Uncovering the Cost of Risk

An ML Exploration of Life Insurance Premiums in the US and UK

Introduction

Why does a 30-year-old man in the UK pay significantly less for life insurance than his counterpart in the US — or why might a 40-year-old non-smoker in excellent health face higher premiums than a 20-year-old smoker? Insurance pricing is often complex and opaque, but uncovering the patterns behind it is crucial. For consumers, greater transparency means fairer treatment and stronger trust in an increasingly algorithm-driven market, as well as the ability to properly conduct a cost-benefit analysis to ensure they are adequately insured relative to their needs and means. While for insurers, establishing rigorous and justified pricing is essential not only for maintaining profitability — by avoiding losses from underpricing risky customers and retaining/attracting customers through competitive pricing — but also for ensuring accurate reserving, as mispricing risks can lead to either under-reserving (threatening solvency, which is a core regulatory concern) or over-reserving (tying up capital that could be more efficiently allocated elsewhere).

By offering financial protection against the economic impact coming from the occurrence of an unexpected death, life insurance fundamentally acts as a peace of mind for the insured and their loved ones, securing the well-being of their loved ones (through income replacement), covering immediate expenses like funeral costs and estate taxes, and critically, ensuring that large, outstanding debts — such as mortgages — and other financial obligations can be met. But this raises a deeper question: why is it especially important to understand how life insurance is priced? The primary distinction is the long-term contractual agreement inherent to life insurance (present in all types but especially pertinent with level cover), with the terms and pricing of an agreement often being set in stone for decades at a time (usually 10-30 years). Because of this long-term commitment, even minor pricing differences can compound into significant financial impacts — reducing disposable income, straining household budgets, and increasing the risk of underinsurance if customers are forced to scale back or cancel coverage altogether. On the flip side, taking the time to secure lower premiums can lead to substantial lifetime savings, freeing up resources for other financial goals such as investing, debt repayment, or building a stronger safety net. Additionally, although term policies can be cancelled at any time at often no additional cost, there are considerable consequences for not obtaining apt cover from the start. Firstly, if someone were to realise their coverage is insufficient after their health has deteriorated (which is relatively common and a symptom of present bias), obtaining additional or replacement insurance becomes difficult or prohibitively expensive — exasperating the already difficult situation. Moreover, any future application would face an age and risk reset — reassessed at an older age and potentially worse health, leading to substantially higher premiums or even denial of coverage.

Regulators in both the UK and the US are very aware of this issue and have implemented a range of measures to ensure clients fully understand how premiums are determined. These regulations also ensure that insurers base their pricing on a complete risk profile, as insurers to face significant consequences if clients are either overinsured, which leads to unnecessary costs (due to higher claims and increased risk of moral hazard), or underinsured, which can result in financial instability for both the client and the insurer. For instance, in the UK, the FCA has introduced guidelines such as the Insurance Conduct of Business Sourcebook (ICOBS), which includes a stipulation that "A firm must take reasonable steps to ensure that the insurance product it recommends or offers is suitable for the customer's demands and needs" (FCA, 2021), which reduces the incidence of a client obtaining inadequate cover. In the US, under Proposition 103, California requires that the methodology used to obtain premiums and justifications for rate changes be provided to the state department of insurance to be approved before implementation and was likely introduced to mitigate over/underinsurance. Despite these attempts, there is still clearly a disconnect between the consumer and the insurer (in both directions), as is evident in a study by the ABI (Association of British Insurers, 2019). In this study, they found that only 29% of customers believed that they understood how their premiums are calculated, that "70% of clients incorrectly assume that gender is taken into account when pricing" (a reality that will become more evident as we go on), and shockingly, that 41% would prefer to keep "information sharing with their insurer to a minimum", even if it means premiums may rise.

Given this persistent lack of transparency, the primary goal of this project is to open the so-called "black box" of life insurance pricing, demystifying pricing model behaviour and highlighting the key factors that influence premiums, subsequently enabling consumers to recognise the trade-offs involved in their policy choices. For insurers, examining a wide dataset of quotes — including cross-country comparisons between the US and the UK — may reveal key differences in pricing models and average pricing within the market. This could help identify gaps in their risk models and enable them to benchmark their offerings against competitors, improving their competitive edge and pricing strategies in both markets. To achieve these objectives, machine learning will be employed due to its ability to handle large, complex datasets and reveal non-linear relationships between variables that traditional statistical methods might miss. By applying interpretable machine learning models, we can not only predict premium prices based on given risk information but also provide insights into which factors most significantly influence pricing, helping both consumers and insurers understand the underlying drivers of premium calculations.

Methodology

To fulfil these goals, life insurance quotes were scraped off independent insurance brokerage platforms (lifeinsure.com in the US, drewberryinsurance.co.uk in the UK), which provide consumer-facing premium estimates based on user-inputted risk profiles across multiple insurers. Given that these quotes originate from actual insurers and are intended for real customers, they will not only prove an authentic depiction of real-world market conditions, but they will also span a diverse range of products across multiple insurers, capturing the intricacies of the differing pricing models, underwriting criteria and risk tolerances into a single representative dataset to efficiently train our model on. Additionally, unlike

proprietary insurer algorithms, which usually incorporate undisclosed risk factors (e.g., geodemographic data, credit score, past claims history), these platforms clearly reveal the one-to-one marginal effects that changing one risk factor will have on premiums, with the caveat that these are base premiums (which will likely be subject to additional underwriting) and therefore may not accurately indicate what the final consumer will actually pay. Issues may also arise from sampling bias, as commission agreements and partnerships between brokers and certain insurers could artificially limit product competition, skewing the data toward specific offerings. However, this concern is mitigated by the fact that brokers are legally required to act with a fiduciary duty to their clients, meaning they must prioritise the best interests of their clients above any financial incentives from insurers, reducing the likelihood of biased or overinflated product recommendations.

Nonetheless, as with any online data collection, platforms are subject to radical and sudden changes, as was experienced firsthand when LifeInsure implemented scraping restrictions shortly after the final data collection had taken place. At the time, no explicit restrictions were in place, and scraping had occurred smoothly over several days without interruption, suggesting that their scraping tolerance had suddenly changed. Of course, this damages the replicability and perhaps somewhat undermines the validity of this study, yet the data collected still remain valid and representative within the context of the period obtained and will therefore aptly provide valuable contributions to our understanding of life insurance pricing.

Variable Name	Description	Data Type	Sample Size	Summary / Distribution	Notes
`Premium (£)`	Monthly premium in GBP, essentially price of holding the insurance policy.	Numeric (Continuous)	14,180	Target Variable	USD values converted using ~0.746 exchange rate (XE, n.d.).
`ln(Premium)`	Natural log of premium.	Numeric (Continuous)	14,180	Target Variable	
`ln(Coverage_Amount)`	Natural log of coverage amount — the max amount payable in case of a claim, a.k.a "sums insured".	Numeric (Discrete)	14,180	ln([100,000, 250,000, 350,000, 500,000, 750,000, 1,000,000, 2,000,000, 5,000,000])	USD values converted using ~0.746 exchange rate (XE, n.d.).
`Term_Length`	Duration of insurance term (in years).	Numeric (Discrete)	14,180	[10, 15, 20, 25, 30]	All terms are level; coverage remains constant over time (no interaction with term).
`Age`	Age of the individual in years.	Numeric (Discrete)	14,180	[20, 30, 40, 45, 50, 60, 65, 70]	Ages calculated based on 1st Jan baseline.
`Is_Male`	Gender of applicant (1 = Male, 0 = Female).	Categorical	14,180	Male: 7,133, Female: 7,078	
`Is_Smoker`	Nicotine use of applicant (1 = Smoker, 0 = Non-Smoker).	Categorical	14,180	Smoker: 7,102, Non-Smoker: 7,078	
`Is_UK`	Country of quote (1 = UK, 0 = US).	Categorical	14,180	UK: 11,894, US: 2,286	

To ensure consistency and comparability across the datasets, the variables selected for this analysis had to be limited to those shared between the two broker platforms. This consequently means certain key assumptions had to be made in cases where variables were not shared to ensure the integrity and validity of the analysis. Interestingly, the US site required certain health information — subjective personal health ratings, along with height and weight (presumably to calculate BMI) — which was not requested in its UK equivalent for the initial quote calculations (this does not necessarily mean the UK quotes would not have required these at inception). To solve this problem, standard baselines were chosen where no loading was likely to take place, such as an average health rating and a healthy BMI of ~24 (height: 5'10, weight: 167 lbs). Similarly, the UK platform specifically asked for employment status and occupation to be provided (which would supposedly flag especially risky jobs and load rates accordingly), so in much the same vein default options of 'Employed' and 'Others - Not Listed' were set as the employment status and occupation respectively. As general assumptions across both tools, all quotes were based on level cover and no medical exams, allowing for increased consistency between them, since the rate of cover reduction in decreasing term policies and the complexity of medical exams required, along with the harshness of their underwriting are all wildly inconsistent. Furthermore, supplemental policy benefits were ignored as they are too numerous and sporadic to realistically analyze, while critical illness cover was also omitted for simplicity, ensuring the focus remains on core life insurance variables.

Of course, studying every possible combination of variables — especially for the UK site where some inputs allow free-text entry — would be far too time-consuming and inefficient. Instead, representative samples were chosen for certain variables, as shown in the table above. For cover amounts, the sample was designed to have higher resolution at lower values, with wider gaps between values as the amounts increased. This approach reflects general market trends — where the majority of consumers typically hold policies within the £100,000–£500,000 range — ensuring that the model is trained with finer detail where most customers fall, leading to better predictive accuracy and more realistic pricing estimates. As coverage amounts increase, the number of policies falls off sharply — following an approximately exponential decay — meaning less granularity is needed to capture the complete distribution towards the higher ends of cover. To address this expected right-skewed distribution, a log transformation was applied, compressing the extreme values and certifying that Random Forest splits remained sensitive to the full spectrum of coverage amounts while giving more weight to prevalent values. Ages and term lengths worked on roughly the same premise, keeping a consistent distribution, with finer granularity being added in places where marginal changes were likely to be disproportionately high. For example, ages 45 and 65 were chosen because they lie in intervals where most people begin to seriously consider purchasing life insurance (implying premiums may be inflated due to excess demand) and age 65 also marks the cut-off where significant underwriting requirements and loadings are imposed, reflecting the heightened risk of health complications at these ages.

Another important thing to note from the table above is the severe imbalance between US and UK samples, with the UK sample containing roughly 5 times more data than the US sample. Consequently, the patterns and relationships the model captures are more reflective of UK market dynamics, with UK-specific consumer behaviours, pricing structures, and underwriting practices exerting a stronger influence on the learned outcomes.

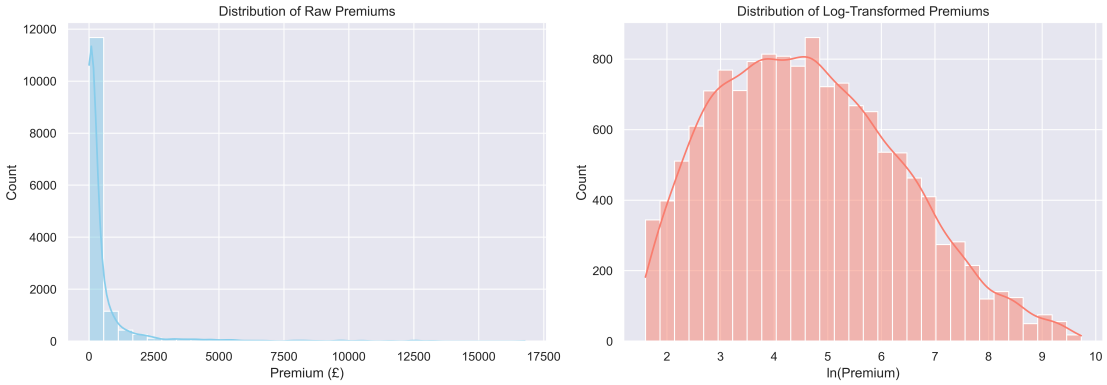
Although the model remains valid and internally consistent within the combined dataset, the relatively small and potentially less representative US sample limits the depth and breadth of US-specific trends captured during training. This raises the possibility that certain effects observed for the US subset may be disproportionately shaped by UK-driven trends or by noise within the limited US data. As such, predictions and inferences related specifically to the US market should be interpreted cautiously to account for the heightened risk of bias and potential underrepresentation.

In this project, Random Forests were used as the primary machine learning model due to their ability to handle large, complex datasets and capture non-linear relationships between variables. A Random Forest is an ensemble method that builds multiple decision trees — where each tree splits the data based on different features (like age or smoking status) — and then predicts the target variable by averaging the outputs of all these trees. By combining the results of these trees, Random Forests improve predictive stability and capture more generalized patterns within the training data, making them well-suited for modelling life insurance premiums, where input-output relationships are often complex. Their robustness to noise and straightforward interpretability make them a practical choice, especially considering the need for the results to be accessible and reliable for a broad, multifaceted audience.

To enhance the interpretability and insightfulness of the Random Forest model, Partial Dependence Plots (PDPs) and SHAP values were computed from its predicted outcomes. PDPs reveal how individual variables influence predictions, holding all other variables constant. Alternatively, SHAP values break down each feature's specific contribution to a given predicted outcome. These tools help clarify the impact of factors like age, smoking status, and coverage amount on premium prices, offering deeper insights into the model's decision-making process. These tools help clarify the impact of factors like age, smoking status, and coverage amount on premium prices, providing a deeper understanding of how the model interprets these variables and their relationships with the target outcome.

In [1]:
```
from IPython.display import HTML # Importing HTML to display figures from our ./notebooks_dev/figures folder.
HTML("""
<div style="text-align:center;">
    <img src="./notebooks_dev/Figures/Fig1-premdist.png" style="width:70vw"/>
</div>
""")
```

Out[1]:

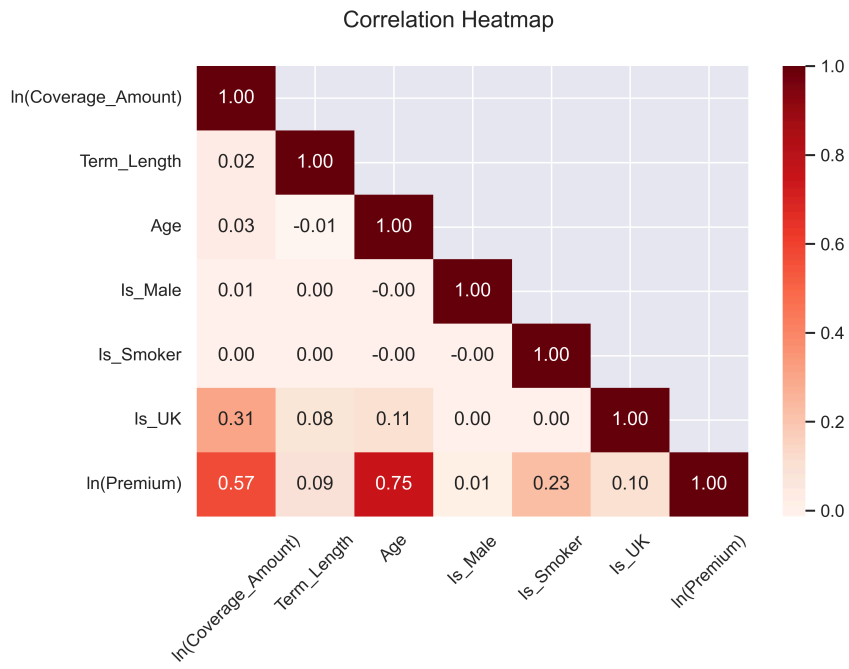


These two plots illustrate the distribution of insurance premiums before and after log transformation. The raw premiums in our dataset are heavily right-skewed, with the majority of values clustered at lower amounts (under £1,000) and a long tail extending to very high premiums (over £15,000) — a typical pattern in insurance because only a small proportion of policies/high-risk individuals warrant premiums that extreme. After applying a log transformation, the distribution becomes much more symmetric and bell-shaped, simultaneously reducing skewness and compressing the range of values. While random forests are generally robust to non-normal targets, highly skewed data can still lead to suboptimal splits, where a disproportionate number of splits focus on rare, extreme values rather than the dense middle range where most premiums lie. Therefore, by transforming the target to a more balanced distribution, the model can partition the data more evenly, allowing it to better capture variations across the full range of typical premiums without being overly influenced by a small number of outliers. Furthermore, log-transforming premiums makes sense conceptually because insurance pricing often operates multiplicatively — risk factors tend to increase premiums by a percentage (loading) rather than by a fixed amount — and the log scale naturally captures this proportional relationship. For these reasons, log-transformed premiums will be used as the target variable during model training, ensuring that the random forest can learn the structure of the data more accurately and produce more stable, interpretable predictions.

Analysis

In [2]:
```
HTML("""
<div style="text-align:center;">
    <img src="./notebooks_dev/Figures/Fig2-corrhmap.png" style="height:450px;"/>
</div>
""")
```

Out[2]:



In the correlation matrix above, we can first observe that there is very little linear correlation (not necessarily interaction) between the explanatory variables, which makes sense since the input features were systematically varied across a wide range of combinations, resulting in minimal dependency between the variables. There was however, one exception to this, namely "Is_UK" seems to interact moderately (that is at least in the context of insurance modelling where hundreds of factors come into play) with coverage amounts, perhaps suggesting that policies originating from the UK might have a higher tendency for higher cover amounts than those from the US. This is surprising, seeing as the US's astronomical healthcare costs would expectedly inflate the amount of cover needed drastically: not only can families be left with huge unpaid medical bills after a death (which must be settled from the estate), but healthcare insurance is often tied to employment, meaning family members may also lose their health coverage — all of which would certainly be taken into account by the insured and the insurer. The more likely explanation becomes evident when looking at the data, as unlike the UK quotes dataset, the US one had much less data present for the upper end of covers ($2,000,000 and$ 5,000,000), mostly because these tended to be medical exam required policies (which makes sense, as insurers need to justify the larger risk and guard against the adverse selection of sick people taking out huge covers), which were explicitly ignored by our scraping.

This bodes well for our model since, although intercorrelated variables are far less problematic for random forests than for other models, minimising them still offers meaningful benefits. The independence of features enhances interpretability, making it easier to understand the individual contribution of variables without needing to account for complex relationships. Additionally, with reduced intercorrelation between variables, the risk of redundant features is minimized — highly correlated inputs are less likely to distort the splits or introduce overlapping signals. As a result, each tree can focus more cleanly on the relevant features, improving both prediction accuracy and the clarity of the model's decision process. Secondly, we can observe that the variables most strongly correlated with premiums are Age and ln(Cover_Amount), with correlation coefficients of 0.75 and 0.57 respectively, while Is_Smoker also shows a notable correlation of 0.23. This is not surprising as these variables tend to have more direct, monotonic relationships with premiums — typically, as age or coverage amount increases, so does the premium, which aligns with general insurance pricing logic. Older individuals tend to pose higher risk, and higher coverage naturally incurs higher costs. Similarly, smokers represent a well-known risk factor, hence their positive association with higher premiums.

In contrast, variables like term_length, Is_Male, and Is_UK show little to no linear correlation with premiums. However, this lack of correlation does not necessarily imply a lack of predictive value. These features may influence premiums in nonlinear or threshold-based ways that a simple correlation metric cannot detect. For example, term_length might exhibit a threshold effect, where premiums are relatively flat for short- to medium-term policies, but increase noticeably once the term exceeds, say, 25 years, due to the insurer's extended exposure to risk. Similarly, Is_Male might show a plateau effect — gender may influence premiums only under specific conditions, such as within certain age bands or policy types, making the overall linear correlation appear negligible, even if the effect is meaningful in practice. Is_UK could follow a piecewise relationship, where the policy being written in the UK has minimal impact on premiums at lower coverage levels or younger ages, but results in a significant pricing adjustment at higher tiers due to regional underwriting policies or regulatory factors.

In short, while correlation is a helpful first step and has provided a general idea of what to expect, it is only a partial lens. The true strength of models like random forests lies in their ability to detect subtle nonlinear patterns — such as thresholds, plateaus, or piecewise effects — that often drive complex outcomes like insurance premiums, even when linear relationships appear weak or nonexistent.

In [3]:
```
HTML("""
<div style="text-align:center;">
    <img src="./notebooks_dev/Figures/Fig3-avgdem.png" style="width:50vw;"/>
</div>
""")
```

Out[3]:

To illustrate how nonlinear effects and interactions can emerge in our data, we will look at our *Average Premium Trends by Demographics* plot, which breaks down premium trajectories by age, gender, smoking status and country. Across all subplots, one pattern is consistent: premiums rise with age, but the relationship is clearly non-linear — with a noticeable inflection beyond age 50. This implies a threshold effect is in play, driven by the fact that the underlying mortality risk increases exponentially with age, something which insurers account for by aggressively increasing premiums at these ranges, to hedge against the growing frequency and costs of claims, thus maintaining their profitability.

This effect is particularly pronounced in the UK, where for non-smokers premiums rise approximately 500% from 50 to 60 and around 120% from ages 60 to 70. This extreme aversion to risk at higher age brackets likely reflects the relatively strict insurance regulations put in place by the FCA in the UK (moreso than the US or even in Europe), one of which includes the FCA's "Fair Pricing & Product Governance" act (FCA, 2021), which seeks to ensure that insurance pricing be proportional to actuarial risk and that there is no "opaque cross-subsidisation" — where insurers charge certain groups more to subsidise the pricing for others. Compare this to the US, where although there are requirements for pricing not to be discriminatory within a group and for them to be actuarially justified, there are no explicit requirements for prices to be proportional to risk under the RBC (NAIC, n.d.), allowing more flexible pricing and risk pooling through methods such as cross-subsidisation (which insurers are not legally required to disclose). It is not immediately evident that the US market is employing cross-subsidisation (mostly because of the scale), but if we look directly at the data, we can see that at the absolute lowest risk profile a consumer in the US is expected to pay on average £9.70, compared to only £5 in the UK. This gives US insurers a larger buffer (especially because this £4.70 difference will be compounded over millions of customers), allowing the large degree of risk smoothing we experience within our graphs, something which is impossible in the UK due to having to price directly in accordance to risk.

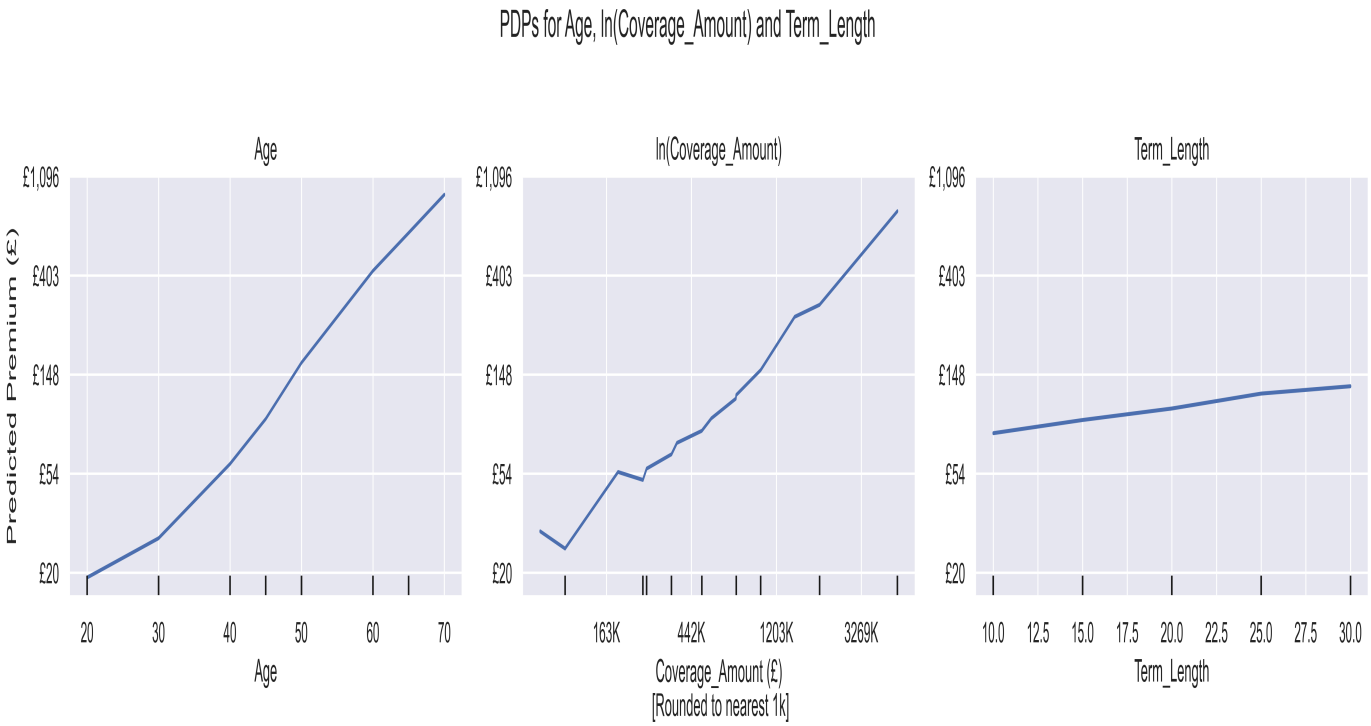
This increased risk pooling flexibility and less rigid obligation for risk pricing is also largely why smokers are not treated as harshly in the US as they are in the UK (roughly a 2000% increase from ages 50 to 60, meaning they can expect to pay ~120% more than non-smokers), as UK insurers must also strictly account for the compounding health risks of being a smoker and at an older age. In addition, differences in solvency requirements between the US and UK also help explain this imbalance — particularly for higher-risk profiles. UK insurers are subject to the more stringent and comprehensive Solvency II framework (Bank of England, 2024), which imposes higher capital requirements based on specific risk exposures, along with strict governance, risk management, and disclosure obligations. By contrast, US insurers operate under the RBC framework, which, while still risk-sensitive, is generally less granular and less demanding in terms of oversight and transparency. This gives US insurers greater flexibility to take on higher-risk individuals at relatively lower prices, since their solvency requirements do not force them to have as much capital to absorb potential losses compared to the UK, contributing to a less risk-averse pricing strategy than is typically seen in the UK.

In terms of gender, we can see that for the UK there is no discrimination in the UK market, which is good news because it means that insurers are properly adhering to the Equality Act of 2010 (The National Archives, 2012), which mandates that gender not be used in directly pricing risk. However, in the US there is no such federal ban, and in fact, gender-based pricing is explicitly permitted and widely used in life underwriting across the US. This culminates in only sight differences between genders (which makes sense because women only outlive men by about 5 years in the US (USAFacts, n.d.)), with slightly higher gender discrimination among smokers, likely reflecting the fact that males tend to smoke more heavily and that hormonal differences make men more susceptible to health risks like heart disease — all of which, and more, would have been captured in the mortality tables insurers use for pricing.

In summary, this visualisation highlights how nonlinearities and interaction effects shape real-world insurance pricing. For example, age has a compounding impact, with premiums rising sharply — especially after age 50 — as insurers price in accelerating mortality risk. Smoking introduces an even stronger threshold effect, particularly in the UK, where premiums can increase twentyfold for older smokers due to stricter requirements to price proportionally to risk. Additionally, we discovered that Gender plays no role in UK pricing, in line with the Equality Act 2010, but has a modest effect in the US, especially among smokers where mortality differences are more pronounced. Cross-country differences also highlight how regulation shapes pricing dynamics: UK insurers face tighter constraints on risk pooling due to FCA and Solvency II rules, while US insurers operate under looser RBC standards, allowing for smoother premium trajectories through opaque cross-subsidisation.

In [4]:
```
HTML("""
<div style="text-align:center;">
    <img src="./notebooks_dev/Figures/Fig4-PDPs.png" style="height:500px;"/>
</div>
""")
```

Out[4]:



While we previously analysed general market trends using line plots, which summarised observed relationships in the raw data (such as average outcomes per feature value), Partial Dependence Plots (PDPs) offer a more nuanced view by revealing how each feature directly contributes to our model's predictions. Unlike line plots, which reflect correlations or trends present in the dataset, PDPs directly estimate the marginal effect of a feature on the predicted outcome — that is, how incrementally changing that variable impacts predicted premiums, ceteris paribus. They do this by varying the target feature while holding all others constant, allowing us to observe the model's learned response to that feature in isolation. This distinction is critical: PDPs show what the model believes the relationship is, which may differ from patterns visible in the raw data, as these can be otherwise affected by confounding variables and interaction effects.

The PDP for age shows a steep, near-linear increase in log premiums (converted to raw premiums for easier interpretation), indicating that as individuals age, their premiums rise exponentially. This mostly mimics what we saw in the line plots discussed earlier, though it is important to note that the PDP's growth rate appears noticeably slower than what we observe for non-smokers in the UK – which is unexpected, since PDPs reflect the effect of age across the full population, including higher-risk cases, and should typically rise faster than a low-risk subgroup like non-smokers. This discrepancy suggests a hidden interaction effect: while the PDP isolates age's marginal impact, it does not account for the real-world interaction that older individuals are also more likely to opt for higher coverage amounts — due to greater financial responsibilities, more dependants, or a desire to leave a larger benefit behind — which would compound with the age-related effects isolated by our PDP and artificially raise premiums for our line plot. Despite this minor caveat, however, the general trend is in line with what we would expect for the same reasons as discussed before.

The PDP for coverage amount shows an almost linear relationship on a log-log scale, with an elasticity of approximately 0.87 (equivalent to the slope on log-log graphs). This indicates that a 1% increase in coverage amount leads to a roughly 0.87% increase in premiums — an almost proportional, but slightly less than one-to-one, relationship. Although this is counterintuitive at first, it is something widely observed within the insurance industry and arises due to nuances in pricing and risk assessment. First, there are significant returns to scale in say providing one policy at £5m rather than 100 policies at £50k, as the insurer's overheads for each policy — such as admin and legal costs — are spread over a higher cover amount, allowing them to provide discounts at these higher covers to increase competitiveness. On top of this, risk assessment is often nonlinear, including aspects such as underwriting thresholds — where beyond a certain cover threshold insurers may require more information or more harshly scrutinise policies (in fact for many brokerages, underwriter input is not given at all below certain thresholds), resulting in less uncertainty for insurers and therefore an increased capacity to discount premiums. Furthermore, there is a prominent selection effect for those buying very high cover premiums, namely, they tend to be wealthier and better educated (and therefore have a propensity to be more health-conscious, have better access to healthcare, tend to have lower stress and are generally less risk-averse) and are subsequently less relatively risky to insure than other demographics.

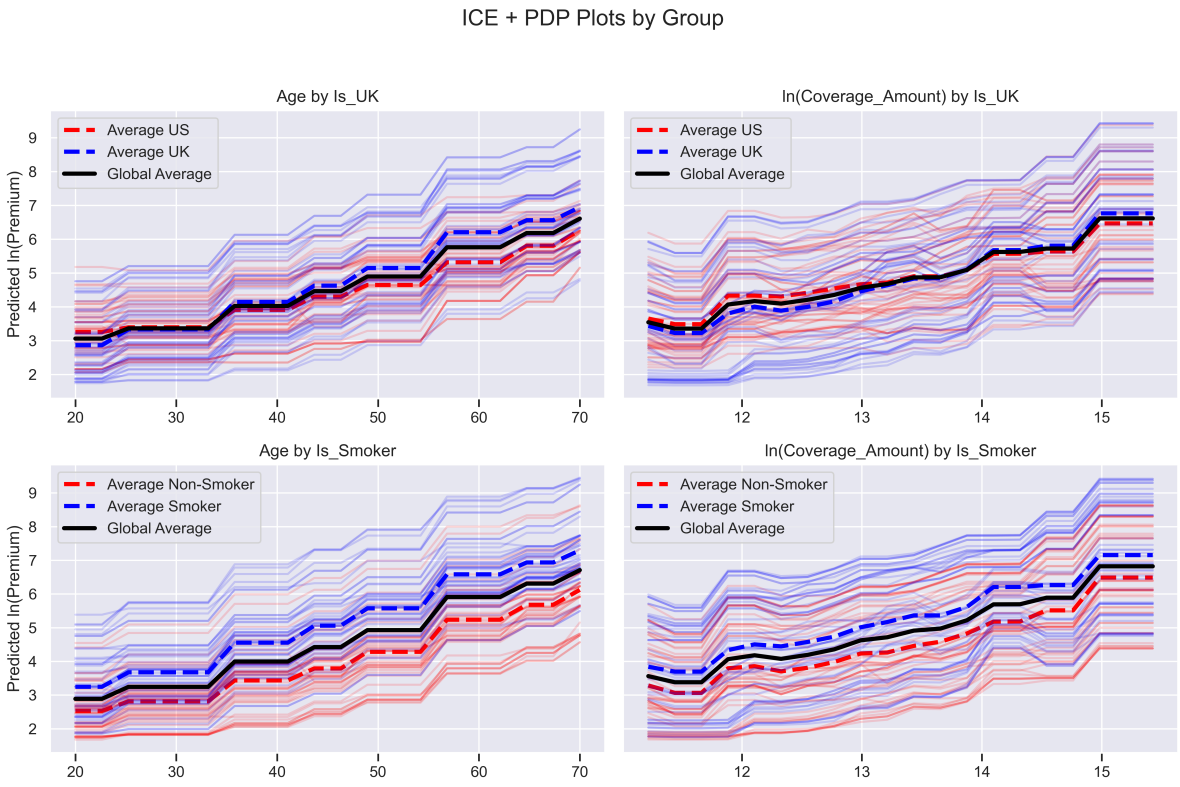
The PDP for term length appears almost flat at first glance, suggesting that term length has little impact on premium pricing. However, this impression is misleading due to the logarithmic scale used for premiums. In reality, increasing the policy term from 10 to 30 years — while holding all other factors constant—results in a substantial 46% increase in premiums. That said, the relationship between term length and premiums is much more linear than that of other features like age. This is because term length extends the duration over which the insurer is at risk, but it doesn't fundamentally change the probability of death in any given year—it just adds more years of exposure. As a result, the increase in premium with term length tends to scale proportionally with time, with insurers typically adding flat premium increases to account for the extended policy — notably, this flat premium increase somewhat levels off beyond a 25-year term, which makes sense because the risk is primarily front-loaded (if a claim occurs early in the term, accumulated premiums may not cover the claims costs) and because costs incurred decades in the future have much lower present values (as per the present value discounting formula). In contrast, age affects the per-year risk itself: mortality rates increase

exponentially as a person ages, particularly beyond middle age. Therefore, a small increase in age can drastically raise the likelihood of a claim, requiring a much sharper increase in premium to offset that risk.

While line plots helped identify general market trends based on observed data, PDPs provide a clearer view of how individual features impact predicted premiums in isolation by controlling for interaction effects. The PDP for age reveals an exponential rise in premiums, aligning with expectations but increasing at a slower rate than seen in real-world subgroups like non-smokers likely due to interactions with other features like coverage amount. The coverage amount PDP shows a near-proportional log-log relationship (elasticity ~0.87), where the lower marginal cost of additional cover reflects economies of scale, underwriting thresholds, and favourable selection among high-coverage applicants. For term length, the PDP appears relatively flat on a log scale but actually reveals a meaningful 46% increase in premiums from 10- to 30-year terms. This relationship is linear rather than exponential because term length adds duration, not per-year risk. The premium effect levels off beyond 25 years due to front-loaded risk and the diminishing present value of long-term liabilities.

In [5]:
```
HTML("""
<div style="text-align:center;">
    <img src="./notebooks_dev/Figures/Fig5-ICEs.png" style="width:75vw"/>
</div>
""")
```

Out[5]:



Individual Conditional Expectation (ICE) plots are similar to PDPs in that they show how a machine learning model's prediction changes as an input varies — holding all other variables constant — but differ in the sense that rather than simply showing the average effect of the input feature they instead visualise this dependence for each sample separately, with one line per sample. This is useful because it allows us to detect heterogeneity in the model's behaviour — specifically, whether the effect of a feature varies across observations — something PDPs may obscure by averaging out these differences. They also allow us to observe whether there are significant interaction effects between our variable and the target (and therefore whether our PDPs are reliable), as in cases with little interaction, only the additive effect should be present — where individual samples are merely parallel shifts of one another, suggesting that the effect of the specific feature examined is constant across all observations, with the baseline prediction (i.e., the starting value of the prediction) changing in accordance to the other features being what actually causes the parallel shift. Observing our graphs, we can see that while age has almost no interaction effects, log-transformed cover amounts have much more variation resulting from interaction effects, with especially high disparities in comparison to the global average (the PDP) at the mid-to-high range of cover amounts (~£500k and ~£1.5m). It is impossible to know for certain what these interaction effects are, however, they do seem to mark specific thresholds for homeowners with large mortgages, or other long-term commitments like business start-up loans, perhaps causing insurers to artificially inflate these brackets because the excess demand for these high cover policies generating too much risk. Alternatively, they may mark thresholds where underwriting becomes more robust (perhaps because they are so common), leading to sudden jumps when compared to the lax risk profiling of the prior threshold.

If we specifically look at how age rating differs across US and UK markets, we can see that at younger ages (from 20-30), US premiums are consistently priced higher than the global average (translates to a difference of about £5), while the UK is consistently priced below it. The two then converge briefly before diverging again at around age 45, at which point the trend swaps: US premiums become significantly cheaper, as much as £260 below the global average. This supports earlier claims that the US market employs opaque cross-subsidisation to great effect, strategically redistributing risk across age bands in a way that dramatically lowers premiums for older individuals. The UK, by comparison, appears significantly more expensive, not necessarily due to inefficient pricing or regulatory rigidities, but because the US approach pulls down the global benchmark, exaggerating the gap. In contrast, the US and the UK markets seem to price the effect of increasing cover by roughly the same amount, with the only significant differences being in the ~£150k to ~£450k range, where the US is slightly above the average and the UK slightly below it — again, likely reflecting the US' cross-subsidisation, as they disproportionately overcharge the low-risk groups which are more likely to fall in these areas (think young homeowners taking out a mortgage on their first home), while insurers in the UK try to reflect the true risk in their pricing.

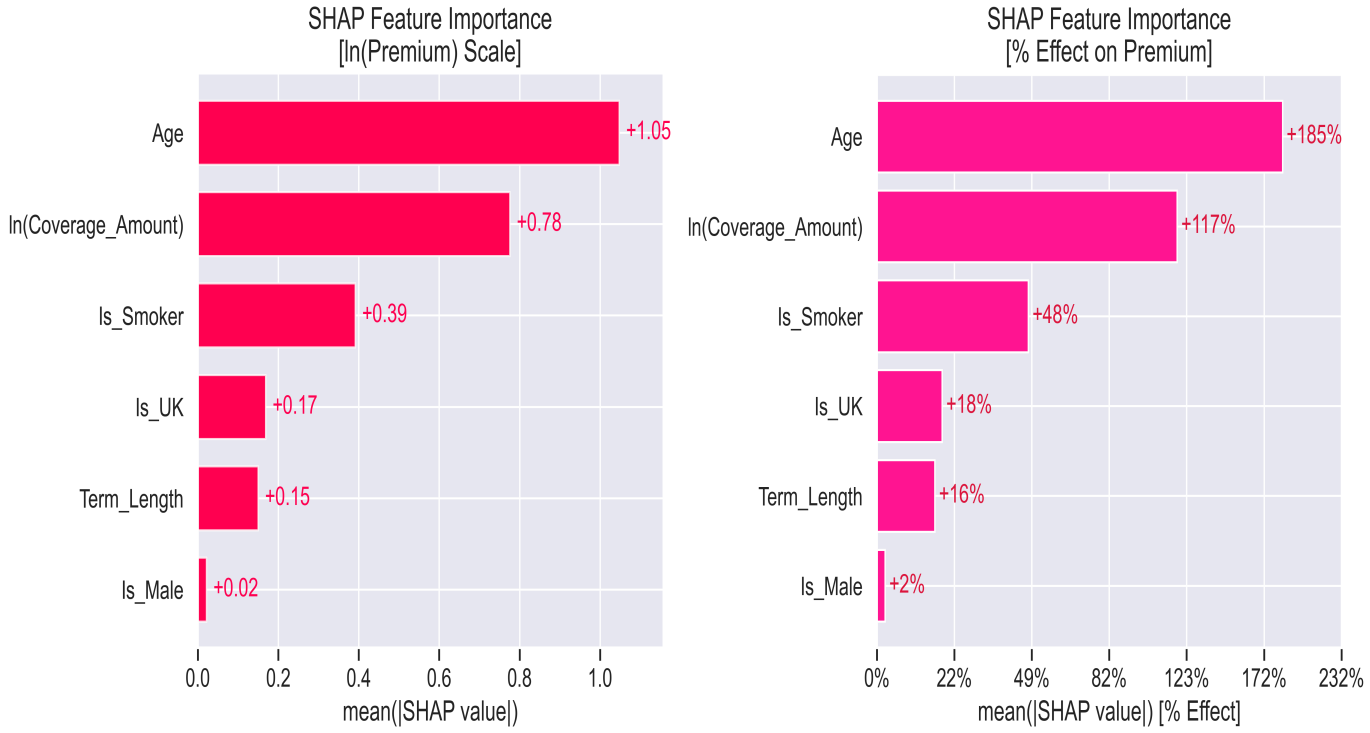
Interestingly though, the risk-smoothing at the higher cover amounts is relatively subdued, supporting our claim that insurers do not interpret risk as scaling completely linearly with cover amounts (the elasticity is still around 0.85 for both these new log-log models), as the risk smoothing the US usually employs has been scaled back to account for this.

If we analyse how being a smoker affects premiums across ages and cover amounts, we observe a consistent pattern: a relatively flat, parallel shift, with premiums scaling upward above the global average for smokers and downward, below the average, for non-smokers. Importantly, these parallel shifts exhibit consistent magnitudes across all ages and cover amounts, suggesting that most insurers (UK and US alike) simply add a percentage loading to premiums for smokers (this does not apply in reverse, it is just that the global average is driven up because of it), rather than scaling the effect about other variables like age or coverage amount. This largely additive treatment does not follow real-world trends however, as being a smoker becomes riskier with age, and the inherently increased risk of death should make insurers more hesitant to price higher covers (although in practice insurers would probably just forego the client). Still, this approach remains common for insurance pricing because it keeps pricing transparent and simple for underwriting, and because these nuances of smoking becoming riskier with age are largely captured within the age variable through actuarial mortality tables (essentially how likely you are to die at a given age given certain demographic factors), at which point the consistent loading for being a smoker is applied after accounting for age, fundamentally leading to the same result with the caveat that the importance of the age variable in our model becomes inflated because of this.

From analysing ICE plots, we observed significant interaction effects in cover amounts, particularly in the mid-to-high range (~£500k and ~£1.5m), possibly indicating risk thresholds for homeowners with large mortgages or long-term loans. In contrast, age shows little interaction effect. When comparing premiums in the US and the UK, we found that US premiums are higher for younger ages (20-30) and significantly lower after age 45, suggesting effective cross-subsidisation. This drives the global average down, making the UK appear more expensive by comparison. For cover amounts, both markets price similarly, with minor variations at lower coverage levels due to differing risk profiles. The impact of smoking is seen as a consistent, flat premium loading across ages and cover amounts, suggesting that smoking risk is accounted for in age-based mortality tables, with insurers applying a simple, additive loading rather than scaling the effect with other variables like age or coverage. This does not mean the interaction nuances of smoking with other variables are completely absent however, as they are largely captured within the age variable through actuarial mortality tables, meaning the actual direct effects of age in our model may be inflated by this underlying interaction effect (although in practice, its role in pricing life insurance in the model is not distorted).

In [6]:
```
HTML("""
<div style="text-align:center;">
    <img src="./notebooks_dev/Figures/Fig6-SHAPimp.png" style="height:500px"/>
</div>
""")
```

Out[6]:



These SHAP global importance plots illustrate how much, on average, each variable impacts predicted premiums, whether in the positive or negative direction. The left plot shows the average impact of each feature on the log of the premium. In contrast, the right plot translates these effects into approximate percentage changes in actual premium amounts. For example, the mean absolute SHAP value for age in the log premium scale means that, on average, age contributes ±1.05 to predicted log premiums, or alternatively, a ±185% change in raw premiums. By highlighting the contribution of each feature, SHAP bar importance plots will be invaluable for understanding and interpreting our model, providing a clear view of which features are most influential and helping to identify the key variables that drive our model's predictions.

From our SHAP feature importance plot, we first observe that age is by far the most prominent factor for predicting premiums, causing premiums to change by an average of ±185% (although direction is not stated this is almost certainly an increase). This is expected, as not only is it the predominant factor for life insurance pricing across every market, but we have also seen its astronomical influence on pricing internally, through our PDP and ICE plots. However, as previously discussed, since the age variable essentially dictates where you fall on insurers' mortality tables, this variable incorporates many interaction effects as part of age-based mortality risk — one of which we saw before with the ages variable likely encapsulating the Is_Smoker variable's interaction with age — thus inflating the actual effect that age directly exerts over premiums, although if we

look at it strictly from a real-world importance perspective this becomes irrelevant since this is precisely what makes it the dominant rating factor, as you become able to easily capture a large proportion of the risk (by leveraging the historical data used to construct mortality tables) via just one variable.

Trailing not too far behind are coverage amounts, which have an average effect of ±117% on raw premiums (i.e., they have an average multiplicative effect of 2.17×). This relatively large impact aligns with real-world expectations — premiums should naturally scale with the amount of coverage since insurers are taking on more financial risk with higher payouts. However, as discussed before, we would expect premiums to grow sub-linearly (i.e., with an elasticity < 1) and in fact, within our model we calculated this as being the case as a 1% increase in cover was shown to cause an equivalent ~0.87% rise in premiums. So, why do coverage amounts appear to multiply premiums by a factor of 2.17×, which seems unintuitive given the previously established sub-linear relationship? The answer lies in the structure of the coverage amount variable — since if we consider that cover amounts tend to increase in large, discrete jumps (e.g., £100k → £250k → £350k), going from £100k to £250k cover essentially means increasing the coverage amount variable by 150%, which should incite an equivalent rise of 130.5% in premiums, which is more in-line with the substantial average shifts in premium that our SHAP global importance model predicted.

Similarly, Is_Smoker also has a significantly large effect, with an average effect of ±48% on premiums (as discussed before, this most likely excludes interaction effects like age) — again because being a smoker increases the risk of health issues and therefore the relative risk incurred by insurers. This might then falsely lead you to assume that being a smoker applies on average a 48% loading to your premiums, but this is not the case, as being a non-smoker will generate negative SHAP values (since SHAP values capture the fact that being a non-smoker will reduce your premiums relative to a baseline), even though real-world pricing would only apply a positive loading to smokers — not a discount to non-smokers — meaning SHAP values will not directly reflect the true loading.

Perhaps the most interesting finding is that policies in the UK contribute an average of ±18% to raw premiums (we will establish the directionality of this later), implying that there are pretty significant pricing differences between the US and the UK. This disparity is very much expected, as throughout this investigation it has been reiterated many times just how disparate the two markets are, this is mostly due to the differing regulations put in place by each respective conduct authority, but differing consumer profiles, needs (consumers have needs for life insurance in the US to cover outstanding healthcare debts, something which is otherwise irrelevant in the UK), and market conditions also play an important — but albeit less significant — role. Term lengths also had a similarly moderate — but still significant — average contribution of ±16% to raw premiums. Although this may seem less than what we would initially expect, it makes sense within the confines of the model and real-world intuition, as for reasons mentioned before, increasing term duration beyond a certain threshold adds little additional risk to insurers and unlike other variables in the model, term length does not directly impact risk — only the duration of the risk, meaning it makes more sense to scale premiums linearly rather than multiplicatively, resulting in a subdued global importance. Finally, we also corroborated the fact that gender has very limited effects on premiums, only contributing an average of ±2%, with the only reason it is not completely irrelevant (which it should be if firms are following the Equality Act of 2010) being the inclusion of US policies into our model, but still even there being male would likely only boost your premium by a few percentage points.

Overall, the SHAP global importance analysis highlights the key drivers of life insurance premiums in our model, with age emerging as the most dominant factor, contributing an average ±185% change in raw premiums — a reflection of how closely it aligns with insurer mortality tables and their embedded risk structures. Coverage amount follows closely behind, exerting an average ±117% effect, which, despite a sub-linear relationship in log-log space, is justified by the large, discrete jumps in typical coverage options. Smoking status also shows a substantial influence, with a ±48% effect, though this should not be interpreted as a direct loading, given the relative nature of SHAP values. Geographic location (UK vs. US) contributes a notable ±18%, confirming the impact of structural and regulatory differences between markets. Term length, while conceptually important, has a more modest effect (±16%), likely due to its linear (rather than exponential) impact on risk exposure. Finally, gender contributes the least (±2%), consistent with regulatory restrictions in the UK and only driven up slightly by the marginal differences in US pricing. These insights collectively reinforce the model's alignment with industry logic while also revealing how variables interact in practice.

In [7]:
```
HTML("""
<div style="text-align:center;">
    <img src="./notebooks_dev/Figures/Fig7-SHAPbee.png" style="height:500px;"/>
</div>
""")
```

Out[7]:

A SHAP beeswarm plot is a powerful visualization for understanding not just which features are important in a machine learning model, but how they influence individual predictions. Unlike a global SHAP importance plot—which ranks features solely by the average magnitude (mean absolute SHAP values) of their impact—a beeswarm plot retains the directionality of each feature's contribution. Each point represents a single prediction, with its position on the x-axis indicating whether the feature pushed the predicted value higher or lower, and the color encoding the feature's actual value. This allows us to see patterns such as whether high values of a feature tend to increase or decrease the prediction, making the beeswarm plot especially useful for understanding how premiums respond to different values of key input variables.

Looking at age, we again observe that it is by far the most important factor, but this time we can truly visualise the true range of its effect. Going from the baseline of zero (which is the mean predicted premium across the entire dataset, and seems to occur at a midrange age of slightly below 45), we can see that at age 20 the predicted premium is expected to be ~82% less than the baseline and at the highest age value 70 observed in the model, premiums are expected to be ~1000% higher than the baseline. On top of this, we can confirm that the effect of age is indeed exponential, as the values for age are arranged in a steadily increasing sequence across the SHAP value axis — indicating an exponential effect when converted from log space into raw premiums. It is also worth noting that the clear separation between age clusters is merely a side effect of age being a discrete variable (this is what causes the single-colour clusters) and the asymmetric sample (responsible for the gaps of missing data, if more granularity was introduced we would essentially see a straight, exponentially rising trend). We can observe this same steadily, exponentially rising trend for log-transformed cover amounts, although we can see here that the extreme cover value of £5m may have caused its global importance to be slightly overstated in the previous graph, as there are fewer intermediate values to bring the mean absolute SHAP value more in-line with what it should actually be (this effect is however likely minimal, and cover would remain the second-most important variable).

Observing Is_Smoker, we can see with complete certainty now that insurers do just apply flat percentage loadings based on whether you are a smoker since both clusters representing the groups of smokers and non-smokers are an equal distance away from the baseline. Crucially, this finally allows us to calculate an estimate for the average percentage loading across both the UK and US life insurance markets, as we can now perform our calculations with the non-smoking group as a baseline (which is how loadings are applied by underwriters in the industry). Taking both groups as having SHAP values of roughly ±0.4, we can then calculate the percentage change as ~122% when going from the non-smoking group to the smoking group. This seems to be pretty in line with general market trends, as comparethemarket — a prominent insurer in the UK — stated that "for every £10 a non-smoker pays each month, smokers could expect to pay between £1 and £10 extra" (Compare the Market, 2025), while in our model the predicted increase is closer to £12.20 (i.e., £22.20 for smokers versus £10.00 for non-smokers), which is a modest overshoot. This discrepancy could be due to factors such as differences in the underlying population, slightly higher risk sensitivity in our model, or regional variations in pricing practices between insurers.

Now we see that the trend with Is_UK is perhaps opposite to what we might initially expect, as the UK seems to be on average 22% cheaper compared to the baseline within our model (the log-scale makes it appear less significant than it actually is), while the US is consistently above the baseline and is very spread out with no real trend. This might seem counterintuitive, as one would expect the lower regulation in the US, combined with greater risk pooling and cross-subsidization, to drive premiums down, making them lower than those in the UK. However, the key differentiating factor lies in how lower-end premiums are priced in the US, where insurers tend to apply more aggressive, risk-sensitive pricing. Since these lower-end policies make up the bulk of the premium distribution, they disproportionately influence the model's average predictions and offset the advantages the US might have at the higher-end policies due to regulatory flexibility or pooling mechanisms.

The SHAP value distribution for Term_Length is clustered near zero, indicating that, on its own, it tends to have a relatively modest and mixed influence on predicted premiums when compared to the substantial effect of other variables like age and cover amounts on predicted premiums. Term length does still have a consistent directional pattern, with higher terms contributing to higher predicted premiums on average, and it seems that unlike what was previously hypothesised, term duration's effect on premiums does not level off at all, it is just the underlying risk to insurers that does (which explains why it tends to increase relatively more linearly rather than exponentially). Still, the log scale does disguise some of the importance this variable can have on predicting premiums, as a high value of 30 and a low value of 10 contributed to an increase and decrease of

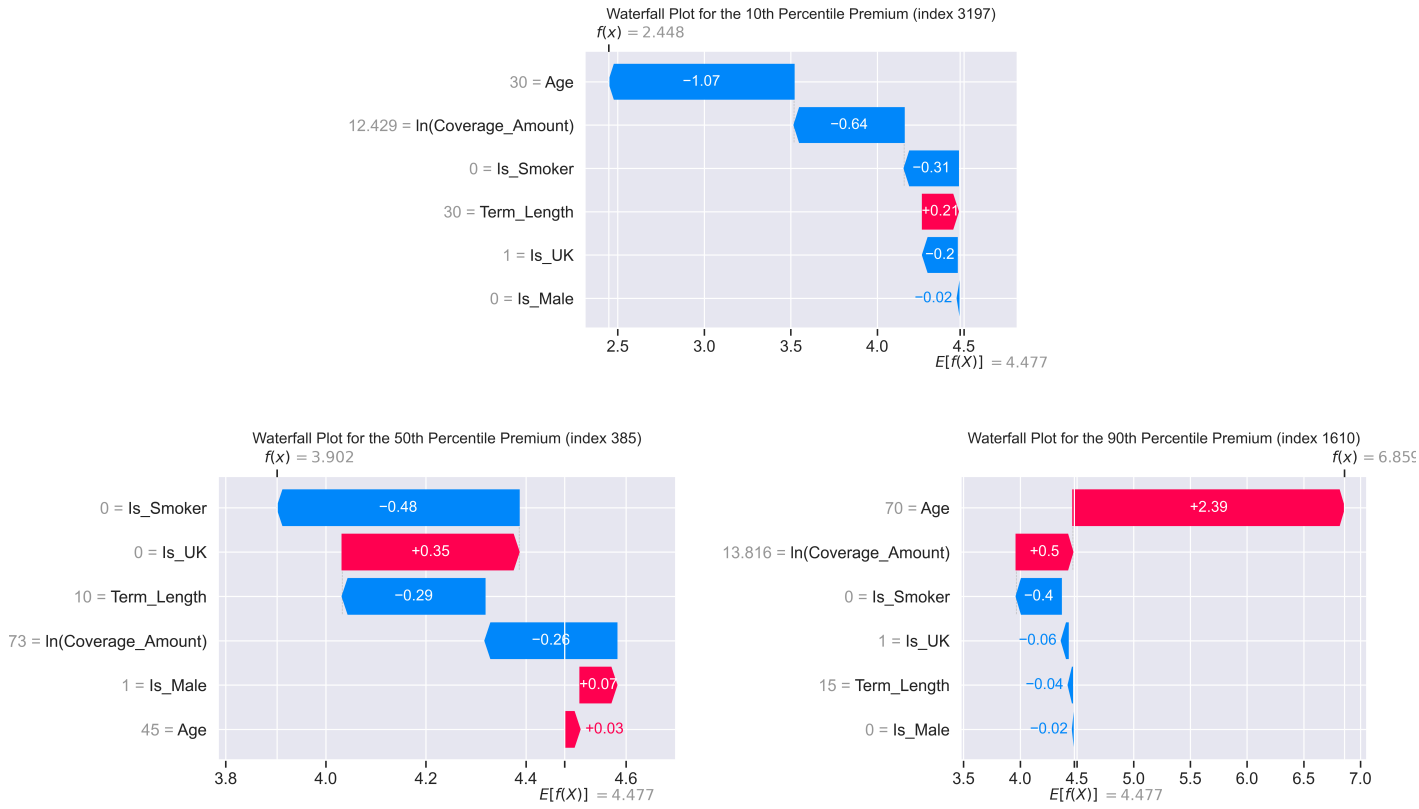
roughly 65% (note that the equal distance from the baseline suggests each successive term cover adds a consistent percentage increase to premiums), respectively.

In summary, the SHAP beeswarm plot reveals how individual features influence life insurance premiums across the dataset. Age is the most dominant factor, showing a clear exponential relationship with premiums: young applicants (e.g., age 20) pay ~82% less than average, while older individuals (e.g., age 70) can pay over 10× more. Log-transformed coverage amount also shows a steadily increasing effect, confirming its strong role in pricing. Smoking status produces a clear binary shift, with an estimated average smoker loading of ~122%, consistent with industry trends. Surprisingly, the UK appears ~22% cheaper than the US, contrary to expectations given the US's regulatory flexibility. This is likely due to the harsher pricing of low-end policies in the US, which dominate the distribution and offset potential cost advantages. Lastly, Term Length shows a modest but consistent impact: longer terms increase premiums in a near-linear, percentage-based fashion, despite appearing less important on the log scale. Note that Is_Male was not directly analysed since it revealed no new relevant insights.

In [8]:
```
HTML("""
<div style="display: flex; justify-content: center; margin-bottom: 20px;">
    <img src="./notebooks_dev/Figures/Fig8-SHAPwater.png" style="width: 45vw;" />
</div>

<div style="display: flex; justify-content: center;">
    <img src="./notebooks_dev/Figures/Fig9-SHAPwater.png" style="width: 45vw; margin: 10px;" />
    <img src="./notebooks_dev/Figures/Fig10-SHAPwater.png" style="width: 45vw; margin: 10px;" />
</div>
""")
```

Out[8]:



A waterfall plot is a visualization tool commonly used to show how individual features contribute to a final prediction. Each bar in the plot represents a specific feature's impact on the model's output, either increasing or decreasing the predicted value relative to a baseline. The cumulative effect of all features can be seen by the end of the plot, allowing for a clear breakdown of the factors driving the prediction. This approach is particularly useful in understanding the local interpretability of a model, where you can trace how individual data points—such as a specific individual's characteristics—affect their predicted premium. Here, we generated waterfall plots for specific sub-sets of the data — indexing on the 10th, 50th and 90th percentile premiums — allowing us to identify if trends change at different premium levels (i.e., low, medium and high).

At the 10th percentile of predicted premiums (Index 3197), the model estimates an actual premium of £11.58, well below the average of £88.13. The most significant contributor to this low premium is the individual's young age (30), which reduces the premium by approximately 66%. A low coverage amount further decreases the premium by around 47%, and being a non-smoker brings it down an additional 27%. Minor downward influences include being a UK resident (about 18% lower) and male (2% lower). Interestingly, this small gender-related difference occurs despite gender-based pricing being prohibited under the Equality Act in the UK, almost certainly a side-effect of the US's gender discrimination being taken into account during model training, ultimately introducing a subtle bias even in UK-specific cases. The only feature increasing the premium is the long policy term (30 years), which adds about 23%. Overall, the low premium stems from a favourable combination of youth, low coverage, and non-smoking status.

In the 50th percentile case (Index 385), the predicted premium is £49.37, still below the average. The non-smoking status reduces the premium by roughly 38%, a shorter policy term (10 years) reduces it by around 25%, and a modest coverage amount decreases it by another 23%. Interestingly, being UK-based increases the premium by 42% in this case, suggesting regional effects may interact with other features. Gender increases the premium slightly by 7% (again, gender discrimination is in fact present in US markets), and age (45) has a minimal effect of about 3% (suggesting

this is the baseline age the model works from). Despite some upward pressures, the premium remains moderate due to the strong downward effects of smoking status, policy term, and coverage.

At the 90th percentile (Index 1610), the predicted premium reaches £952.60, far exceeding the average. The primary driver is the individual's age (70), which increases the premium by nearly 990%. A very high coverage amount adds around 65%. Although the individual is a non-smoker, reducing the premium by 33%, is not enough to offset the upward pressure. Minor effects include being a UK resident (6% lower), a 15-year policy term (4% lower), and gender (2% lower). Overall, this high premium is overwhelmingly explained by advanced age and substantial coverage needs.

At the 90th percentile (Index 1610), the predicted premium surges to £952.60, far exceeding the average. The dominant factor is the individual's age (70), which inflates the premium by nearly 990%. A very high coverage amount adds another 65%. Though the individual is a non-smoker, leading to a 33% reduction, this is insufficient to counterbalance the extreme upward pressures. Minor downward influences include being a UK resident (6% lower), having a 15-year policy term (4% lower), and being female (2% lower), the latter again reflecting the gender bias of the US-influenced training data. Overall, this high premium is primarily driven by advanced age and large coverage needs.

Collectively, these three cases illustrate and reiterate the consistent trends in how different features affect life insurance premium predictions. Age stands out as the most influential factor, particularly at the higher end of the premium distribution. The coverage amount, captured through its logarithmic transformation, also exerts a steady and interpretable impact across the spectrum. Being a non-smoker consistently reduces premiums, underscoring the importance of health-related risk factors in underwriting models. Other variables such as country and gender exhibit smaller and sometimes inconsistent effects, while term length has a moderate, context-dependent influence on pricing.

Conclusion

The primary objective of this project was to demystify the decision-making process of a life insurance pricing model using advanced interpretability techniques and to assess whether the model's behaviour aligns with established real-world underwriting logic. This objective was largely achieved through the use of a Random Forest model, supported by a suite of visualisation tools. By leveraging SHAP values — including both global feature importance and detailed beeswarm plots — along with Partial Dependence Plots (PDPs), Individual Conditional Expectation (ICE) curves, and local SHAP waterfall charts, the analysis moved beyond simple surface-level analysis, which hides important trends due to being unable to separate interactions and accurately predict nuanced premium variations based on individual applicant characteristics. These tools provided a transparent and structured view of how life insurance premiums are determined and adjusted across a wide range of applicant profiles.

Perhaps the most striking and consistent insight was the overwhelming dominance of age as a pricing factor — an unsurprising result given its foundational role in mortality tables, which inherently capture a wide range of interaction effects. Not only did global SHAP values rank it as the most important feature by a significant margin, but the beeswarm plot clearly illustrated its exponential relationship with premiums: a 20-year-old can expect to pay approximately 82% less than the average premium baseline, while a 70-year-old faces a loading of nearly 1000% above that baseline. The SHAP value distribution revealed clearly ordered strata by age group, reflecting the discrete encoding of the feature and illustrating how risk escalates predictably with age. This pattern was consistently mirrored in both the PDPs and ICE curves, aligning closely with real-world underwriting, where mortality risk — and thus pricing — increases markedly with each passing age band, especially in later decades of life.

Smoking status, while binary, showed an equally clear and interpretable pattern. Throughout the study, we consistently found that being a smoker consistently increases premiums, applying what seemed to be a flat percentage loading, which we were able to derive an estimate of 122% for, from the beeswarm. This directly mirrors the underwriting principle of applying smoker loadings as fixed percentage multipliers, and it allowed us to quantify the average financial impact of smoking across the dataset. This estimate aligned well with public figures from insurers like Compare the Market, suggesting that the model's calibration is reasonable and consistent with market expectations.

Coverage amount, especially when log-transformed, also followed expected economic logic. The PDP and ICE plots revealed a smooth, monotonic increase in premiums as cover rose, and the SHAP beeswarm plot confirmed this trend — though the importance to our model in the global SHAP plot was slightly overstated due to the presence of extreme values like £5 million. This reflected the inherently multiplicative nature of pricing larger coverage amounts: premiums scale up, but not completely proportionally (we found that they only scaled up by an equivalent of ~0.87). Term length showed a modest but meaningful contribution. Contrary to the hypothesis that its effect would plateau — reflecting insurers' diminishing marginal risk over longer durations — SHAP values showed a linear contribution: each additional 10-year term increased premiums by roughly 65% or more. This suggests that while mortality risk may flatten, the model learns that the insurer's exposure over time still drives consistent premium increases.

The variable Is_UK produced one of the more surprising findings. Initially, one might expect US policies — with their looser regulations and competitive pricing — to be cheaper. However, SHAP results revealed the opposite: UK applicants were, on average, paying 22% less. Upon closer inspection, this was attributed to how aggressively US insurers price low-end policies, which form the bulk of the dataset. These policies, while cheap, reflect higher sensitivity to risk factors, pushing up average US premiums and offsetting any regulatory cost advantages. This insight underscores how pricing dynamics can differ not just by regulation, but also by market structure and insurer behavior.

Gender (Is_Male), while globally less influential, surfaced in subtle but important patterns. Despite gender-based pricing being prohibited in the UK, our SHAP waterfall plots showed small but consistent premium differences — suggesting residual bias from the US data used in model training. For example, being male reduced premiums by about 2% in the 10th percentile case. This highlights how bias can inadvertently persist in models trained on multinational data, even when not explicitly allowed by law. On the contrary, we were able to study some of the effects that gender discrimination has on US pricing in one of our waterfall plots, where the 50th percentile premium (which was a US quote) experienced a ~7.25% loading to premiums because the insured was a male, which although not too extreme, is in line with market expectations and stays consistent with the fact that although significant, the life expectancy disparities between genders are unlikely to affect mortality rate too drastically.

The final layer of analysis came through SHAP waterfall plots at three premium percentiles — low (10th), median (50th), and high (90th). These gave us a case-by-case breakdown of how feature contributions shift across the distribution. At the 10th percentile, low premiums were driven primarily by young age, low coverage, and non-smoking status. At the median, downward influences were similar, though with some unexpected upward pressure from UK residency — highlighting interactions that global metrics can miss. At the 90th percentile, high age and large cover amounts overwhelmingly dominated, confirming that at the upper end of the premium scale, traditional risk factors drive cost nearly singlehandedly, with other features exerting only marginal influence.

Collectively, these findings demonstrate that the model not only captures the key drivers of life insurance pricing but applies them in ways that align with established underwriting logic — both in terms of feature importance and directional effects. The use of interpretability tools enabled validation at both the global and local levels, surfacing deeper patterns and assumptions — some expected, others surprising — embedded within the model's decision-making. With this foundation in place, the next step is to apply these insights — whether to audit existing pricing frameworks, refine product design, or enhance transparency in customer-facing tools. On a consumer level, I hope that this study has helped to demystify some of the complexities behind insurance pricing — shedding light on how factors like age, coverage amount, and smoking status influence premiums, and illustrating that these decisions are often grounded in consistent, data-driven logic rather than arbitrary outcomes.

References

Project Repository Link: https://github.com/freitas-andrew/life-insurance-scraper-ml-study

FCA (2021). ICOBS 6.1.1 - Suitability. FCA Handbook.
Retrieved from https://www.handbook.fca.org.uk/handbook/ICOBS/6/1.html

Association of British Insurers. (2019). Consumer attitudes towards data and insurance.
https://www.abi.org.uk/globalassets/files/publications/public/data/britain_thinks_consumer_data_insurance_report.pdf

XE.com. (n.d.). XE Currency Converter. Retrieved April 21st, 2025,
https://www.xe.com/currencyconverter/convert/?Amount=1&From=USD&To=GBP

Financial Conduct Authority (2021). PS21/11: General insurance pricing practices – amendments.
https://www.fca.org.uk/publications/policy-statements/ps21-11-general-insurance-pricing-practices-amendments

National Association of Insurance Commissioners. (n.d.). Risk-Based Capital (RBC) for Insurers Model Act.
https://content.naic.org/sites/default/files/model-law-312.pdf

Bank of England (2024). Review of Solvency II: Adapting to the UK insurance market.
https://www.bankofengland.co.uk/prudential-regulation/publication/2024/february/review-of-solvency-ii-adapting-to-the-uk-insurance-market-policy-statement

The National Archives (2012). The Equality Act 2010 (Amendment) Regulations 2012 No. 2992.
https://www.legislation.gov.uk/uksi/2012/2992/pdfs/uksi_20122992_en.pdf

USAFacts (n.d.). Do women live longer than men in the US?
https://usafacts.org/articles/do-women-live-longer-than-men-in-the-us/

Compare the Market (2025) - Life Insurance For Smokers.
https://www.comparethemarket.com/life-insurance/content/for-smokers/?