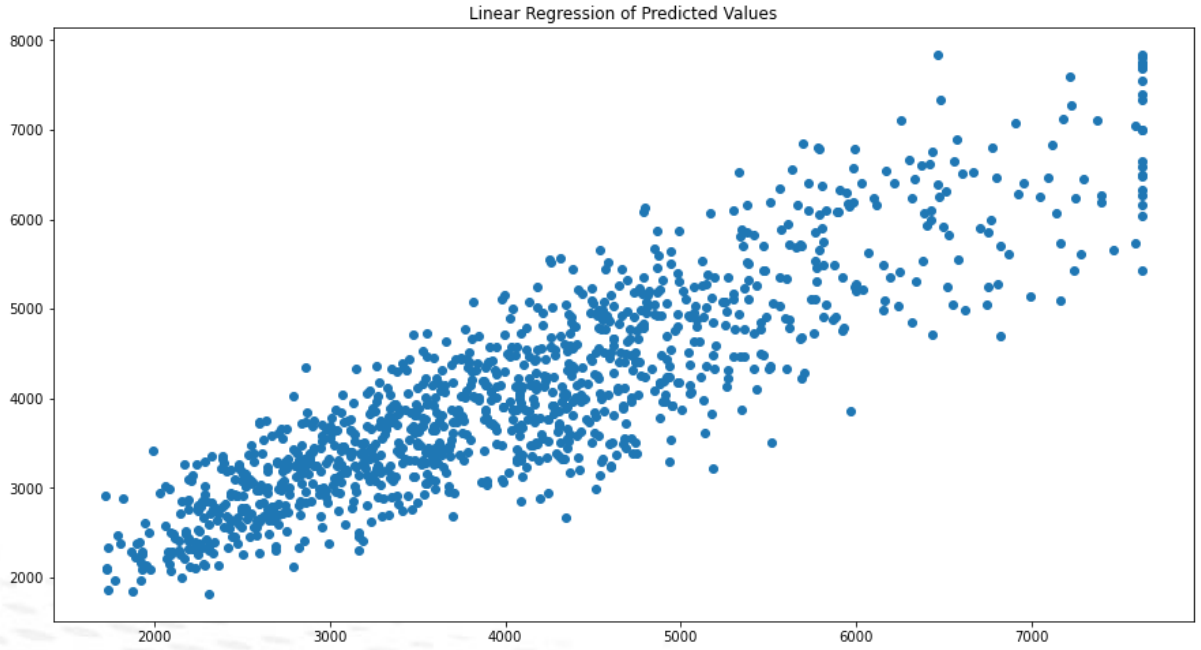
	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	612.550689	585.514819	0.800806	0.801482
Decision Tree Regressor	0.000000	725.006753	1.000000	0.695626
Random Forest Regressor	189.614010	519.044211	0.980913	0.843997
ANN Regressor	225.889011	701.144120	0.972912	0.715332

- Model Outputs (With Model Tuning):

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	612.550689	585.514819	0.800806	0.801482
Decision Tree Regressor	495.463438	569.694730	0.869679	0.812065
Random Forest Regressor	527.410585	572.885614	0.852331	0.809954
ANN Regressor	28.117642	670.444991	0.999580	0.739715

- Initial LR results

	R-Squared	RMSE
Training	0.8068152802160813	600.5900784990952
Testing	0.7825646087670782	621.5274260080358



- Here, R-squared (R^2) is a statistical measure that **represents the proportion of the variance for a dependent variable** that's explained by an independent variable or variables in a regression model.

Model Selection

- From the previous results, it is evident that Linear Regression is a better model.

- Why Linear Regression?
  - Post removal of variables causing multicollinearity, Linear Regression provided a good R-squared value and similarly a high adjusted R squared value. Hence a good percentage of variance can be successfully explained by our model.
  - A very important factor being the train and test set accuracy scores are ~80% and consistent.
  - Unlike other models where overfitting and inconsistency in the performance metrics can be observed. Linear Regression model does not show these inconsistencies in the observation.

  (Here by overfitting we mean, the model is performing very good for training set and giving poor results for the testing set)
  - The LR model makes it easier to understand the model, multicollinearity in the data. Also, unlike other model its computational time is quick therefore we can run it multiple times whereas ANN and Random Forests needs capable machines as they are very time consuming models. Might have to wait for hours and in our case they still don't perform better than LR.

Note: 100 % accuracy cannot be achieved in real life data as there is always some unexplainable factors and noise that's always present in our data.

Model Evaluation

The Equation

(1092.35) * Intercept + (21.65) * Age + (22.62) * CustTenure + (46.51) * ExistingProdType + **(6.25) * NumberOfPolicy** + **(0.03) * MonthlyIncome** + (33.05) * Complaint + (40.23) * ExistingPolicyTenure + **(0.0) * SumAssured** + **(-2.31) * LastMonthCalls** + **(7.56) * CustCareScore** + (22.69) * Channel_Online + **(3.5) * Channel_Third_Party_Partner** + (-616.86) * Occupation_Large_Business + (-474.97) * Occupation_Salaried + (-581.64) * Occupation_Small_Business + (26.68) * EducationField_Engineer + (-177.27) * EducationField_MBA + (-92.61) * EducationField_Post_Graduate + **(2.33) * EducationField_Under_Graduate** + (25.19) * Gender_Male + (-493.36) * Designation_Executive + (-481.42) * Designation_Manager + (-277.42) * Designation_Senior_Manager + **(-2.96) * Designation_VP** + (-48.2) * MaritalStatus_Married + (29.66) * MaritalStatus_Single + (-188.88) * MaritalStatus_Unmarried + (62.35) * Zone_North + (193.51) * Zone_South + (50.0) * Zone_West + (141.95) * PaymentMethod_Monthly + (112.03) * PaymentMethod_Quarterly + (-79.92) * PaymentMethod_Yearly

- From the equation the variables with a low or no coefficient value depicts that the variable is very important to the independent variable's prediction. As the coefficients value increase it shows the variable has become comparatively less significant.

- The variable significance can be explained using the * method, where * depicts highly significant, ** less significant, and *** and **** least significant.

Variables	Significance
SumAssured, MonthlyIncome	*
LastMonthCalls, CustCareScore, Channel_Third_Party_Partner, EducationField_Under_Graduate, Designation_VP, NumberOfPolicy	**
Age, CustTenure, Channel_Online, EducationField_Engineer, Gender_Male, MaritalStatus_Single, Complaint, ExistingPolicyTenure, MaritalStatus_Married, Zone_West, Zone_North, PaymentMethod_Yearly, EducationField_Post_Graduate	***
Occupation_Large_Business, Occupation_Salaried, Occupation_Small_Business, EducationField_MBA, Designation_Executive, Designation_Manager, Designation_Senior_Manager, MaritalStatus_Unmarried, Zone_South, Paymentmethod_Monthly, PaymentMethod_Quaterly	****

- R-Squared Obtained from final Linear Regression Model: 0.806
- Adjusted R-Squared Obtained from final Linear Regression Model: 0.805

- Decision Trees, Random Forest, and ANN (Before Hyperparameter Tuning) :
  - It can be observed that all the 3 models have overfitting problems where we have ideal accuracies of ~100% for our training set. However the models are performing poorly on our testing set having accuracies ~70% – 84%. There is a major accuracy difference between the training and testing set which is not acceptable for predictions.
  - If the accuracy difference is greater than 6-10% it is advised to not accept the model as the predictions can be unreliable.

- Decision Trees, Random Forest, and ANN (After Hyperparameter Tuning) :
  - After Hyperparameter Tuning Decision Trees and Random Forest models showed no overfitting errors.
  - The training accuracies were ~85% and testing accuracies were ~80%.
  - ANN still showed no improvement in results and was still overfitting.

- Although the Decision Trees and Random Forest were performing good, I went with Linear Regression as it gave more stable results and Variable importance could be calculated more easily from the Linear Regression Equation and stats-model performed to predict the results.

Insights from Analysis

- Company wants to predict the ideal bonus and what is the engagement for high and low performing agents respectively.
- From the model, the high performing agent we will find variable significance, for eg, Sum Assured is highly significant here and highly correlated to our target variable.
- SumAssured is highly significant as the agent performing good is the one which is getting more profit for the company selling more or high value policies.
- If the Designation is VP the person buys more policy or high value policies.
- Therefore, for high and low performing agents, we will train them, suggesting them to purchase or get policies with high sum assured as it is very significant to our model.
- Another important feature is Customer tenure where the agents need to focus on the customers who've a tenure ranging between 8-20 this where the majority of the customer are.
- Focusing on customers with greater monthly incomes as greater the monthly income, greater is the possibility of the customer buying a higher valued policy.
- From the Linear Regression Equation we can find insights and remove all the least significant variables

Recommendations

- For High Performing Agents we can create a healthy contest with a threshold.
- Where, if they achieve the desired sum assured, they are eligible for certain incentives like latest gadgets, exotic family vacation packages and some extra perks as well.
- For low performing agents, we can introduce certain feedback upskill programs to train them into closing higher sum assured policies, reaching certain people to ultimately becoming top/high performers.
- Apart from this, we need more data/predictors like Premium Amount, this will help us to solve the business problem even better as well have more variables to test upon thereby having more accurate results in real time problems like this.
- I also feel another predictor can be added as customers geographical location or Region and not just the zones as people living in rural areas are less likely to buy a policy whereas those living in a highly developed location are likely to be belonging to the upper class and should be targeted.
- Similarly, another predictor can be AgentID can be introduced which will make it easier to observe the high and low performing agent trend.