**Supervised Learning Insights**

We have used supervised learning to identify the general characteristics of policyholders for each product type. This allows us to get a general idea of what a policyholder who buys each policy looks like. Then we use unsupervised learning to split the data into clusters and understanding what products are well suited together. This allows us to compare the characteristics of NEOS Life’s policyholders and see the differences to a) General characteristics of each product b) Against major competitors. This is done by comparing various individual variables across underwriters, market share analysis and market penetration analysis.

**Logistic Modelling**

Not using because only good for some insurance types. Additionally, better alternatives available that capture non linear patterns and the complex relationships between numeric and categorical variables.

**1. Life Insurance**

* **Accuracy**: 86.94%
* **Sensitivity (Recall)**: 86.78%
* **Specificity**: 87.33%
* **Positive Predictive Value (Precision)**: 94.34%
* **Negative Predictive Value**: 73.06%
* **Balanced Accuracy**: 87.05%
* **AUC-ROC**: 0.8705
* **Insights**: The model for life insurance shows excellent performance across all metrics, with particularly high precision, indicating strong predictive capability and reliability in identifying potential purchasers.

**2. Trauma Insurance**

* **Accuracy**: 72.79%
* **Sensitivity**: 69.19%
* **Specificity**: 75.86%
* **Positive Predictive Value**: 71.00%
* **Negative Predictive Value**: 74.25%
* **Balanced Accuracy**: 72.53%
* **AUC-ROC**: 0.7253
* **Insights**: The trauma insurance model has moderate accuracy and balanced accuracy, indicating some challenges in prediction, which might be improved by more complex modelling or feature engineering.

**3. Total Permanent Disability (TPD) Insurance**

* **Accuracy**: 85.04%
* **Sensitivity**: 88.70%
* **Specificity**: 78.78%
* **Positive Predictive Value**: 87.71%
* **Negative Predictive Value**: 80.33%
* **Balanced Accuracy**: 83.74%
* **AUC-ROC**: 0.8374
* **Insights**: The TPD model performs very well, particularly in terms of sensitivity and positive predictive value, making it highly effective for targeting potential customers.

**4. Income Protection (IP) Insurance**

* **Accuracy**: 76.18%
* **Sensitivity**: 76.38%
* **Specificity**: 76.00%
* **Positive Predictive Value**: 74.45%
* **Negative Predictive Value**: 77.84%
* **Balanced Accuracy**: 76.19%
* **AUC-ROC**: 0.7619
* **Insights**: The IP insurance model shows good accuracy and balanced accuracy but indicates potential areas for improvement in model sophistication or feature utilization.

**General Recommendations:**

* **Model Improvements**: Trauma and IP insurance models could benefit from advanced modelling techniques such as Random Forests or Gradient Boosting to handle possible non-linear relationships and complex interactions better.
* **Feature Engineering**: Across all models, consider examining and possibly enhancing the feature set, including creating interaction terms or incorporating more contextual data.
* **Model Deployment**: The models for Life and TPD insurance are strong candidates for deployment given their high accuracy and predictive values. Efforts could focus on refining the Trauma and IP models to match or exceed these performances.

**KNN Modelling**

Not using because even though the model itself performs super well, it does not provide any insights into variable importance that would help us in providing the necessary recommendations to NEOS Life. Additionally, it is a very computationally expensive modelling process.

**Summary of kNN Model Performance:**

1. **Life Insurance:**
   * **Accuracy**: 93.3%
   * **Sensitivity**: 95.09%
   * **Specificity**: 88.93%
   * **AUC-ROC**: 0.9201
   * High performance across metrics, especially in sensitivity and precision.
2. **Trauma Insurance:**
   * **Accuracy**: 87.7%
   * **Sensitivity**: 86.92%
   * **Specificity**: 88.36%
   * **AUC-ROC**: 0.8764
   * Good balance in accuracy and sensitivity, slightly lower than Life insurance.
3. **Total Permanent Disability (TPD) Insurance:**
   * **Accuracy**: 90.81%
   * **Sensitivity**: 92.92%
   * **Specificity**: 87.21%
   * **AUC-ROC**: 0.9006
   * Very high sensitivity, excellent for targeting potential customers.
4. **Income Protection (IP) Insurance:**
   * **Accuracy**: 88.38%
   * **Sensitivity**: 88.01%
   * **Specificity**: 88.71%
   * **AUC-ROC**: 0.8836
   * Strong balanced accuracy, slightly better specificity compared to Trauma.

**Rationale for Considering Alternatives Despite Strong Performance:**

While kNN models demonstrate high accuracy and effectiveness in prediction across various insurance products, they have significant limitations that might prompt the use of other methods such as Random Forests:

1. **Lack of Feature Importance Insights**:
   * **Key Limitation**: kNN does not provide any information on which features are most influential in making predictions. This is a critical drawback for strategic decision-making where understanding what drives insurance purchases is necessary.
   * **Impact**: Without insights into feature importance, it's difficult to tailor marketing strategies, product design, or customer engagement based on the most impactful variables.
2. **Computational Efficiency**:
   * **Processing Time and Resources**: kNN can be computationally expensive, especially with large datasets, because it involves calculating the distance between instances for each prediction.
   * **Scalability Concerns**: This computational demand increases with the size of the data, which can make kNN less practical for real-time predictions or very large datasets.
3. **Model Transparency and Actionability**:
   * **Black Box Nature**: While not as opaque as deep learning models, kNN still offers little transparency regarding how decisions are made, unlike models that provide coefficients or decision rules.
   * **Strategic Implementation**: For insurance companies, understanding the reasoning behind model predictions can be as important as the predictions themselves, especially for aligning with business strategies and regulatory requirements.

**Conclusion:**

Given these considerations, **Random Forests** are a compelling alternative. They not only provide robust predictive accuracy often comparable to or better than kNN but also offer detailed insights into which variables are most important in predicting outcomes. This feature makes Random Forests particularly valuable for applications where strategic insights into data are crucial, such as in developing new insurance products, targeting potential customers more effectively, or complying with industry regulations that require explanation of decision-making processes.

Transitioning to Random Forests or similar models would allows us to retain high predictive performance while gaining critical insights into the data, thus enhancing both the actionability of the analytics and the strategic value of our modelling efforts.

**Naive Bayes**

Similar reasons to the kNN model

**Naive Bayes: Key Characteristics**

1. **Model Simplicity**: Naive Bayes is based on Bayes' Theorem and assumes that predictors are independent of each other given the response variable. This simplicity can be advantageous for quick modeling and when the assumption of independence roughly holds.
2. **Probabilistic Framework**: It provides probabilities of outcomes, which can be very useful for understanding the likelihood of different classifications.

**Limitations of Naive Bayes in Context:**

1. **Lack of Feature Importance**:
   * **Insight into Drivers**: Naive Bayes does not inherently provide insights into the relative importance of features. It treats all features as independently contributing to the outcome, which doesn't help in understanding which features are more influential than others.
   * **Strategic Decision Making**: Like kNN, the lack of information on feature importance with Naive Bayes makes it less useful for strategic decision-making where knowing the key drivers can influence product adjustments, marketing strategies, or risk assessments.
2. **Independence Assumption**:
   * **Model Accuracy and Realism**: The assumption that all predictors are independent given the outcome can lead to less accurate predictions when this condition is not met, as it often isn't in real-world data where features can be correlated.
3. **Interpretability**:
   * **Model Understanding**: While Naive Bayes is generally more interpretable than some more complex models like neural networks, it still does not provide the kind of intuitive, decision-rule-based explanations that methods like decision trees or Random Forests offer.

**Comparing to Random Forests:**

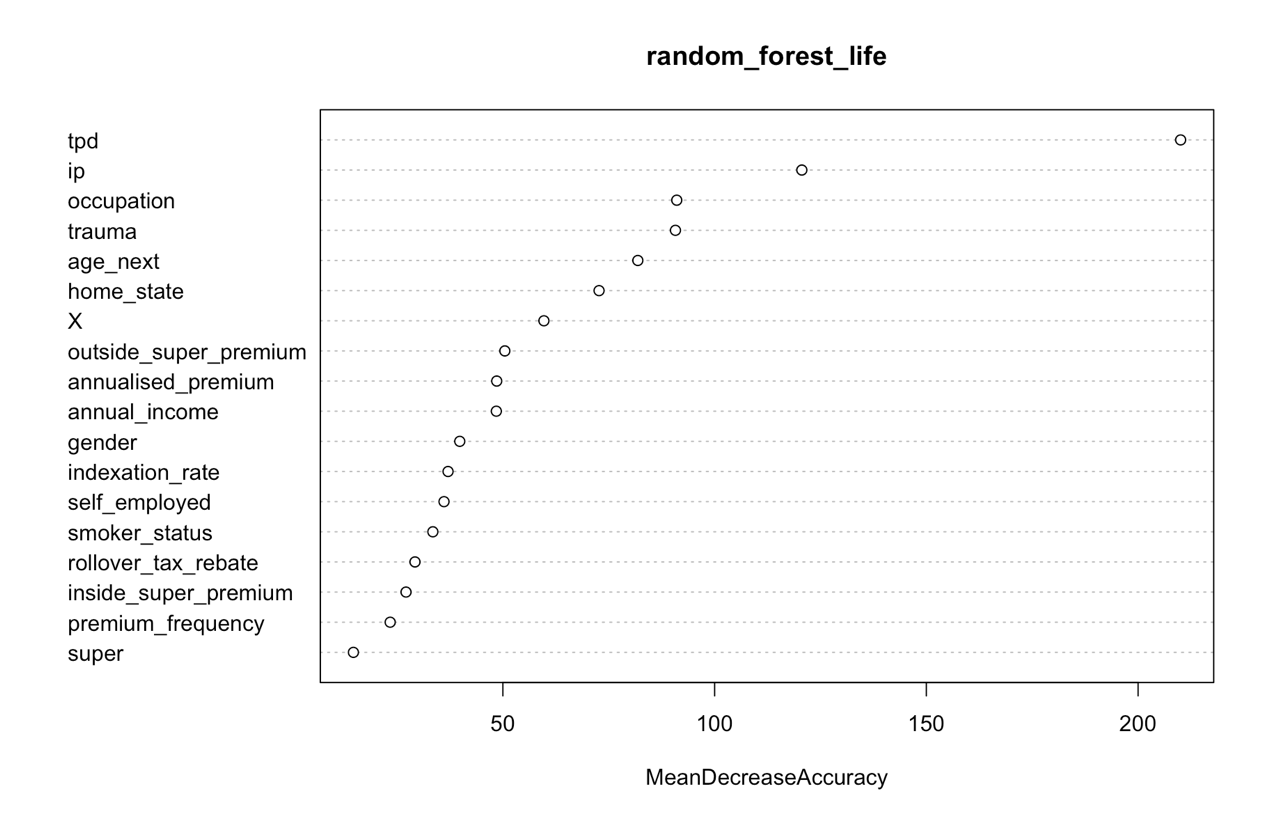
* **Random Forests** not only handle the correlation between features effectively but also provide outputs like feature importance scores, which are invaluable for interpreting the influence of each predictor on the decision-making process. This makes Random Forests particularly effective for detailed data analysis and strategic planning.
* **Model Flexibility and Robustness**: Random Forests are less sensitive to the distribution of the data and can model complex interactions between variables, which are often present in real-world scenarios.

**Conclusion:**

While Naive Bayes can be quick and effective for certain types of classification tasks, particularly when the predictors are indeed independent, its limitations make it less suitable than Random Forests for applications where understanding the influence of different features is crucial. Random Forests provide both high predictive accuracy and valuable insights into which factors most strongly influence the outcomes, thereby supporting more informed business decisions and strategic planning. If deeper insights into data and more nuanced modeling of relationships are needed, Random Forests or similar models would typically be a better choice.

**Random Forests**

**1. Life Insurance**

* **Accuracy**: 95.21%
* **Sensitivity**: 95.82%
* **Specificity**: 93.73%
* **Positive Predictive Value**: 97.38%
* **Negative Predictive Value**: 90.21%
* **Balanced Accuracy**: 94.78%
* **AUC-ROC**: 94.78
* **Top Variables**: TPD, IP, Trauma, Age, Occupation.
* The model demonstrates exceptional predictive accuracy and reliability, effectively distinguishing between purchasers and non-purchasers of life insurance.
* Variable importance chart

**2. Trauma Insurance**

* **Accuracy**: 92.89%
* **Sensitivity**: 96.20%
* **Specificity**: 90.06%
* **Positive Predictive Value**: 89.21%
* **Negative Predictive Value**: 96.52%
* **Balanced Accuracy**: 93.13%
* **AUC-ROC**: 93.13
* **Top Variables**: IP, Life, Age, Outside Super Premium, TPD.
* This model shows strong performance across all metrics, with high sensitivity and balanced accuracy, making it highly effective for identifying potential trauma insurance customers.
* Variable Importance ChartA graph with black lines

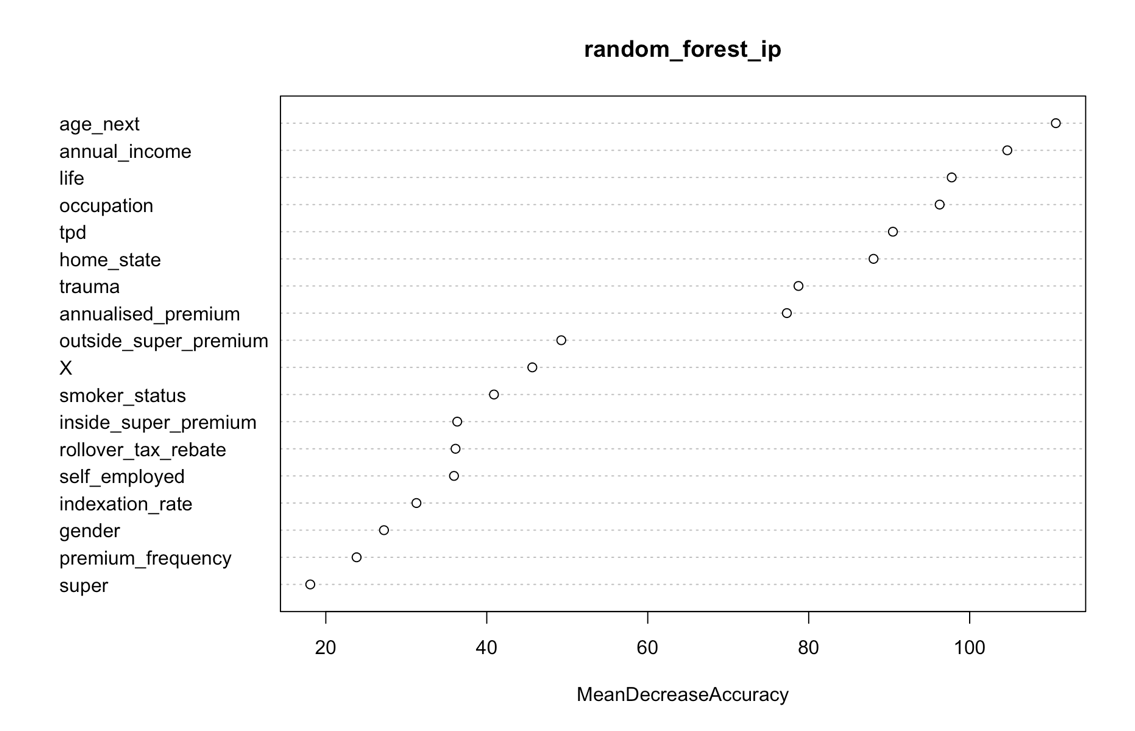
  Description automatically generated

**3. Total Permanent Disability (TPD) Insurance**

* **Accuracy**: 93.44%
* **Sensitivity**: 94.78%
* **Specificity**: 91.16%
* **Positive Predictive Value**: 94.82%
* **Negative Predictive Value**: 91.10%
* **Balanced Accuracy**: 92.97%
* **AUC-ROC**: 92.97
* **Top Variables**: Life, Age, IP, Occupation, Home State.
* The TPD insurance model excels at predicting purchasers with high accuracy, demonstrating strong sensitivity and specificity, suitable for targeted interventions.
* Variable Importance ChartA graph with different numbers

  Description automatically generated with medium confidence

**4. Income Protection (IP) Insurance**

* **Accuracy**: 92.88%
* **Sensitivity**: 93.38%
* **Specificity**: 92.43%
* **Positive Predictive Value**: 91.87%
* **Negative Predictive Value**: 93.84%
* **Balanced Accuracy**: 92.90%
* **AUC-ROC**: 92.90
* **Top Variables**: Age, TPD, Life, Occupation, Home State.
* This model effectively predicts IP insurance purchases with high accuracy and provides insights into the significant factors influencing the decision.
* Variable Importance Chart

**Overall Conclusion**

The Random Forest models for all four insurance products not only provide outstanding predictive performances but also offer crucial insights into the factors driving these predictions. Each model demonstrates high reliability, with excellent sensitivity and specificity metrics, making them valuable tools for both operational and strategic applications in the insurance industry. These models can significantly aid in understanding customer behaviour, refining marketing strategies, and tailoring product offerings to meet specific customer needs.