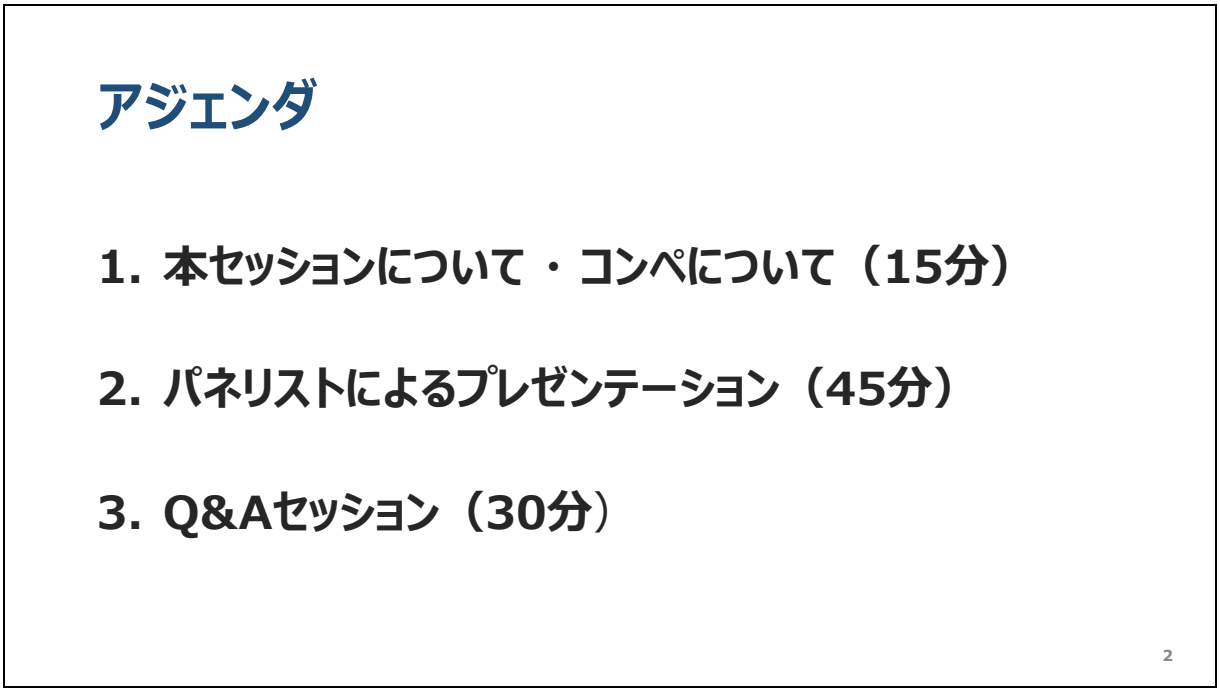
データ分析コンペティションとアクチュアリー

Guy Carpenter　藤田　卓君
Guy Carpenter　崎村　沙羅氏
Guy Carpenter　久下　康太朗君
アメリカンホーム医療・損害保険　荒井　智弘君
三井住友海上火災保険　伊藤　健太君

藤田　皆さん、こんにちは。「データ分析コンペティションとアクチュアリー」のセッションにお越しいただき、ありがとうございます。

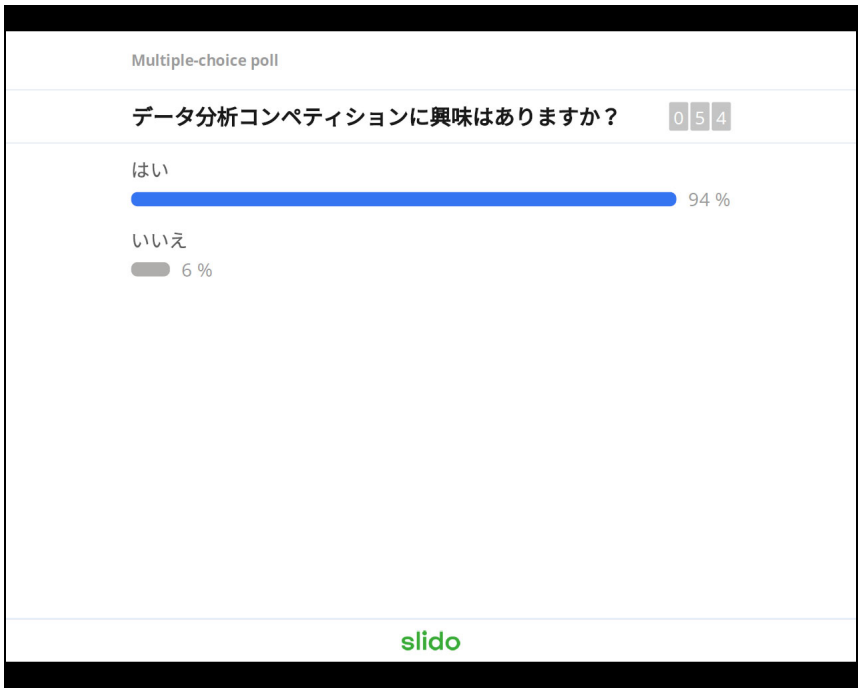


本セッションは 90 分間のパネルディスカッションで、４人のパネリストをお招きしております。また、大きく３部構成になっており、まず私の方から今回取り上げるコンペの概要についてお話しします。次に、パネリストの方々からコンペに参加された経験談を交えながら、構築されたモデルや最終結果についてお話しいただきたいと思います。最後に、ＱＡセッションを設けております。

このセッションでは 2020 年度に開催された自動車保険に関するデータ分析コンペティション「Insurance Pricing Game」の参加者から、コンペの概要・採用したモデルと結果について発表したのち、アクチュアリー業務におけるモデリング手法の活用の可能性やコンペの参加等についてディスカッションを行うというものになっています。このコンペは、AIcrowd という団体が運営したものです。

ここ最近、データサイエンスや機械学習にますます注目が集まる中、日本アクチュアリー会の会員で Kaggle Master や Kaggle Expert、また当該分野の第一線でご活躍されている方も少なくないと思います。そのような中、私自身、データサイエンスに興味はあって、いろいろ手は出しているのですけれども、コンペに関しては今回の「Insurance Pricing Game」が初めてでした。コンペ初心者としてのフレッシュな視点や、実体験について皆様に発信してシェアすることは、意義があると思いますし、また、このコンペ自体も世界的かつ大々的に開かれた保険アクチュアリー系コンペでしたので、データ分析コンペティションとアクチュアリーを議論する上ではうってつけのテーマだと考え、このセッションを開催する運びとなりました。

さて、ここでご視聴いただいている皆様に、Slido を通して、いくつかご質問させてください。まず「データ分析コンペティションに興味はありますか？」



もちろん皆さん、「はい」を期待しているのですけれども、ありがとうございます。50 人以上の方に、投票いただきました。90%以上の方は興味があるということでした。

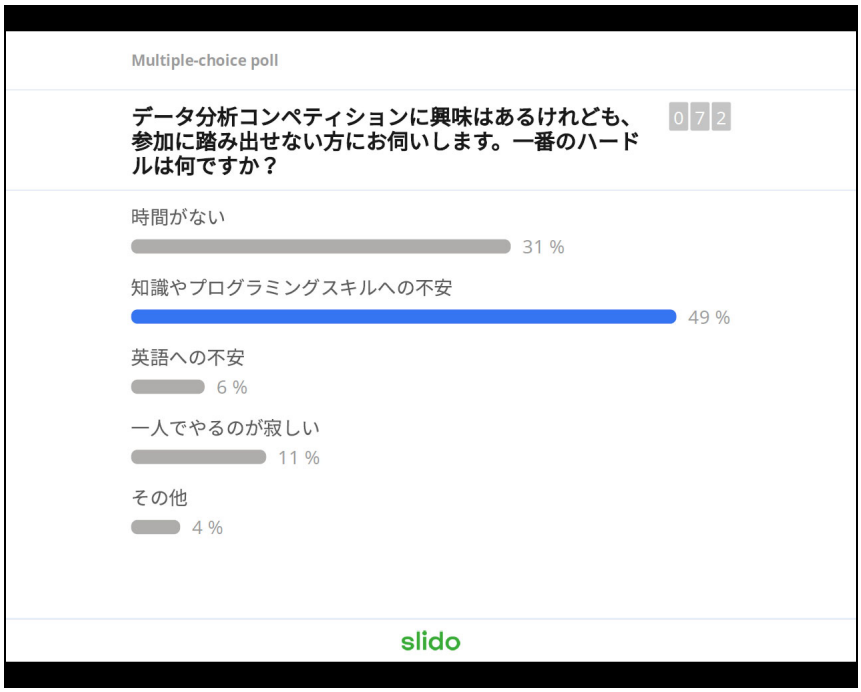
続いて、「データ分析コンペティションに参加したことはありますか？」という質問です。

「いいえ」が８割９割程度になってきました。これは、面白いですね。皆様の周りも、このぐらいの分布でしょうか。

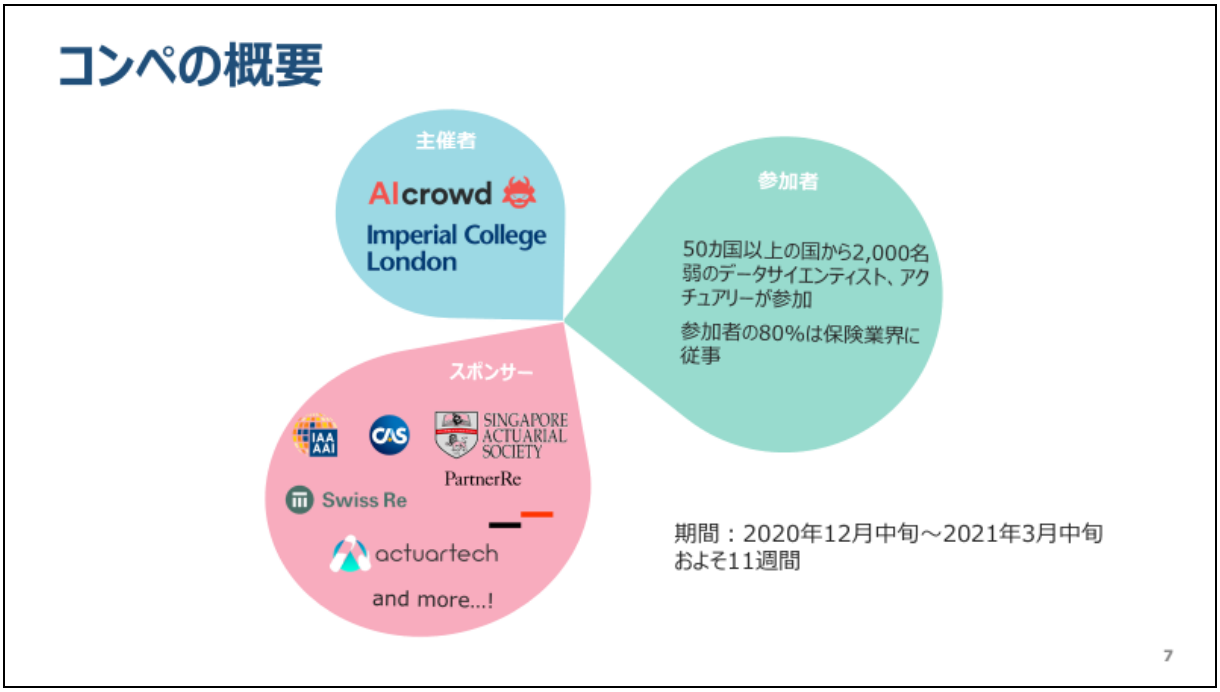
崎村　思ったより、参加していらっしゃる方が多いですね。

藤田　そのような印象ですか。とはいえ、やはり「いいえ」が多数派ですね。70 人近くの人に投票いただきました。「いいえ」が 87%、「はい」が 13%という結果になりました。

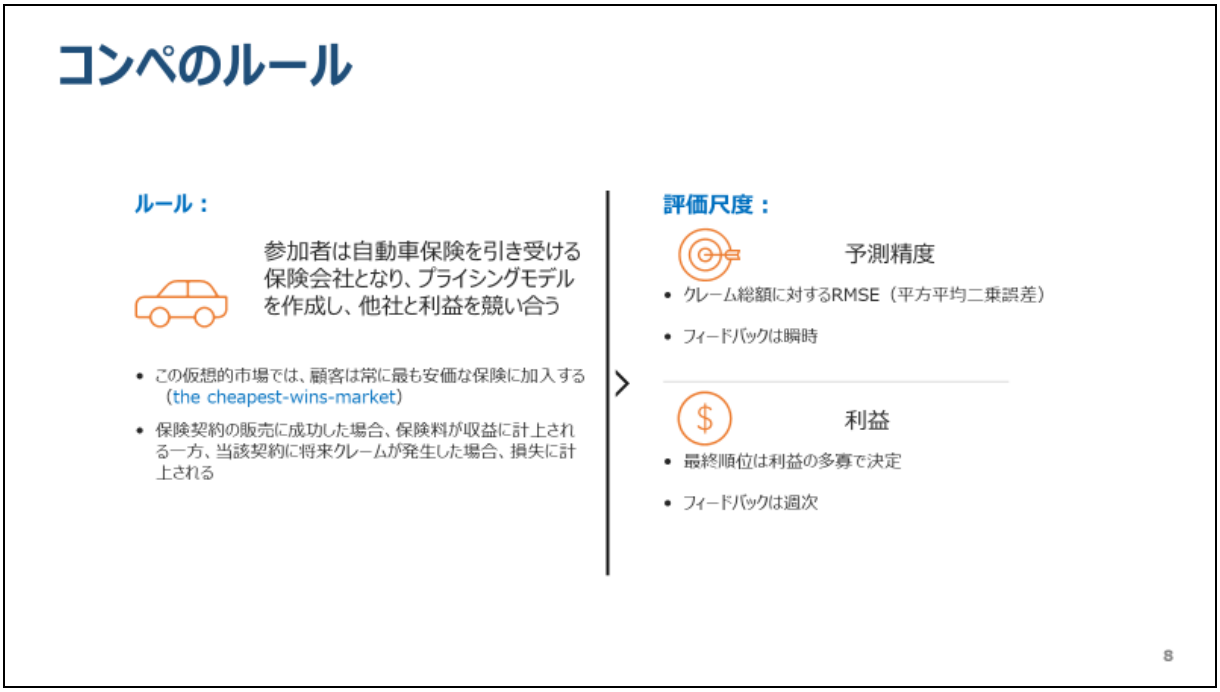
では、もう一つだけ、ご質問させてください。「いいえ」の方々が多かったので、気になる点です。「データ分析コンペティションに興味はあるのだけれども、参加に踏み出せない方にお伺いします。一番のハードルは何ですか？」選択肢は「時間がないから」、「知識とかプログラミングスキルへの不安があるから」、「基本的に１人でやるのが寂しくて、仲間が欲しいから」、そしてコンペは基本的に英語で開かれているので「英語への不安があるから」、最後に「その他」、です。



「知識やプログラミングスキルへの不安」というものが最多です。2 番目は、「時間がない」。確かに、普段の業務等やりながら、休日や時間外のところをどう工面するのかということが一つの課題になりますね。この質問も 70 人近くの方に投票いただいて、非常に興味深い結果が出てきました。こちらの結果は後ほどパネルディスカッションで、是非参考にさせていただければと思います。



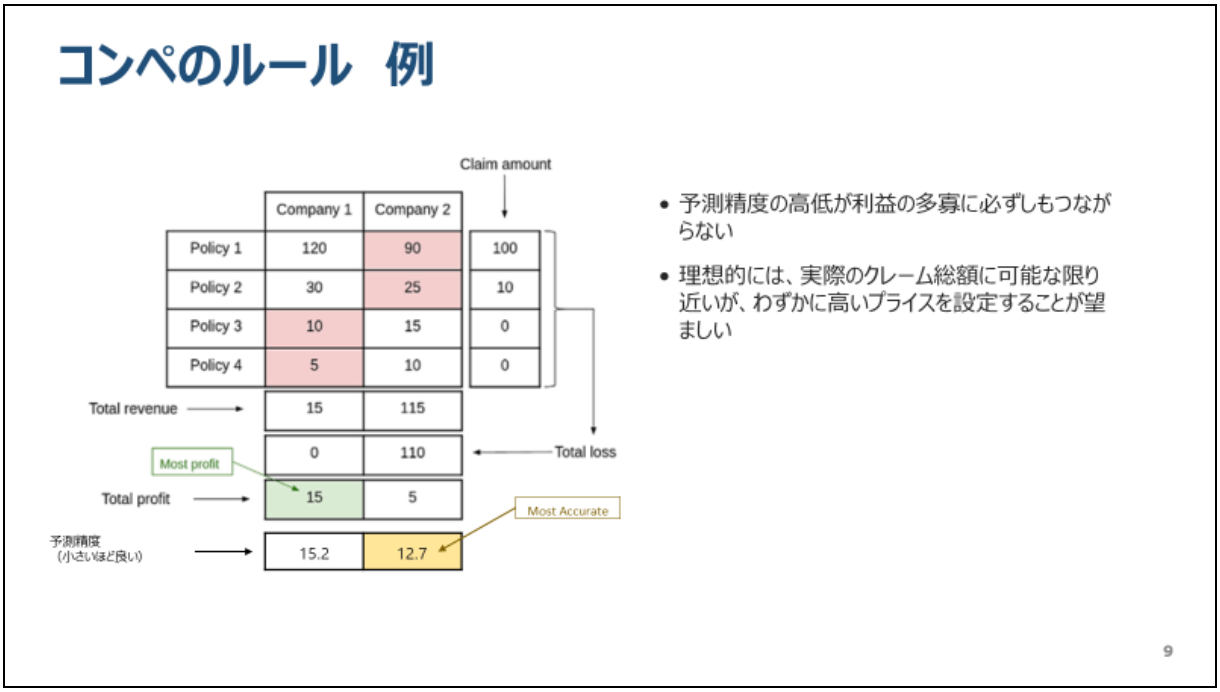
では、「Insurance Pricing Game」について、解説させていただきます。データ分析コンペでは、企業などが抱えているビジネスや研究上の課題をテーマとしたタスクが取り上げられて、参加者は主催者により投稿されたデータを分析し、様々なモデルを構築して競い合います。提出されたモデルや結果は、予め定められた指標により順位付けされて、コンペによっては、上位の人に賞金やタイトルを付与するものもあります。このコンペは、AIcrowd と Imperial College London という団体が運営主体となり、IAA（国際アクチュアリー会）や CAS（米国損保アクチュアリー会）などのアクチュアリー団体、また、スイス再保険等の再保険会社、インシュアテック企業である actuartech などがスポンサーとなり開催されました。参加者は 50 カ国以上の国から総勢 2,000 名弱集まり、その多くは保険業界で働くアクチュアリーやデータサイエンティストでした。期間は 2020 年の 12 月中旬から 2021 年の 3 月中旬、およそ 11 週間に亘り、賞金は総額 1 万 2,000 ドルという規模のものでした。



コンペのルールです。参加者もしくは参加チームは、自動車保険を引き受ける保険会社となります。まず、学習データとして自動車保険のクレームデータが与えられます。それからプライシングモデルを構築して、他のチームとテストデータに対する利益を競い合うことになります。つまり、参加者全員により、仮想的な保険市場が構成されることになります。この市場では顧客は最も安い保険料を提示した保険会社の保険に加入するという the cheapest-wins-market ルールに従います。保障内容や過去の保険会社の実績など、現実世界の保険市場で比較され得るその他の要素は考慮されません。引き受けた保険契約に対する保険料は、全額が収益計上され、契約にクレームが発生した場合は、保険金支払が生じ、損失計上されます。その差額が、当該保険契約の利益となります。

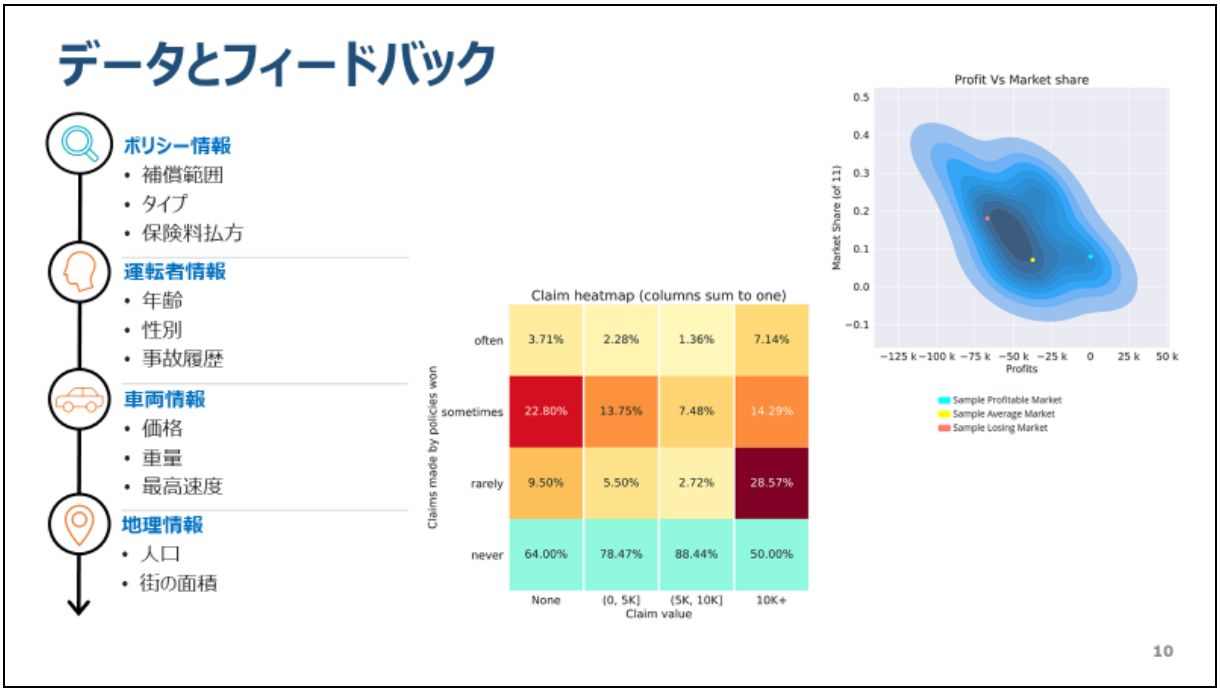
評価尺度は、二つあり、一つはモデルの予測精度を示すもの、すなわちクレームの予測金額と実績金額の RMSE（平方平均二乗誤差）です。もう一つの評価尺度は利益額です。ただし、最終順位は利益額のみで決定されます。

また、参加者は提出したモデルのフィードバックを得ることができます。利益額は週に一度、RMSE については提出の都度、確認することができます。



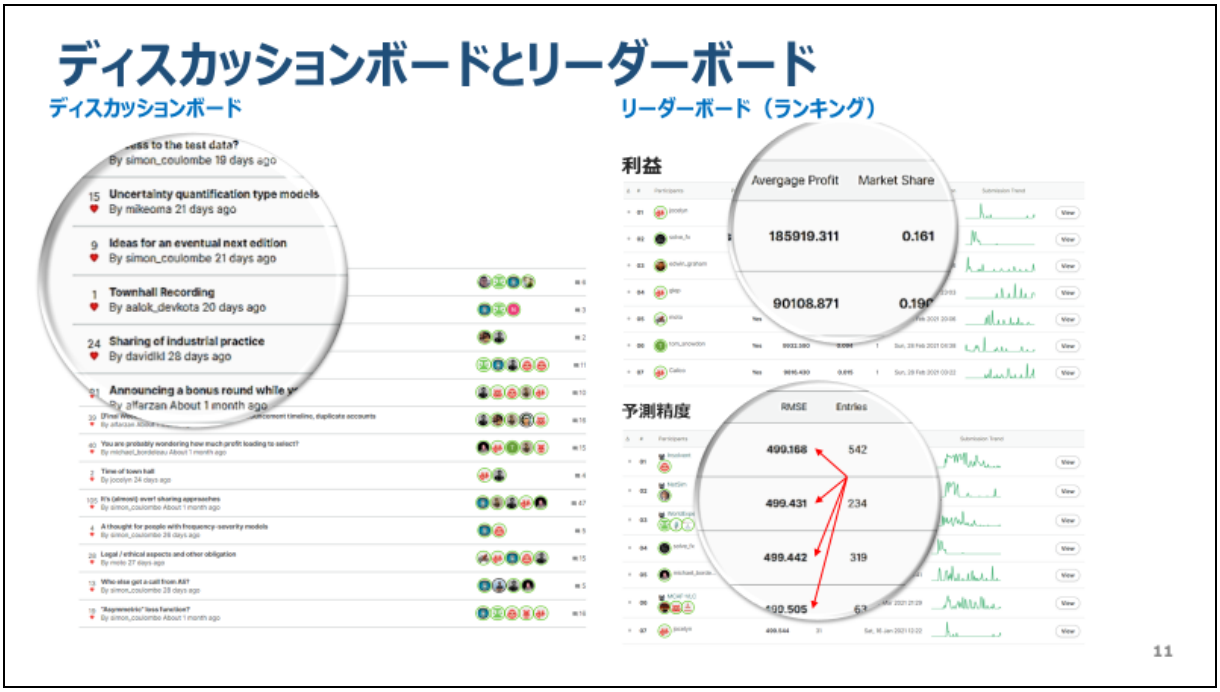
ここで、the cheapest-wins-market について、具体例を用いて解説します。このスライドは、Company１と Company２の２社が４つの保険契約、Policy１から Policy４に対して、プライシングを行った例になります。例えば、一番上の Policy１に着目すると、Company１より Company２の方が安い保険料を提示していることが分かります。したがって、Policy１の顧客は、Company２の保険に加入することになります。そして Company１にはゼロ、Company２には 90 の収入が計上されます。一方で、Policy１には、100 というクレームが発生しました。そうすると、その分 Company２には損失が計上されて、結果的には、Policy１に対する Company１と Company２の利益は、それぞれゼロとマイナス 10 となります。つまり Company１の利益の方が高いことになります。

同様に Policy４まで見ていくと、予測精度は Company２の方が良いのですが、利益は Company１の方が高いため、Company１の勝利となります。つまり、予測精度の高さは利益の大きさに必ずしもつながらないことを示す例になります。上位を狙うためには、保険契約ごとの期待クレーム額を適切に見積もって、まずは、保険契約を勝ち取ることができるような競争力のあるプライスを提示して、その上で、利益を上げる必要があります。



このコンペで用いられたデータは、あるヨーロッパの保険会社の自動車保険契約、約 10 万件を対象として、それらに対する連続した過去５年間のクレームデータを匿名化したものです。データの特徴量は、補償内容等の契約情報、運転者の年齢や性別等の運転者情報、車両価格や車両重量等の車両情報、契約者が住む土地の人口や面積等の地理情報を含む数十個のもので構成されています。コンペの規約上、これ以上の詳細は話すことができませんが、ご承知おきください。

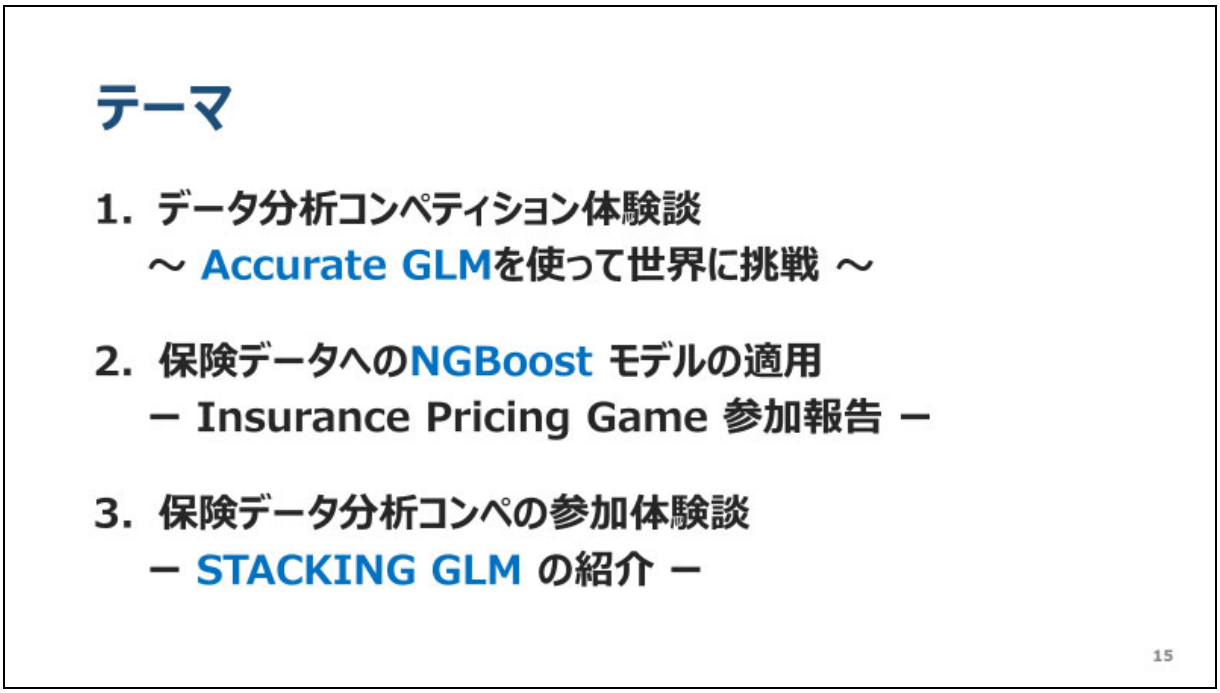
またスライドの右側に示すように、提出したモデルのパフォーマンスに関するフィードバックが毎週得られます。例えば、右上は参加した市場における利益および市場シェアの分布、中央下図は横軸にクレーム額のレンジ、縦軸にはその引き受けた度合いを表したヒートマップ、それ以外にも種々の KPI などが得られます。これらをもとに、翌週の戦略を考えていくことになります。



また、ディスカッションボードが設置されており、コンペのルールやデータに関する疑問点の解消、モデリングに関する議論を行うことができます。そして、リーダーボードでは、自身の現在の順位を把握したり、他のチームの状況を見ることができます。

こうして、およそ 11 週間という時間が過ぎて、最終締切日を迎え、そこから約２週間後、本コンペに関するタウンホールがオンラインで開かれました。そこでは、最終結果や優れたモデルの紹介、パネルディスカッションなどが行われました。タウンホールは、YouTube 上で公開されていますので、ご興味ある方は、是非、お時間のあるときにご覧ください。

では、私からの説明は以上です。ここからは、パネリストによるプレゼンテーションに移っていきたいと思います。パネリストの方々をご紹介します。まず、アメリカンホーム医療・損害保険株式会社の荒井智弘様、続いて、三井住友海上火災保険株式会社の伊藤健太様、最後に、ガイカーペンタージャパンの崎村沙羅様、久下康太朗様、以上の４名をお招きしております。どうぞよろしくお願いします。スライドに簡単にプロフィールも掲載しましたが、それぞれの一番下に着目ください。皆様この「Insurance Pricing Game」が初めてのコンペ参加でしたので、初参加の感想や、どのようなものだったのかという経験談を、共有していただければと思います。



まず、荒井様から「データ分析コンペティションの体験談～Accurate GLM を使って世界に

挑戦～」続いて、伊藤様から「保険データへの NGBoost モデルの適用―Insurance Pricing Game 参加報告―」という題でご発表いただきます。最後に「保険データ分析コンペの参加体験談―STACKING GLM の紹介」を﨑村様からご発表いただければと思います。

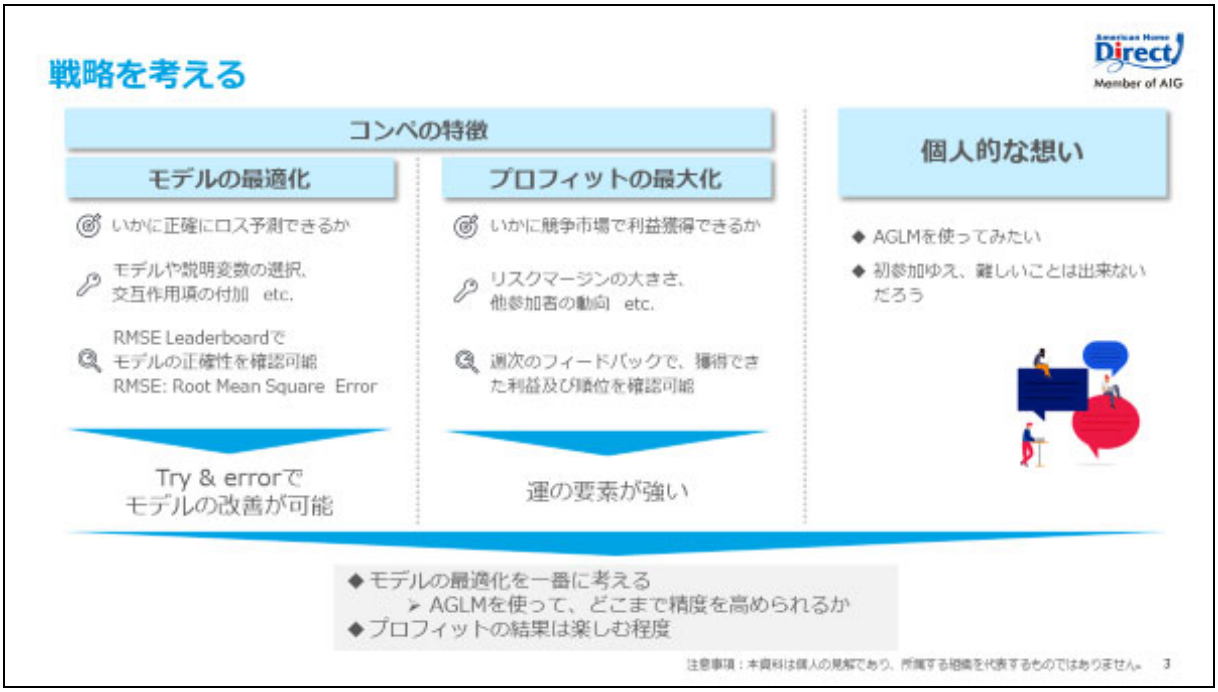
それでは、荒井様、よろしくお願いいたします。

荒井　皆さんこんにちは。アメリカンホーム医療・損害保険株式会社の荒井と申します。よろしくお願いします。私からは、データ分析コンペに今回参加してみて、その経験談などをお話しさせていただければと思います。

まず、データ分析コンペに参加したきっかけですが、副題にも書きましたとおり、Accurate GLM に興味を持っていたことでした。Accurate GLM 自体は、オーガナイザーの藤田様が 2018 年の年次大会で発表された内容でして、当時、とてもいいモデルだなと思って聞いていていたところ、ちょうどいいコンペがありましたので、これを機に参加してみようという経緯です。実際、AGLM は、とても使いやすく、コンペ初心者であったとしても、それなりに精度がいいモデルを作ることができたので、その点を中心にお伝えできればと思っております。

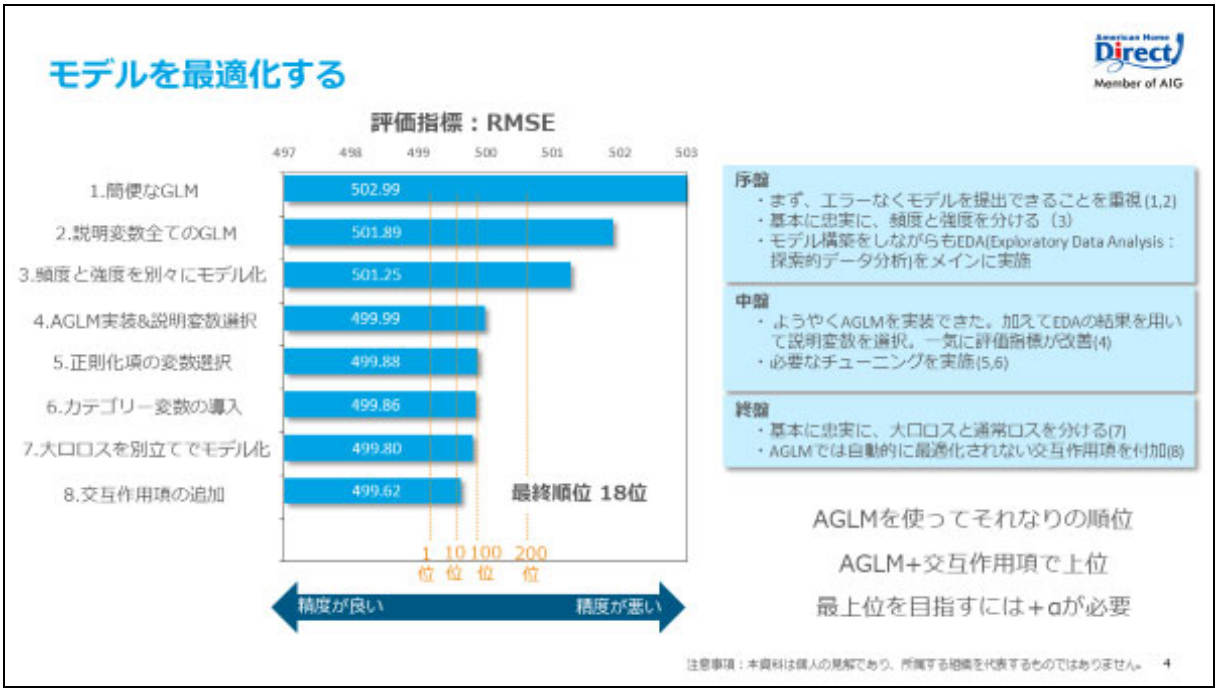
まず、私自身が初参加でしたので、初参加者なりの不安や対応など、そのようなところをまとめました。実際、データ分析のコンペの Web サイトを見ると、そもそも何を求められているのか、何を提出したらいいのか、それらを把握するだけでも結構大変なことです。さらに中身を見ていくと、XGBoost など格好いい言葉がたくさん並んでいて不安にもなりますし、自分自身のスキルも、R や Python を多少は使っているけれどもといったレベルでしたし、いろいろな障壁があって、なかなか踏み出せない分野だと思っております。

しかし、実際に参加してみて感じたことは、そこまで思ったよりハードルが高くないということです。まず、データを提出するということが最初のハードルですが、コンペ主催者もいろいろな人にコンペに参加してもらいたいという思いがあると思いますので、チュートリアルがとても充実しています。なので、簡単なモデルを提出するだけに限って言えば、それほど苦労はなかったように思います。また、XGBoost などそのようなモデルを知らなくても、アクチュアリーが慣れ親しんでいる一般化線形モデル、その発展系の AGLM を使えば、十分世界で通用するということも分かりました。サイト内でのディスカッションボードに R や Python などのコードを公開している方もいらっしゃるので、それも非常に勉強になりました。参加した結論としては、とても有意義で、良かったと思います。なお、今回のプレゼンのテーマである AGLM について詳しく説明することは省きますが、簡単に言えば、一般化線形モデルに「離散化」、「O ダミー変数」、「正則化」、という三つのコンセプトを組み込んだモデルになります。



では、早速内容に入っていきます。まず、戦略について、です。藤田様のイントロダクションにもあったとおり、このコンペの特徴は二つに分かれると思います。一つが、保険金のロスをいかに予測できるかというモデルの最適化の問題、もう一つが、保険料にマージンを乗せて、どのぐらい利益が取れるかというプロフィットの最大化の問題になります。一つ目のモデルの最適化にあたって RMSE という評価指標で正確性を順位付けしていくのですが、RMSE はモデルを提出すると直ちにフィードバックが得られます。なので、何度も Try & error を繰り返すことで、モデルを改善することが可能です。一方で、プロフィットの最大化に関しては、リスクマージンをいかに乗せるか、また、他の参加者がどのような料率設定をしているのかなどにも依存しますので、運の要素が強いという特徴があります。

加えて、AGLM を使ってみたい、初参加なので難しいこともできないだろうなどといった個人的な想いもあり、王道路線を取ることにしました。つまり、スライドの一番下に書いたとおり、モデルの最適化を一番に考えよう、ということです。なので、AGLM を使って、どこまで精度が上げられるのか挑戦してみることにしました。一方、プロフィットの最大化の方は、フィードバックを見ながらマージンを決定して、楽しむ程度で考えていました。

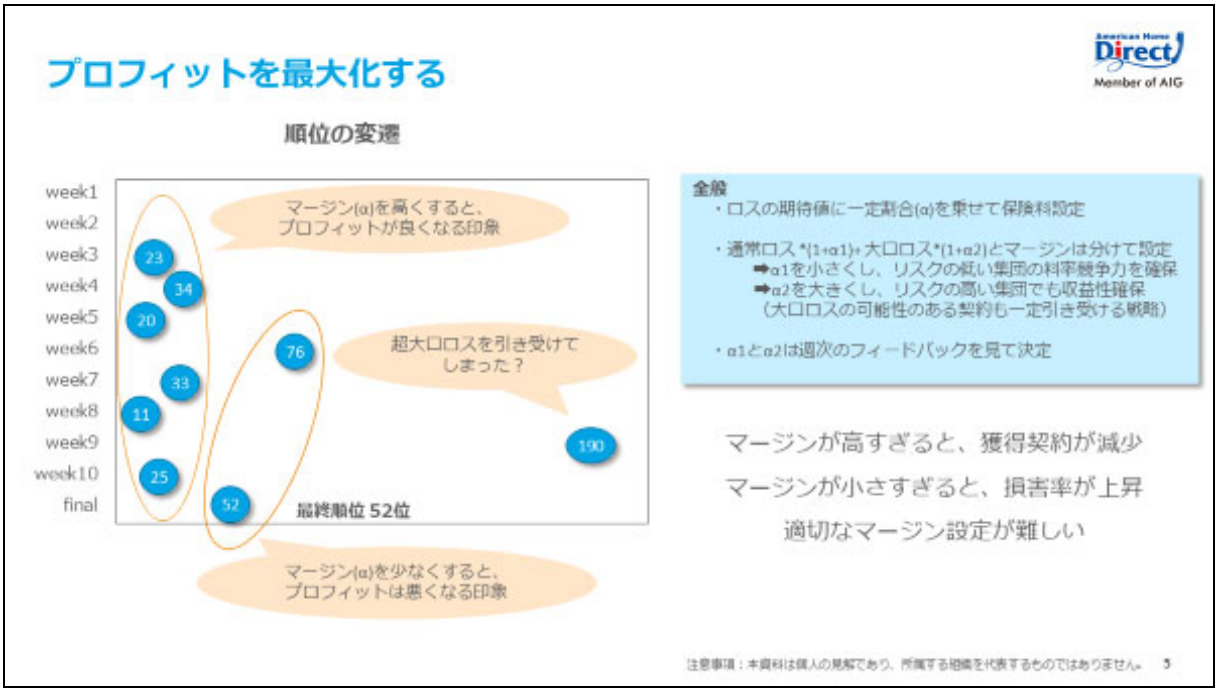


このスライドは、今回、自分がメインテーマにしたモデルの最適化について、です。左の図が、作成した各モデルに対する評価指標を表しています。ただ、必ずしもこの順番でモデルを開発していったわけではなく、どの改善指標で、どのぐらいの改善度合いがあったかということを、コンペ後に分かりやすくまとめたものになります。したがって、例えば、最後の８番で交互作用項の追加をしていますが、その際に、４番にある説明変数を調整したりなど、行ったり来たりしている部分もありますので、ご承知おきいただければ思います。

モデルを作る過程では、やはりエラーなくモデルを提出するといったところが最初は一番難しいところでした。まずは簡単な GLM や頻度、強度を分けたモデルを作りつつ、EDA と言われるデータ分析などを中心に行っておりました。その後、ようやく AGLM を実装することができたのが４番のあたりで、評価指標がかなり改善しています。この３番から４番の評価指標の改善からも AGLM の強さや、使い勝手の良さが、分かっていただけると思います。

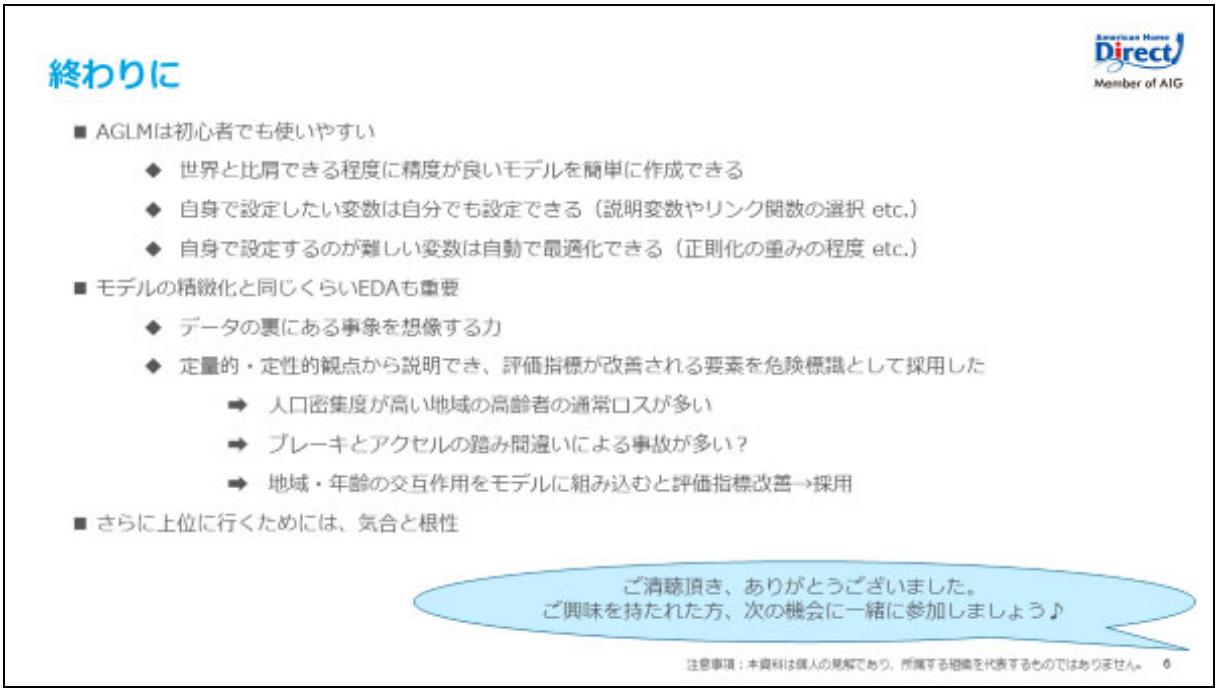
AGLM の良さは三つの要素がバランスよく組み込まれているところだとは思いますが、私自身の考察としては､O ダミー変数化のところが一番強力ではないかと思います｡O とは Ordinary の頭文字で､順序を意味しています｡自動車保険の場合､年齢によってロスの変化が大きくて、若年層はロスが高くて、中年層はロスが低くて、高年層がまた高くなってと、そのように曲線を描きます。そのような特性をどうやってモデルで表現するのかというところが、結構悩みどころです。AGLM の O ダミー変数のような、順序の要素を含んだモデルを使うことで実態をうまく表現できるのだろうと思っています。

AGLM の実装後も、パラメーターのチューニングをしていきましたが、どうにもこうにも評価指標があまり良くならず悩んでいたところ、とある恩師の方から、「AGLM の場合、交互作用項を自分で判断して入れないといけないですね」というアドバイスをいただいて、交互作用項のデータ分析を行いました。分析した交互作用項をモデルに入れ込んだところ、かなり評価指標が改善され、最終的には値は 499.62、順位では 18 位、初心者としてはそれなりに良い順位が取れたのではないかと思います。まとめると、AGLM を使うだけでも、一定程度の順位になる上、そこに交互作用項を入れると、さらに上位でも戦えるということが分かりました。



続いては、プロフィットの最大化の問題になります。こちらは、ロスの期待値がまずあって、どの区分にどの程度のマージンを含めるかが、勝負の分かれ目になります。個人的には、不確実性に応じてリスクマージンを設定するという、いわゆるリスクベースプライシングをやりたかったのですが、時間が取れなかったため、簡便的な手法を採用しました。具体的には、通常ロスと大口ロスを分けて、それぞれに対して違うマージンを乗せるというような形で挑戦してみました。どのぐらいマージンを乗せるかは、フィードバックを見て決定していました。

左の図が順位の変遷になります。ご覧のとおり、かなり変動が大きい印象です。利益を上げようとマージンを高くすると、契約が減少します。一方で、契約をたくさん取ろうとマージンを少なくすると、損害率が上昇してしまう。モデルが正しいのであれば、損害率の上昇は抑えられるような気がするのですが、実際にはモデルに使うデータと、競争に使うデータが違うという、いわゆるパラメータリスクがありますので、このような結果になったのではないかと考察しています。



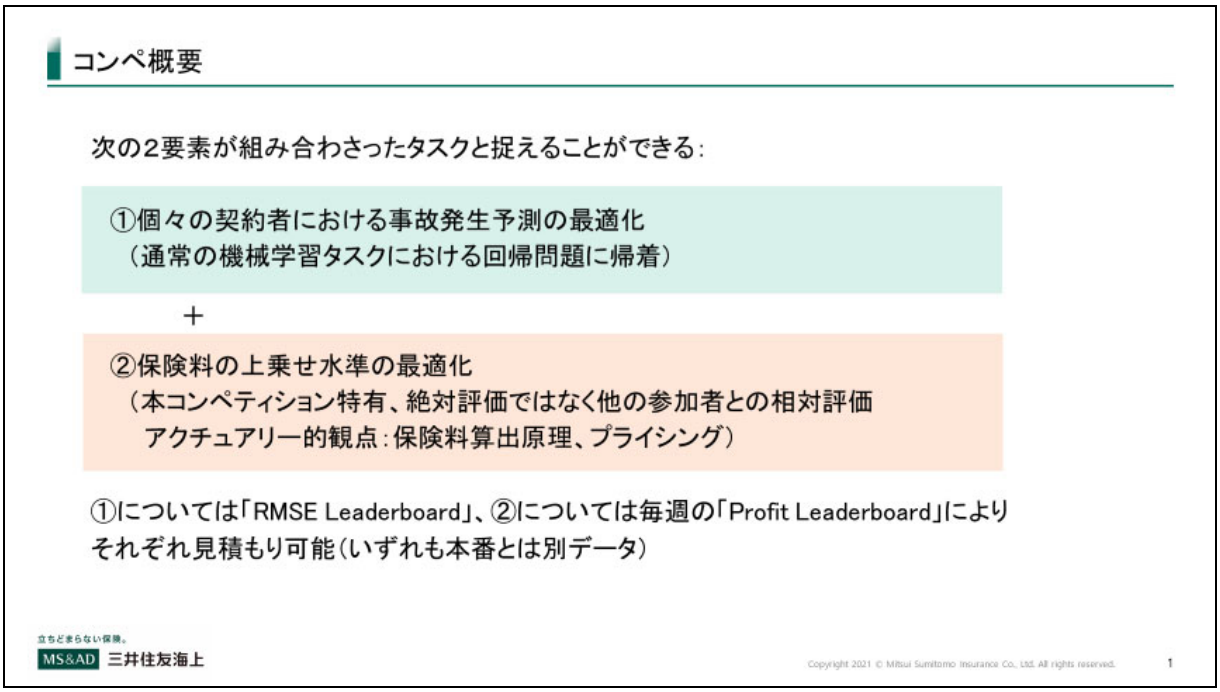
最後に、まとめです。これまでお話したとおり、AGLM はコンペ初心者でもかなり使いやすく、簡単に精度が良いモデルを作ることができます。私自身、AGLM で良いと思うところは、バランスの良さだと感じました。例えば、説明変数を決めるときに、ステップワイズなどを使えば自動で選択することもできますが、今回、私はどの変数を選択するかは自分で行いたいと思いました。一方で、正則化の重みなどは、自分で決定することはなかなか難しくて、AGLM を使って自動で設定することができます。そういった、自分で設定したいところと、自動で最適化したいところが自由に選択できるという汎用性の高さは、AGLM の良いところだと思います。

もう一点感じたことは、モデルの精緻化を進める上で、EDA も重要だと思っています。EDA の定量的な結果に対して、定性的に説明できる要素をモデルに入れ込むというようなプロセスで今回やりました。そうすることによって、モデルの改良度や改良スピードが速くなったかなという印象を受けています。

最後、反省点になりますが、さらに上位に行くためには、より多くのモデルを提出して、評価指標に用いるデータの中身を探りに行く、そのような気合と根性が必要なのかなと感じました。私からは以上になります。ご清聴頂き、ありがとうございました。続きましては、伊藤様の発表になりますので、よろしくお願いします。

伊藤　それでは、私の発表に移らせていただければと思います。先ほどご紹介をいただきました、三井住友海上の伊藤と申します。本日はよろしくお願いいたします。

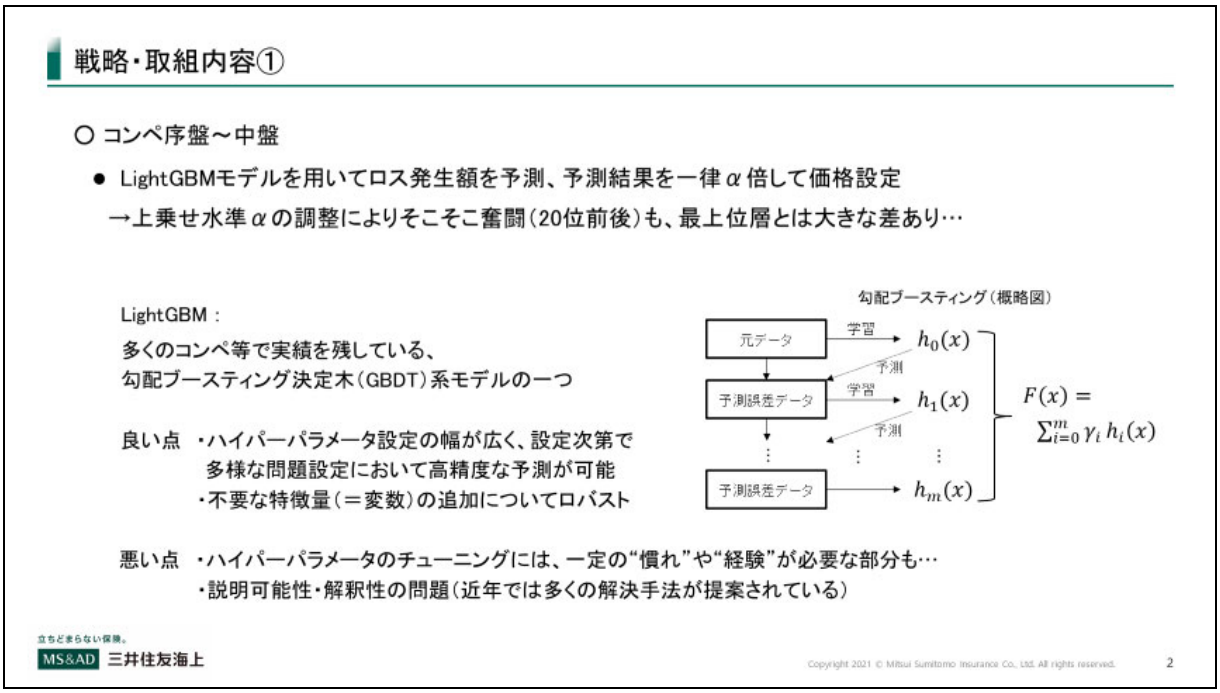
まず、このコンペに参加した経緯をご説明すると、私も、データサイエンスや機械学習といった分野は、流行していることもあり興味を持っていたのですが、コンペとなると、先ほどのアンケートの多くの皆さんと同じように、少しハードルが高いと思っておりました。そのような中で今回、保険のデータを使った、さらにアクチュアリー向けのちょうどよいコンペがあるということで、思い切って参加してみた次第です。私の発表では、コンペに取り組んだ中での戦略や最終的に使ったモデル、結果などについて具体的にお話ししていきたいと思います。



先ほど、藤田様、荒井様からもご説明があったとおりですが、このコンペは自動車保険のプ

ライシングを行うコンペとなります。私はこのコンペを、次の二つの要素が組み合わさったタスクとして捉えました。一つ目は、①個々の契約者における事故発生予測の最適化、各契約においてどのくらい事故が起きるのかを予測することで、これは、典型的な機械学習における回帰の問題に帰着できるかと思います。二つ目は、②保険料の上乗せ水準の最適化で、これが、このコンペティション特有の面白い点ですが、他の参加者よりも利益を出すというところが最終的な評価指標になっているため、①の事故発生予測をもとに、保険料の水準を最適化していくということが必要になります。当然、他の参加者がどのくらい保険料を上乗せするのかということが影響しますので、絶対評価ではなく相対評価になります。また、プライシングという意味では、アクチュアリー的な観点も含まれているかと思います。

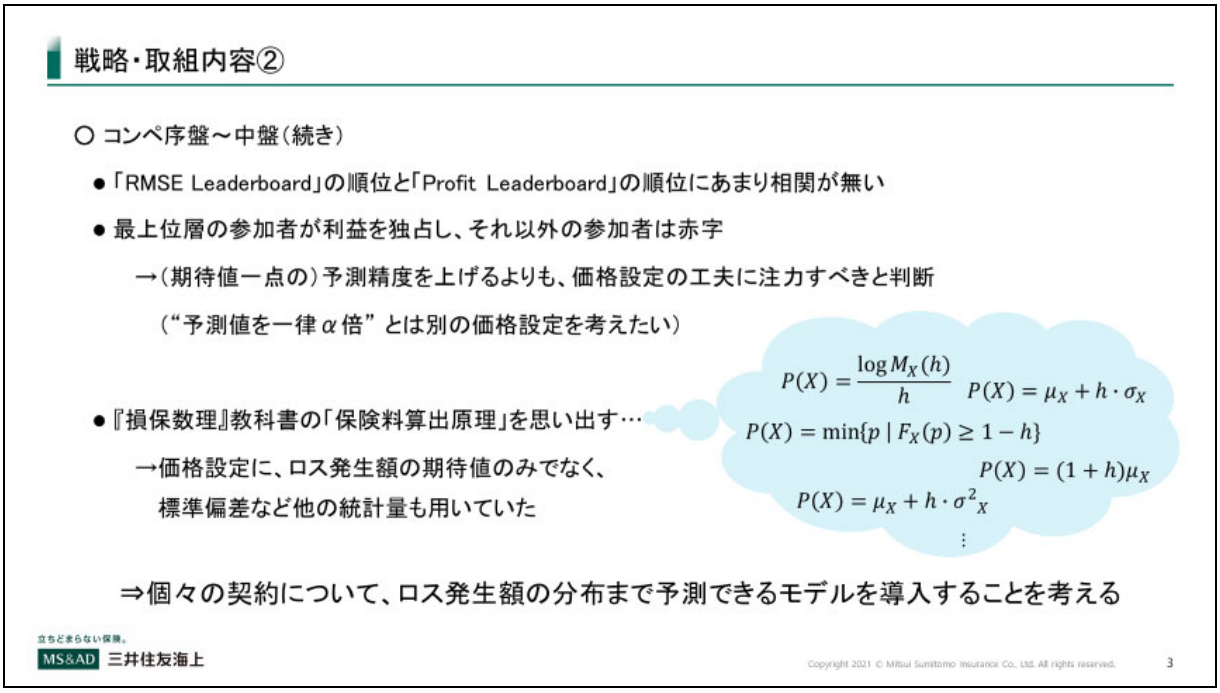
モデルに対するフィードバックとして、①に対しては「RMSE Leaderboard」、誤差の平均の順位表が、②に対しては毎週の「Profit Leaderboard」、利益の順位表があります。これら二つの Leaderboard によって、それぞれ最適化の程度の見積もりが可能となっています。これらのフィードバックを活かしながら、自分のモデルを毎週バージョンアップしていきました。



では、具体的な取り組み内容の説明に移ります。まず、コンペの序盤では、LightGBM というモデルを用いてロス発生額を予測した上で、その予測結果を定数倍して価格設定するという方針で進めました。LightGBM とは、聞きなれない方もいらっしゃるかもしれませんが、多くのコンペ等で実績を残している、勾配ブースティング決定木と呼ばれるモデルの一種です。勾配ブースティング決定木は、決定木を組み合わせたモデルでして、同じように決定木を組み合わせるモデルにランダムフォレストというものがありますが、ランダムフォレストがたくさんの決定木を一斉に並列的に組み合わせるのに対して、ブースティングは、たくさんの決定木を数珠つなぎのように直列的に組み合わせていくモデルと考えていただければよいかと思います。勾配ブースティング決定木系のモデルの利点として、ハイパーパラメーターの設定が大変なものの、設定次第では、多くの問題設定において、特に精度という点においては、非常に高いパフォーマンスを発揮できるということがあります。今回は、先ほどの①の通り、予測精度

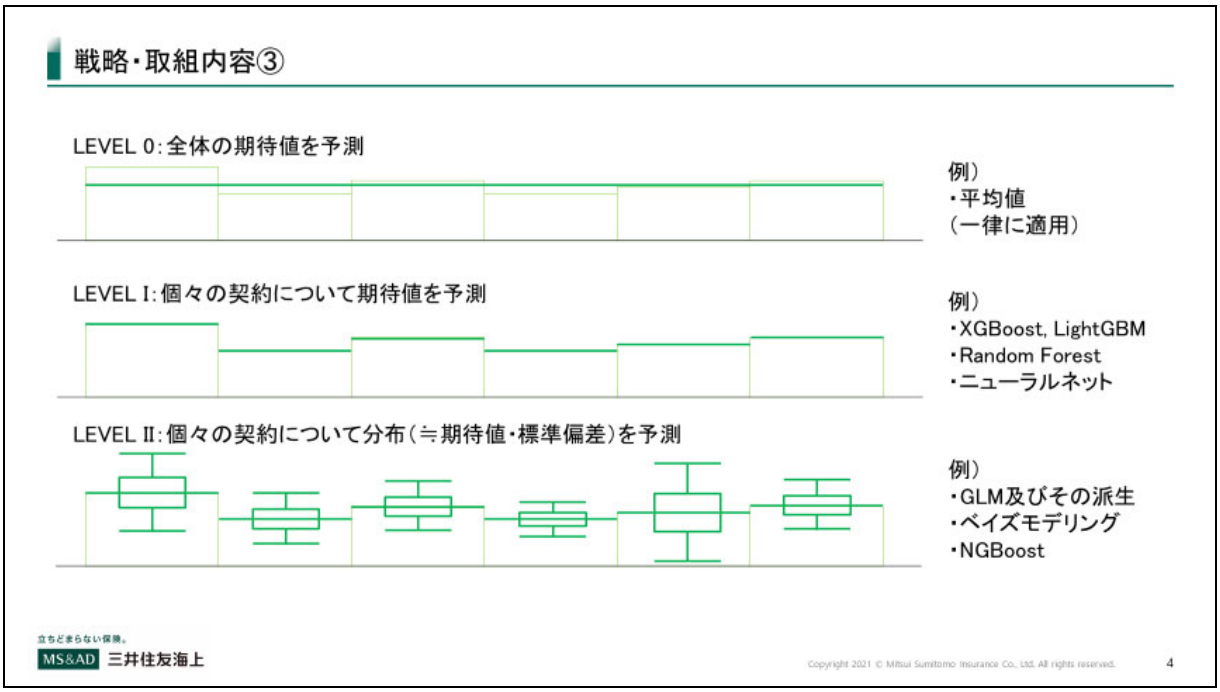
を高めることが一つの重要な目標になりますので、最初のうちはこの LightGBM モデルを用いて取り組みました。

このモデルでは、毎週の「Profit Leaderboard」において、予測結果の定数倍部分を調整することによって、ある程度は善戦できていたものの、最上位の層とは大きな隔たりがありました。より上位に行きたいと思ったとき、何か別のことをしないといけないだろうと思い、次の戦略を考えることにしました。

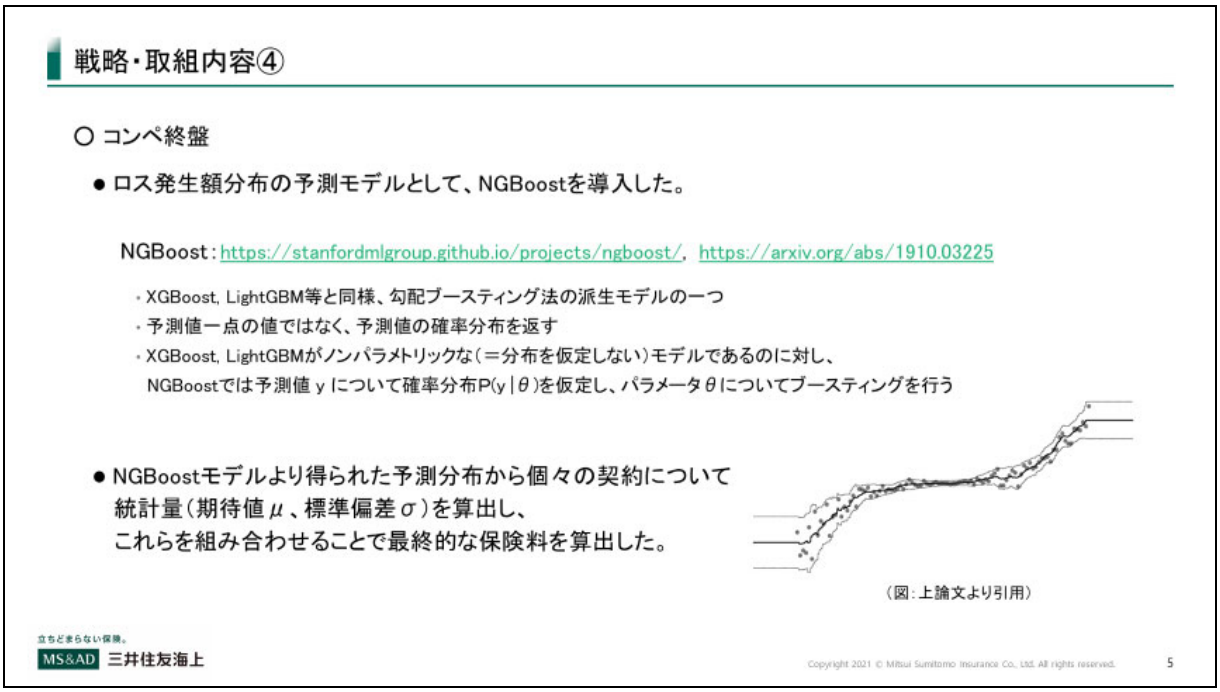


次の戦略を考える中で、いくつかの分析を行った結果、「RMSE Leaderboard」、精度の順位と「Profit Leaderboard」、利益の順位、この 2 つに思ったよりも相関がないということに気付きました。また、週によっては、最上位層の参加者が利益を独占して大きく利益を上げている一方、それ以外の参加者はほとんど赤字、ということもありました。

これらを踏まえ、予測精度を追求していくこともちろん大事ですが、そちらの方面で頑張るよりも、価格設定の工夫に注力すべきだろうと判断しました。特に、予測結果を一律定数倍するのとは別の価格設定方法を考えたいと思いました。そうした中で今回は、アクチュアリー向けのコンペということもあり、アクチュアリー会、損保数理の教科書の、保険料算出原理の章を思い出しました。保険料算出原理では、価格設定の際にロス発生額の期待値だけではなく、例えば、標準偏差や%値などといったような、他の統計量も用いていたかと思います。これを踏まえ、私のプライシングモデルにも、個々の契約についてロス発生額の分布まで予測できるようなモデルを導入して、その分布から標準偏差などの統計量を計算し、これも価格設定に反映させることを考えました。

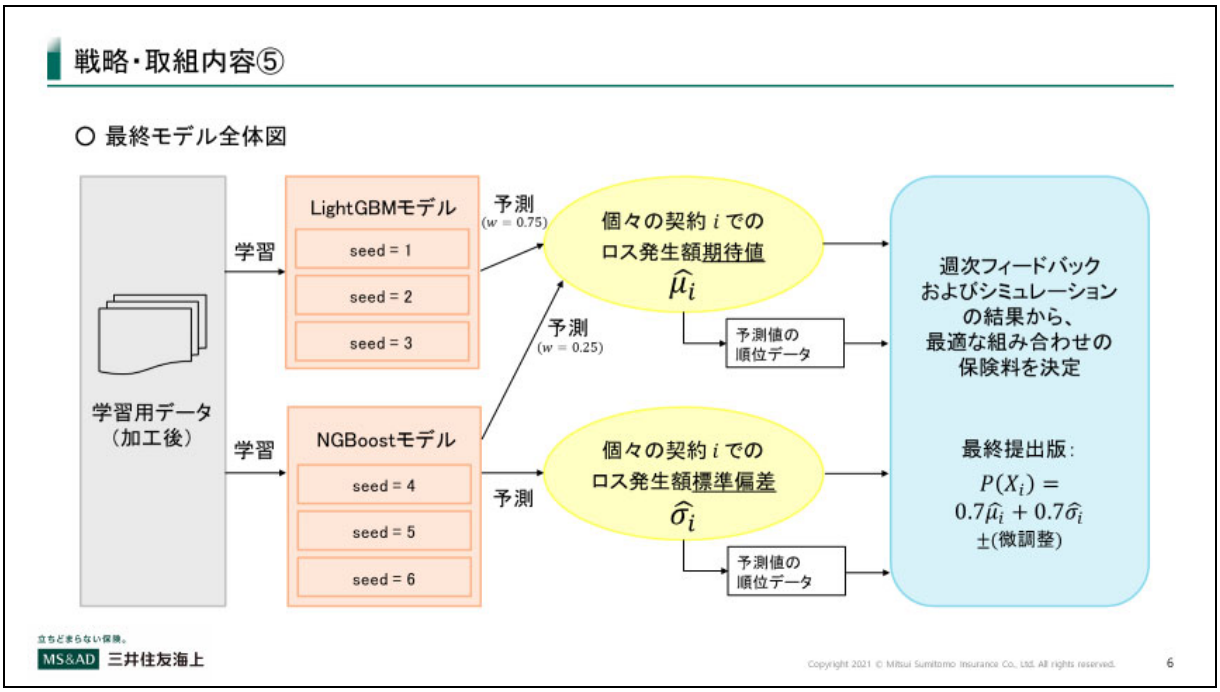


より直感的に、こちらの図で考え方を説明させていただければと思います。まず、一番上の最も単純なモデル（LEVEL 0）として、全体の期待値を一律に予測するモデルを考えることにします。そこから１歩進めたもの（LEVEL I）として、個々の契約についてそれぞれ期待値を予測するようなモデルが考えられます。これは、先ほどご説明した LightGBM や、スライドの右の例に書いてあるようなモデルも含まれるかと思います。今回私がやろうとしたことは、さらにもう一段階進めて（LEVEL II)、個々の契約について分布まで予測できないかということです。個々の契約について分布を予測したいとなったときに、いくつかの方法が考えられますが、例えば、誤差の分布の仮定を残したまま GLM をやったり、その派生形であったり、あるいは、ベイズモデリングなども候補になりますが、今回は、NGBoost というモデルを使って、これを達成しようと試みました。

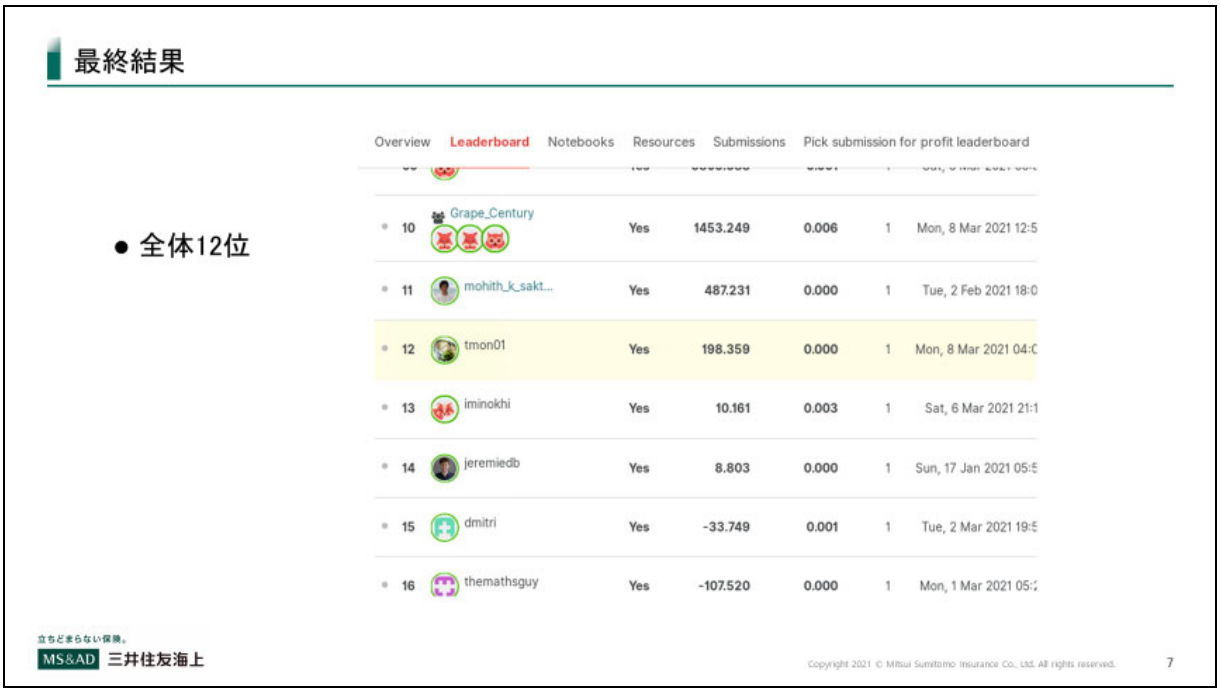


いまご説明した通り、コンペの終盤には、ロス発生額分布の予測モデルとして、NGBoost モデルというものを導入しました。これは、XGBoost や LightGBM といった勾配ブースティング系モデルの派生形の一種となります。特徴としては、一点の予測値ではなく、予測値の確率分布を返すようなモデルになっており、まさに今回やりたい目的と一致しています。XGBoost や LightGBM がノンパラメトリック、つまり、分布を仮定せずに予測値 y をそのまま当てにいくようなモデルであるのに対して、NGBoost では、この予測値 y について、まずパラメトリックな確率分布を仮定して、そのパラメーターθについてブースティングを行うモデルになります。2019 年頃に出てきたモデルです。よろしければこちらのリンク先をご参照ください。

スライド右下の図が直感的なイメージに近いのですが、黒い実線が期待値一点の予測値だとすると、そこからの散らばり具合まで予測できるようなモデルとなっています。このモデルから得られた予測値の分布から、個々の契約について、期待値・標準偏差を算出し、これら統計量を組み合わせることで、最終的な保険料を算出しました。

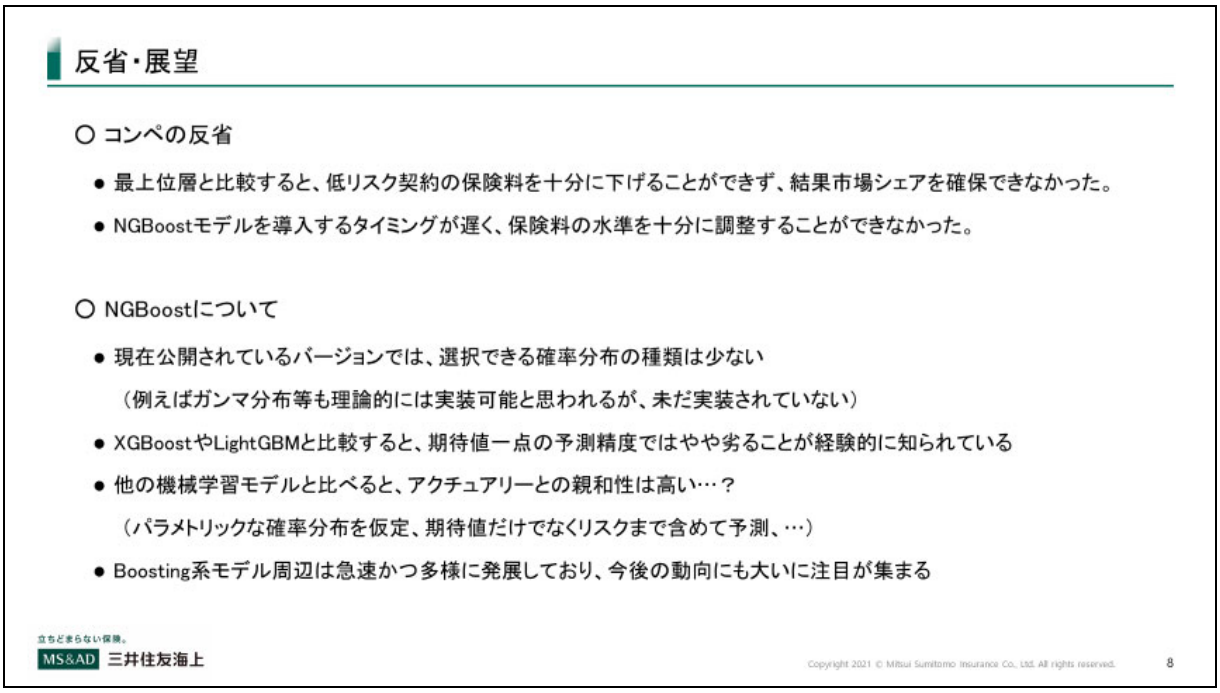


最終モデルの全体図です。いま説明した NGBoost が、ちょうど下半分の赤四角になっており、個々の契約 i でのロス発生額の期待値と標準偏差をそれぞれ予測しています。また、上半分には前半に説明した LightGBM モデルがありますが、こちらは期待値の予測を補強するために用いています。こうして、二つのモデルから予測した黄色の丸二つ、個々の契約 i でのロス発生額の期待値と標準偏差を組み合わせることによって、最終的なプライシングを行いました。また、この組み合わせるときの係数としては、週次の Profit Leaderboard（利益のフィードバック）の結果や、手元でのシミュレーションの結果を踏まえて、最終的には（0.7, 0.7）と設定しました。



この最終モデルを提出し、全体で 12 位という結果が得られました。参加人数が 2,000 人弱

で、うち何人かはチームを組んでいるところで、初参加にしてはまあまあ良い順位が出せたのではないかと思っております。ちなみに、10 位に３人チームが見えますけれども、これが実は日本人のチームで、こちらが次の発表になっております。楽しみにしていただければと思います。



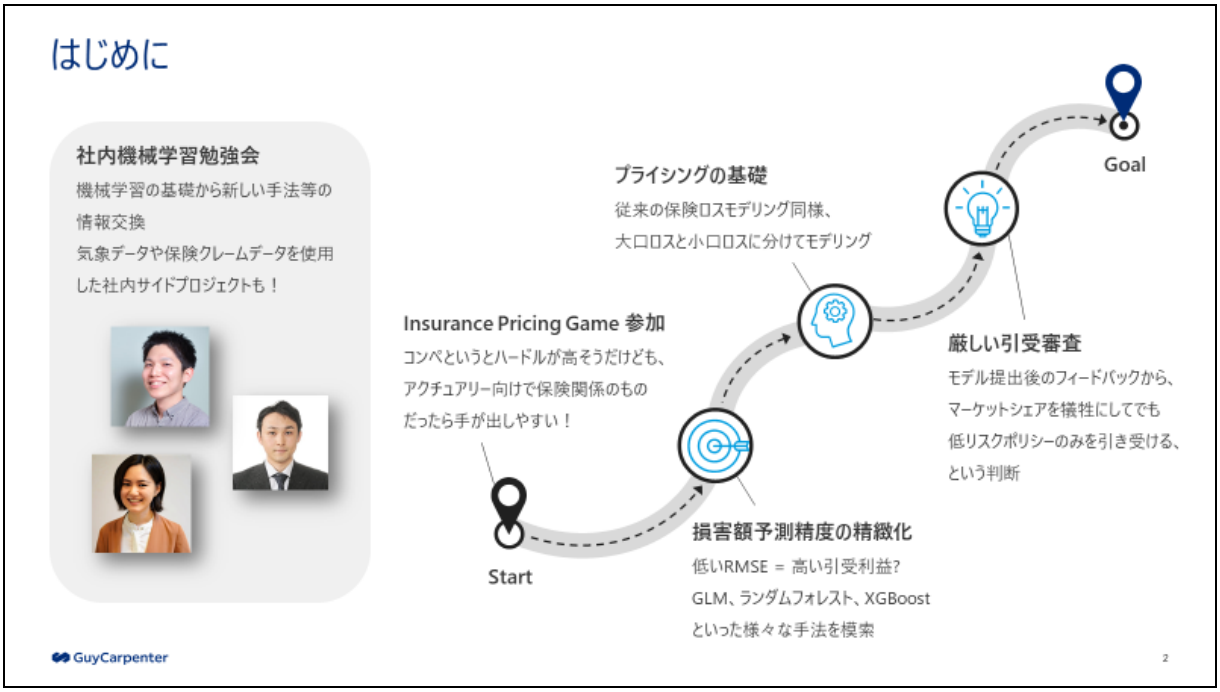
コンペの反省としては、結果論ではありますが、最上位層と比較するとリスクの低い契約の保険料を十分に下げることができなかったため、市場シェアを確保できず、あまり利益を上げられなかったというところがあります。また、途中から先ほどの NGBoost モデルを導入したのですが、導入するタイミングがもう少し早ければ、もっと多くの回数の利益のフィードバックに参加でき、係数の調整などもよりできたのではないかと思っています。

最後に、今後の展望として、特に NGBoost についてコメントしたいと思います。まず、現在公開されているバージョンでは、選択できる確率分布の種類が少ないという点が挙げられるかと思います。理論的には、例えばガンマ分布や、他の保険のモデリングで使われるようなパラメトリックな分布も適用可能と思われますが、現在公開されているバージョンにおいては、実装されている分布の種類はあまり多くありません。また、XGBoost や LightGBM といった、コンペでよく使われているような勾配ブースティングモデルと比較すると、確率分布を経由する分、期待値一点の予測精度ではやや劣るということが経験的に知られています。

ただ、この二つを考慮した上でもなお、われわれアクチュアリーが慣れ親しんでいる、パラメトリックな確率分布を仮定している点や、期待値のみならず、分布、リスクまで含めて予測できるという点を踏まえると、他の機械学習モデルと比べたとき、相対的にアクチュアリーとの親和性は高いと言えるのではないかと考えています。

最近はディープラーニング系のモデルもますます台頭してきており、Boosting 系モデル周辺は少し引け気味にはなっているのですけれども、それでも依然として、急速かつ多様に発展している分野でないかと思っており、今後の動向にも大いに注目が集まる分野だと考えています。これで、私のパートの発表を終わります。

崎村　では、最後に私たちガイカーペンターチームの今回のコンペの体験談とモデルについて、紹介させていただこうと思います。私たちは３人チームで参加をしました。今回が初めてのコンペだったのですが、初心者で初めてでも、これだけ楽しく参加できるのだということを感じながら聞いていただけたらと思います。よろしくお願いいたします。

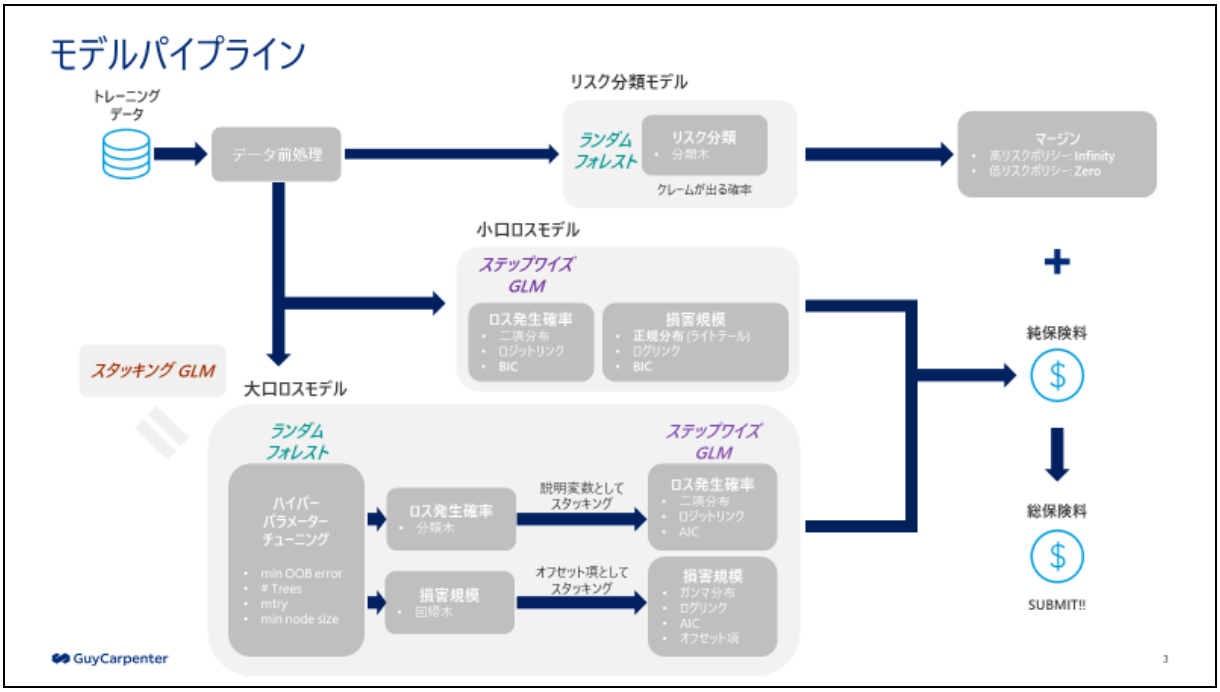


はじめに、私たちのチームについて、紹介します。私たちはアクチュアリーの藤田さん、自然災害リスクを分析されている久下さん、ジュニアアクチュアリーの崎村、の３人で構成されています。われわれは元々機械学習に興味があり、基礎を一緒に勉強したり、新しい手法の情報交換などを行っています。また、持っているデータを使用して、実際にどのような機械学習

手法を使っていけるかといった、社内でのサイドプロジェクトも行っております。

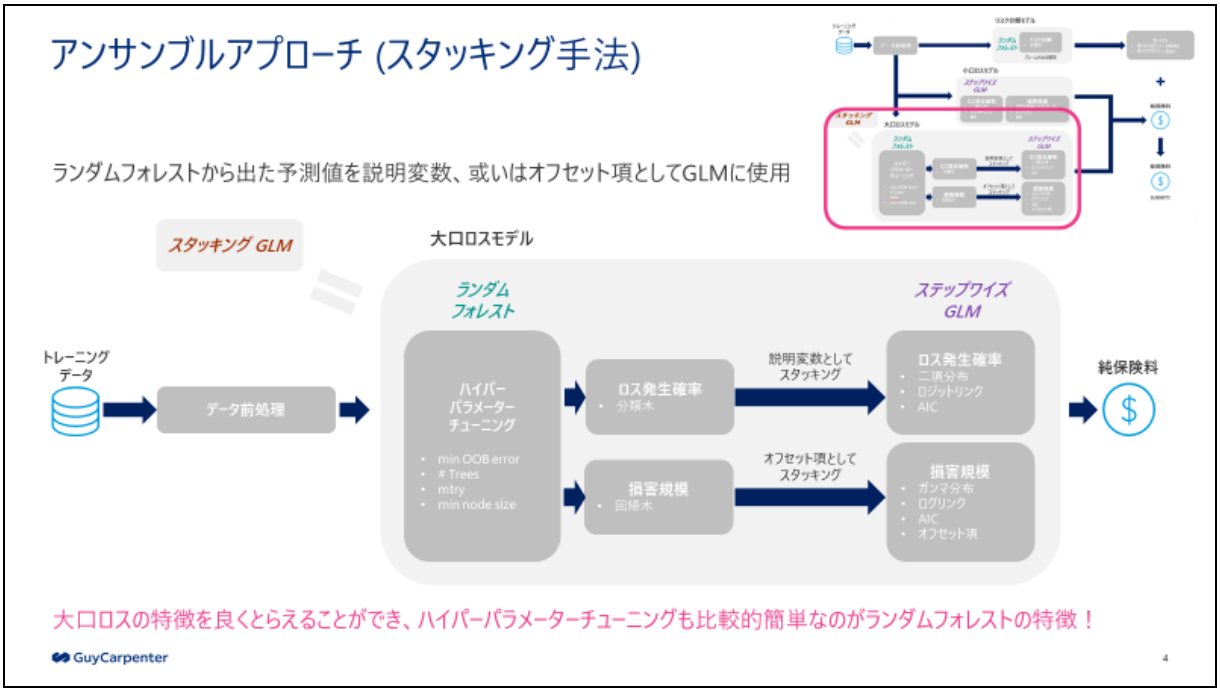
今回の「Insurance Pricing Game」に参加した経緯として、やはり、コンペはなかなかハードルが高く、画像処理といったテーマもどうしても手が出しにくい印象なのですが、今回は、アクチュアリー向けのコンペである上、データも保険関係、かつテーブルデータということで、少し手が出しやすいのではないかということで参加しました。初めは、取りあえず登録してみよう、という感じで始めたのですが、いざ始めてみると、なかなかはまってしまい、とても良い経験になったと思います。

では、どのような流れでモデルを構築していったかを紹介させていただきます。最初は、予測精度にフォーカスしており、GLM やランダムフォレスト、XGBoost といったような様々な手法を使ってみて、どのように予想精度を上げていけるかというところに時間を使っていました。しかし、途中で、保険ロスのモデリングをするときは、大口ロスと小口ロスに分けてモデルしますねということを思い出し、モデルにも反映することにしました。そのあとは、マージンについて考え始めました。毎週の練習試合から得られる順位やフィードバックを見ていますと、意外とマーケットシェアを持っていないチームが上位にいるということに気が付き、そのような方たちは、低リスクのポリシーを引き受けていて、大きなリスクはそれほど引き受けていないのではないかといった想像をしたので、われわれも似たような戦略を取った上で、最終モデルを提出しています。



では次に大まかなモデルのパイプラインの説明です。まず、データの前処理を行ったら、大口のロスモデルと小口のロスモデル、それぞれ別のモデルを作成しています。そこから出てきた純保険料を使用していまして、このモデルとは別に、リスクを分類するモデルというものを作っています。このモデルでは、クレームが発生する確率を予測しており、その予測結果をもとに、高リスクのポリシーと低リスクのポリシーにグループ分けして、低リスクか高リスクかによって、マージンの乗せ方を変えています。では、ここから大きく三つに分けて、もう少し

細かくポイントについてお話しさせていただきます。

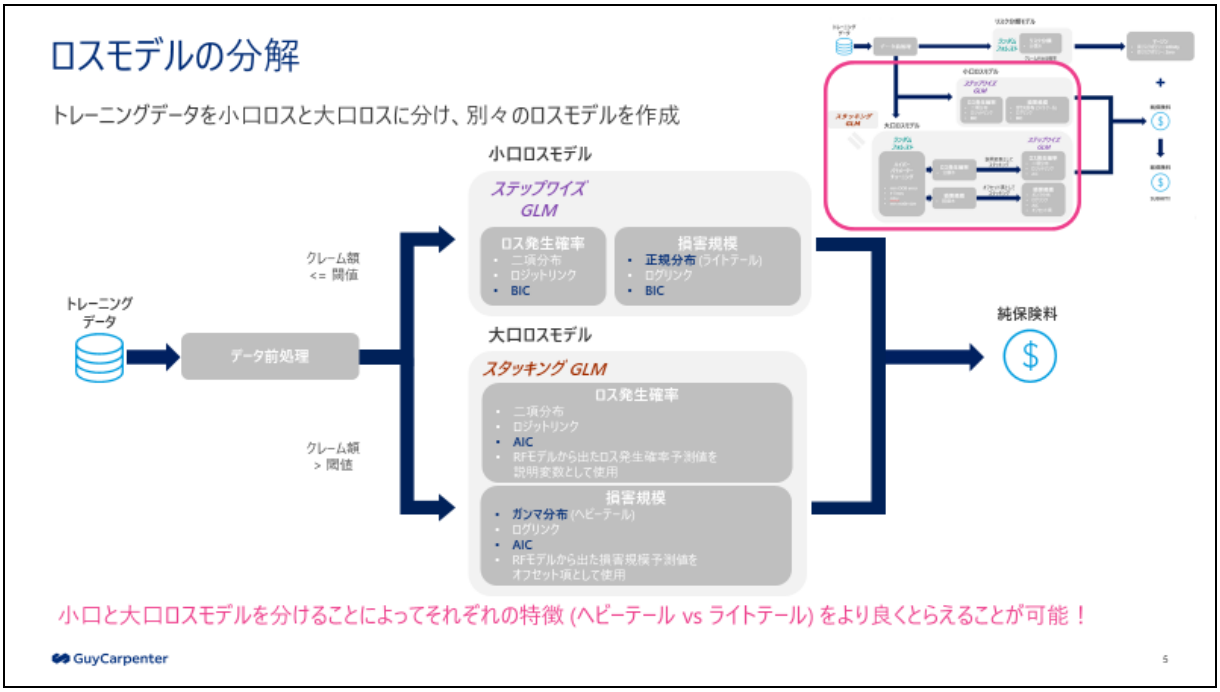


まず一つ目は、アンサンブルアプローチです。アンサンブルアプローチとは、複数の機械学習モデルをミックスしてモデル作成するものです。その中でも特にスタッキング手法という、一つのモデルの予測結果を、他のモデルの中に組み込む手法を用いています。今回、大口ロスモデルを作るときに、このスタッキング手法を使いました。まず、ランダムフォレストを使用したモデルを作成しています。先ほど、伊藤さんに少し紹介していただいた、いくつかある決定木系モデルの中でも、多くの決定木を並列的に考えるモデルです。今回はＦＤ法を使用し、ロスの発生確率と損害規模、それぞれ別モデルを作り、その予測結果を、ステップワイズ GLM の中に組み込んでいます。

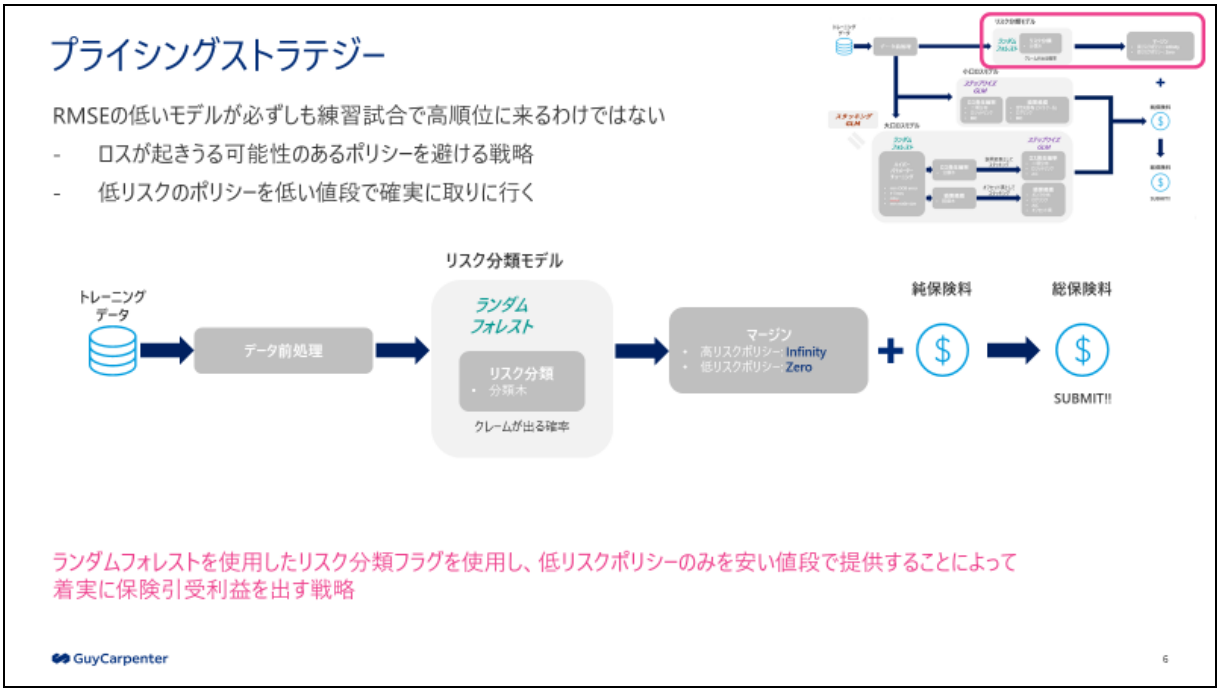
ここの組み込み方なのですけれども、まず、損害規模に組み込んでいます。やはり、大きくてまれな大口ロスの特徴はなかなか GLM のみではキャプチャーすることが難しいため、ランダムフォレストによって、その欠点を補うようなモデルとなっています。ランダムフォレストは、非線形の部分をうまくピックアップしてくれますので、予測結果を GLM の中のオフセット項として使用し、さらに GLM の線形性も使ってサポートするという形で損害規模の予測に使っています。

一方、ロスの発生確率に関しては、せっかくランダムフォレストモデルを作ってみたので、試しに GLM の説明変数の一つとして組み入れてみたところ、なかなか効いてくる変数でしたので、説明変数として、スタッキングしています。

この組み合わせを今回はスタッキング GLM と呼んでおります。このランダムフォレストを組み合わせることで、大口ロスの特徴をうまく捉えるモデルとなっていますし、また、ハイパーパラメータチューニングも比較的簡単なため、今回ランダムフォレストを採用しています。



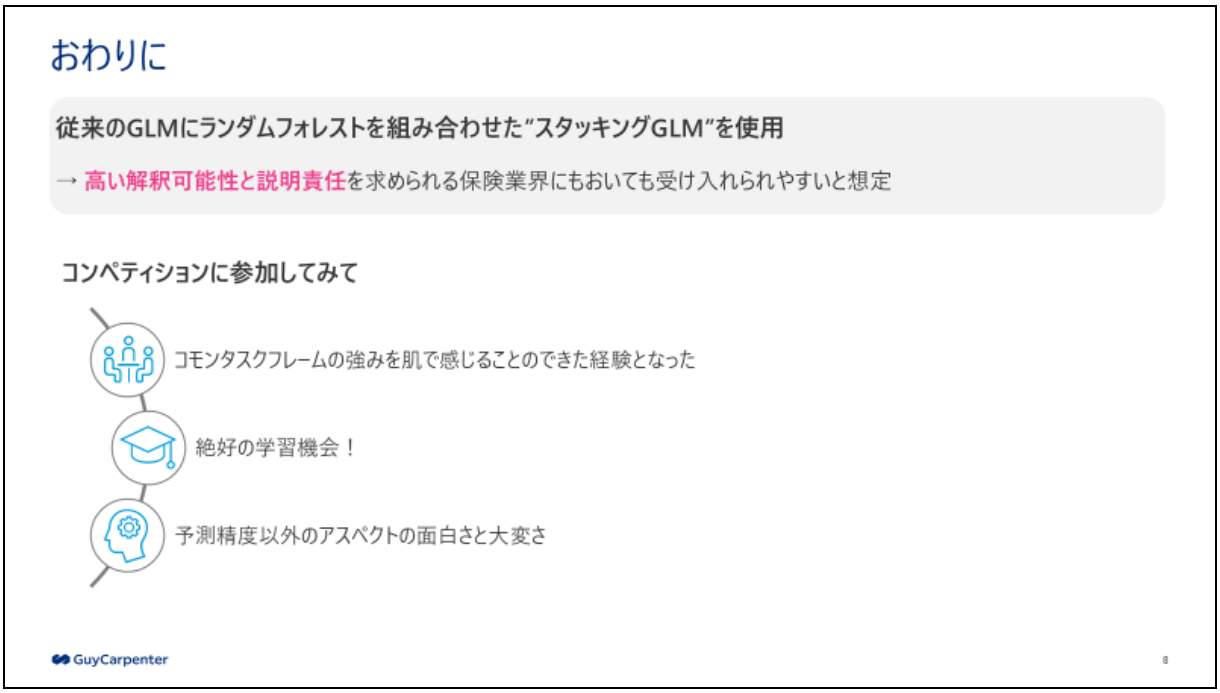
次に、ロスモデルの分解です。先ほど、少し申し上げましたが、小口ロスと大口ロスで異なるモデルを作っています。特に、小口モデルにはステップワイズ GLM を使用し、大口モデルには、先ほど説明した、スタッキング GLM を使用しています。こうすることによって、それぞれの違った特徴をうまく捉えることができるのではないかと考えておりまして、小口の方では、例えば損害規模の正規分布の仮定をおいていますが、大口ロスモデルの方ではガンマ分布といった仮定を置いており、テールが長いものを採用しています。また、どちらでもステップワイズ GLM を使用して、説明変数を自動で選ぶといったプロセスを経ていますが、そのときに、小口ロスモデルでは、比較的シンプルで安定しているモデルを作りたかったため、BIC を指標とした一方、大口ロスモデルでは、あまり説明変数の量を減らしすぎないよう、ペナルティが少し低めの AIC を指標とした、といった点で差異を設けました。



最後に、プライシングの戦略です。RMSE の小さいモデルが必ずしも利益を出しているというわけではなかったこと、また、毎週のフィードバックを見る限り、高順位に来ている人たちが意外とマーケットシェアを取っていないということを踏まえ、われわれのチームも低リスクのポリシーを狙う戦略を取っています。そのために、ランダムフォレストを使ってリスクの分類モデルというものを構築しました。ここではクレームが発生する確率を予測し、その確率を基に、高リスクなものと低リスクなものにグループ分けを行います。ここではかなり保守的に分類しており、ほぼほぼロスが出ないであろうというものだけを低リスクポリシーに入れて、それ以外は全て高リスクポリシーの方に振り分けています。

高リスクポリシーに対しては、無限大のマージンを設定し、「私たちは、このポリシーは売りません」といった意思表示をする一方、低リスクポリシーに対しては、マージンを全く乗せずに純保険料で販売するという少し極端なリスク選択をしています。つまり、ロスが出ないであろうポリシーのみをこつこつ安く売って利益を出すといった戦略を取っています。

このようなモデルを作った結果、全体の中で総合順位 10 位をいただきました。また、主催者から優秀なモデルの一つとして選ばれ、グローバルのタウンホールにて、モデル紹介もしました。タウンホールでは、他のチームのモデルについても聞きましたが、われわれのモデルの特徴として、割とシンプルなモデルだという部分が特に目立っていた点だと思います。他のチームは、10 個以上のモデルを合算したアンサンブルモデルになっていたり、機械学習の手法も 4 個、5 個と組み合わせていたりするチームも多いのです。その中で、比較的シンプルなモデルで十分戦えたというのは、なかなか面白い結果だったのではないかと思います。



最後にまとめです。今回、従来の GLM にランダムフォレストを組み合わせたスタッキング GLM を使用しましたが、やはり、アクチュアリーの皆様が聞き慣れている GLM を使っています

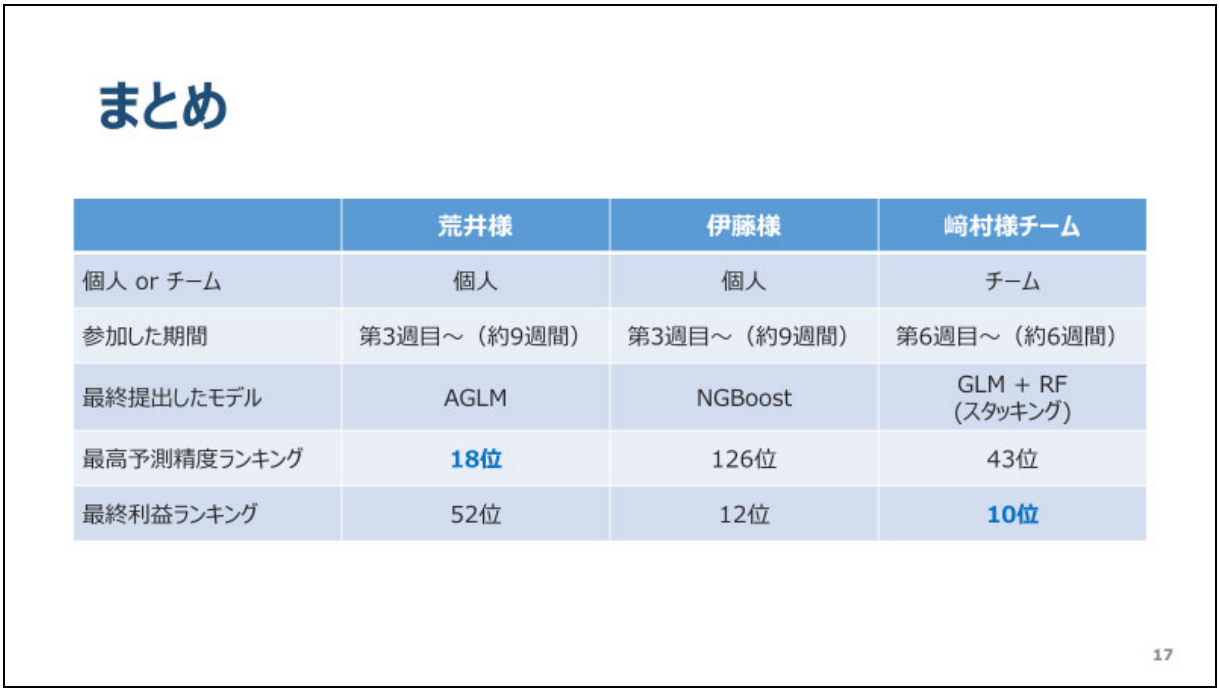
ので、解釈可能性が高いこと、説明責任を求められる保険業界においては、とても受け入れられやすいものではないかと考えています。プライシングの中に機械学習を取り込んでいくということは、なかなか難しいところだと思うのですが、実は、ヨーロッパなどでは、スタッキング GLM のように、GLM の中に少し忍び込ませるというような形で機械学習が使われてきているようで、これからどんどん増えていくのではないかと思っています。

コンペティションに参加した感想として、まず一番にコモンタスクフレームの強みを実感することができました。ディスカッションボードというものがあり、そこでは EDA の結果が出ていたり、あるいは、コードをそのまま貼付けている参加者もいたり、活発なディスカッションが行われていて、みんなで協力していいモデルを作っていきましょうという文化が感じられる場所でした。また、自分にとってもいい勉強の機会になりました。やはり、実際にデータを使って自分の手を動かしてやるということは、机上で勉強することとはまた違った経験や知識を吸収することができるのではないかと思っています。最後に、機械学習というと、どうしても予測精度をいかに高めるかという点にフォーカスしがちなのですが、今回のコンペでは、他のチームと戦うという相対評価の要素が加わり、面白いと同時に、とても大変なところだと感じました。

以上で、私たち、ガイカーペンターチームの発表とさせていただきます。ご清聴ありがとうございました。では、司会の藤田さんにバトンタッチしたいと思います。

藤田　パネリストの皆様、ありがとうございました。大変、興味深く聞かせていただきました。

次のＱＡセッションに移る前に、３名の方々の内容を簡単にまとめたいと思います。



	荒井様	伊藤様	﨑村様チーム
個人 or チーム	個人	個人	チーム
参加した期間	第3週目～（約9週間）	第3週目～（約9週間）	第6週目～（約6週間）
最終提出したモデル	AGLM	NGBoost	GLM + RF (スタッキング)
最高予測精度ランキング	**18位**	126位	43位
最終利益ランキング	52位	12位	**10位**

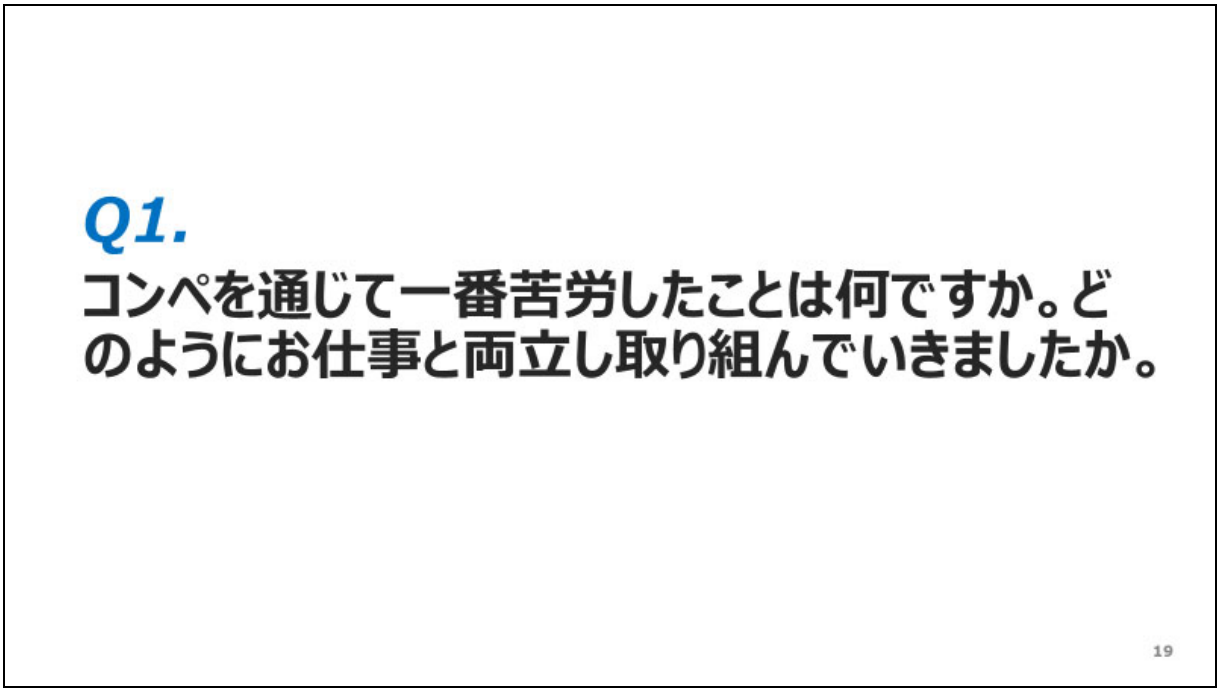
今回、世界 50 か国以上の国々から 2,000 名弱の方が参加したコンペの中で、日本の３チームがどのようなモデルを作ったのか、その結果どうだったのかを、表にしてみました。多くのコンペと同様、今回のコンペは個人でもチームでも参加することが可能になっていました。荒井様と伊藤様は個人で、﨑村様のところは３人チームで参加されたということです。このコン

ペ自体は全 11 週間に亘って行われましたが、荒井様・伊藤様は第３週目から９週間ほど参加されています。﨑村様のチームは６週目からですので、約６週間の参加期間ということです。

最終的に提出したモデルは、偶然にもそれぞれ別のモデルになっています。荒井様が AGLM、伊藤様が NGBoost、﨑村様のチームが GLM とランダムフォレストのスタッキングモデルです。興味深いのは、内２チームが長年アクチュアリーに親しまれている GLM ベースのモデルを使用しています。荒井様は GLM から派生した AGLM を使っており、﨑村様チームも GLM をランダムフォレストに組み込んで使用しています。伊藤様は Boosting をベースとした手法を使っていますが、プレゼンの中でもコメントされたように、この NGBoost というモデルは、y そのものを予測するだけではなく、その分布すなわち不確実性も予測するモデルなので、アクチュアリーにとって親和性が高い手法なのでは、ということです。ふたを開けてみると、３チームとも、アクチュアリーにとって親和性の高いモデルを選ばれたことがわかりました。

表の下２行が、順位の最終結果です。予測精度のランキングは、荒井様が 18 位、伊藤様が 126 位、﨑村様チームが 43 位で、荒井様の AGLM が 18 位と非常に好成績を収めています。そして、利益のランキングは、荒井様は 52 位、伊藤様が 12 位、﨑村様チームが 10 位で、皆様、本当に初心者とは思えないくらい上位にランクしていることが分かります。初心者でもこのぐらいまで目指せるということを、裏付ける結果ではないかと思います。冒頭にアンケートを取らせていただきましたが、コンペに興味はあるけれどもまだ参加したことはない方々の背中を後押しできれば幸いです。

では、ＱＡセッションに移っていきたいと思います。パネリストの皆様、改めましてよろしくお願いします。まず、私から三つほど質問をご用意いたしました。



一つ目です。「このコンペを通じて一番苦労したことは何ですか。どのようにお仕事と両立し取り組んでいきましたか」、視聴者の方々からも「休日を利用してコンペに参加されているのは大変ではないですか」という質問や、「どのぐらい時間をかけましたか」という質問を頂いております。では、荒井様からよろしくお願いします。

荒井　ありがとうございます。コンペを通じて一番苦労したところというと、やはり、モデルの提出です。チュートリアルの中で、モデルの提出の仕方の解説があるので、簡単なモデルは誰でも提出できます。ただ、複雑なモデルを作ろうとすると、やはり、Ｒや Python などの知識が相当必要だということを感じています。具体的に言いますと、今回のコンペでは、各契約に対してロスがいくらですということを、テーブルデータとしてで出すのではなく、Ｒや Python などのモデルを提出する形式なのです。提出の枠組みにかなり制約があり、自分のやりたいことを、その枠組みの中で表現するということが難しかったです。例えば、頻度と強度を分けるという二つのモデルにするだけでも、かなり大変でした。

藤田　ありがとうございます。やはり、初めてコンペに参加をする際、そのような仕様の面は、なかなか苦労される部分ですね。伊藤様はいかがでしょうか。

伊藤　私も同じく、モデルの提出で苦労しました。私は Python を使ったのですけれども、いざモデルを送ったとしても、主催者側で動かしたところエラーで止まっていますというような案内が来たり、うまくライブラリを読み込んでくれなかったり、そのようなプログラミングの面で苦労したというところが一番大変だったかと思っております。仕事との両立に関しては、私は主に土日を使ってコンペに取り組んでいたのですけれども、そうですね、先ほどご紹介があったように、３週目ぐらいから平均して、土日のうちどちらか１日ぐらいを毎週使ってやっていたように記憶しております。

藤田　ありがとうございます。提出してエラーが返ってきたときに、どのように解決されたのですか。

伊藤　そうですね。かなり詳細にエラーコードまで教えていただけましたので、それを Google で検索したり、何回も送ってみたりして、何とか自力で解決していったという感じです。

藤田　なるほど。﨑村様、いかがでしょうか。

﨑村　週に１回練習試合のような形で、利益のフィードバックが得られましたが、その解釈がなかなか難しかったです。週に１度、一つのモデルしか提出することができないため、毎週、どんどん良いモデルにしていこうとあちこちを変更したため、どの部分がどう寄与したのかを見ることが大変でした。また、他の参加者のモデルもどんどん変わっていきますので、相対評価の中で、自分のしたことが何か改善につながったのか、たまたま周りのモデルが何か違うことをしたために結果が変わったのか、その判断も難しかったですね。

藤田　ありがとうございます。そうですね、やはりコンペは、ある意味ゲームなので、定められた制約条件の中で、どのように原因を分析して、解決していくのかということは、なかなか難しい部分ではあると同時に、面白さや醍醐味の一つなのだろうと思います。
　視聴者の方々から関連する質問を頂いております。「休日を利用してのコンペの参加という話がありましたが、今回のコンペでは、1 日当たり大体どのぐらいの時間をかけて参加されましたか」というご質問です。﨑村様、いかがですか。

﨑村　やはり、仕事との兼ね合いで、ほとんど作業ができない週もあれば、平日も含めて休日も、がっつりとこのコンペに取り組むことができた週もあり、週によってばらばらだったかなと思います。ただ、毎週末、締め切りが日曜日でしたので、やはり、その直前はどうしても、できるだけいいものを提出したくて、ついつい夜な夜なやってしまいました。

藤田　何時間ぐらいやりましたか。具体的に覚えていますか。

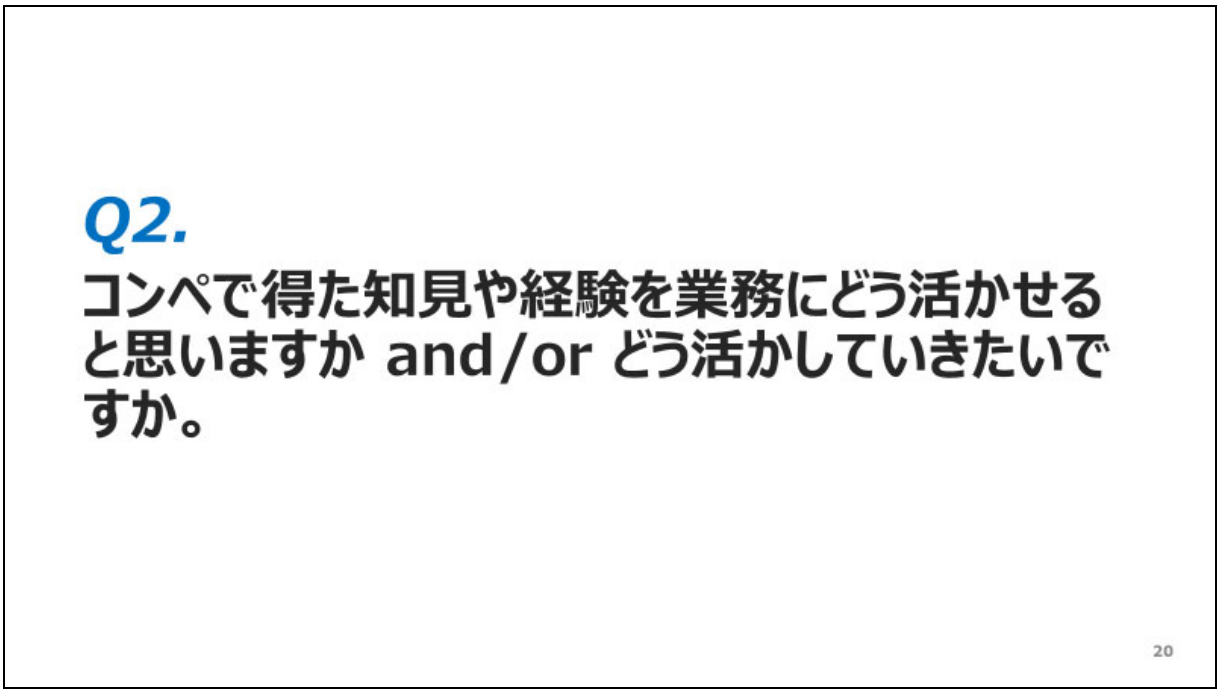
﨑村　どうでしょう。私は、少し朝が弱いので、午後からがっつり、夜中ぎりぎりまで取り組んでいたという日もありました。

藤田　コンペに参加すると、夜型になってしまうということも、よく聞きますね。伊藤様、いかがですか。

伊藤　私も同じような状況でしたが、土日がメインだったかと思います。やり出すと、結構夜遅くまでやってしまうこともあったかと記憶しています。

藤田　ありがとうございます。荒井様、いかがでしょうか。

荒井　私も休日ほぼほぼ、具体的に言うと 12 時間や 13 時間など、かなりの時間を使っていました。暇な人と思われてしまうかもしれないのですが、ちょうど、コンペに参加したのが去年の 12 月中旬からですので、新型コロナの影響で外に出られない状況でもあり、一気にぎりぎりやった感じです。



藤田　ありがとうございます。それでは、次の質問に移りたいと思います。二つ目「コンペで得た知見や経験を業務にどう活かせると思いますか。もしくは、どう活かしていきたいですか」、では、こちらも荒井様から、よろしくお願いします。

荒井　やはり、高度なモデルを実際の業務に使うことは、なかなか難しいところがあると思いますが、一方で今回コンペに参加したことで、EDA の幅が広がったと思います。コンペでいい成績を取ろうとすると、どうしても EDA をかなりやり込むことになります。具体的にいいますと、今回、交互作用項を結構分析したのですけれども、データを入れれば、すぐに可視化してくれるようなＲのパッケージがあることが分かり、それで分析していました。業務でも今後同じような状況になった時、「Ｒを使えばいいや」というような選択肢が増えるという意味では、とても良かったのではないかと思います。

藤田　EDA とは、まさにデータそのものをくまなく調べるという作業で、アクチュアリーにとっても、クレームデータを使ってロスモデリングをしていく際など、普段の業務で行うところと思うので、そちらの知識は確かに拡がりますね。伊藤様、いかがでしょうか。

伊藤　私もまた同じで恐縮なのですが、やはり、データを実際に触ってみるという経験は、業務にも直接活きてくると思います。データにどのような特徴があるのだろう、どのように考えたらいいのだろうということを探していく、まさに EDA になりますが、これは、アクチュアリーの業務も広い意味ではデータ分析であると捉えると、かなり業務に活かせるところではない

かと思います。また、知見という点では、コンペに参加する前までは、保険のプライシングは、様々な規制などを仮に全部考えなければ、比較的機械学習による予測を活用しやすい分野なのではないかという印象を持っていました。しかし、実際に自分でやってみると、あくまでも今回のデータやルールの中での話ですが、考えていたほど簡単な話ではない、という知見が得られたので、良かったと思います。

藤田　ありがとうございます。久下様、いかがでしょうか。

久下　少し話をさせていただきたいと思います。今回のパネラーの方々は、アクチュアリー会の中でということで、コンペティションは「初めまして」とおっしゃっていましたが、統計処理に関してはよくよくご存じの方々であったと感じております。今、聞いていらっしゃる方の中で、本当に初心者なのですという方がいらっしゃったら、私の意見も是非、聞いていただければと思います。

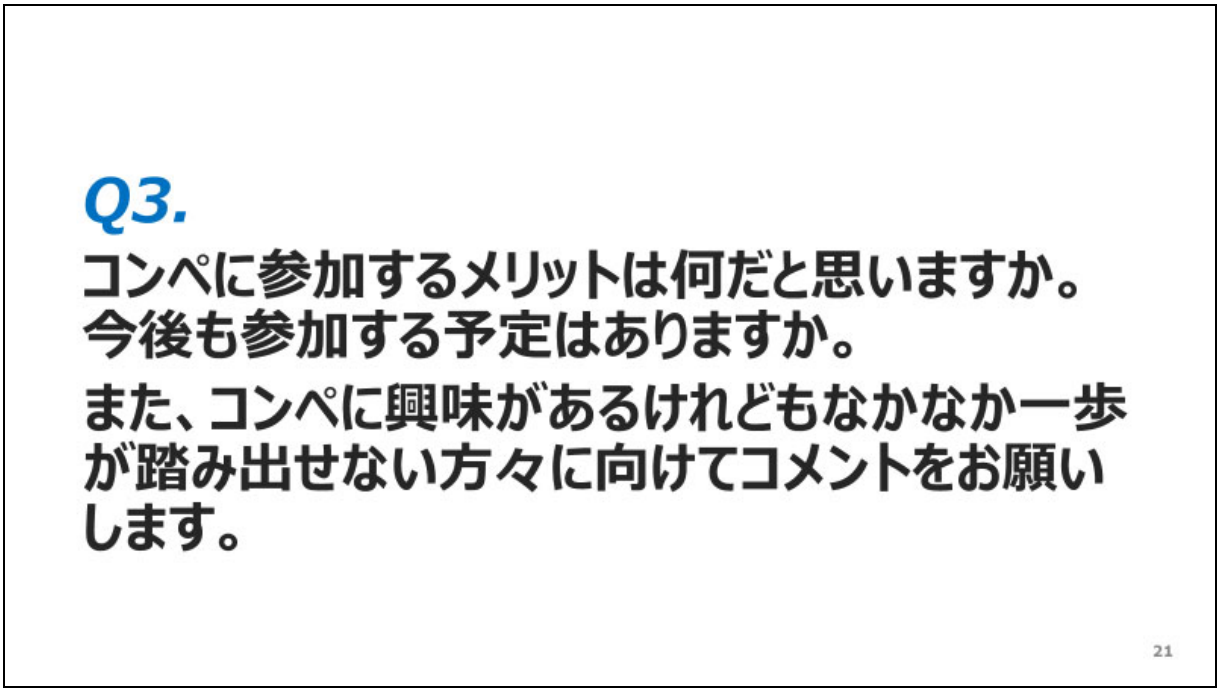
私のバックグラウンドとしては、自然災害リスクに関する、地震や台風に関するエンジニアで、その中で統計処理なども使っておりますが、そこで、アクチュアリーの技術が、何かしらコラボレーションできないかということで、会社の中で３人チームとなって、このコンペティションに参加した次第です。私自身は統計学に関して、本当に初心者で、今回コンペに参加して実際のデータを使いながら、どのような統計の処理の手法があるのかということを学べただけでも、非常に有意義であったと思います。私のエンジニア側の話で申し上げると、やはり非線形の動きが肝になります。地震や台風のリスクは線形では図れませんので、非線形の動きというものを統計的にどのように処理すればいいのだろうと考えたとき、GLMや基礎的な統計処理手法だと、非線形の動きというものをなかなか捉えられないというところを感じていたところです。もちろん、GAMのような非線形を捉えようという統計処理手法もありますが、分布を一定仮定しないとならないなど、その分布を設定することにもかなり時間を要してしまう欠点があります。

今回、機械学習というテーマでこのコンペに参加し、ここまで見てきたように非線形の動きを捉えられる統計処理手法もたくさんあるのだということを知ることができただけでも良かったです。その処理手法に関して、うまくいきますというものだけではなくて、実はうまくいかないですというところも、データを使いながら感じることができたということは、勉強の機会になっただけでも、とても良かったと思います。非線形のモデルというものを、自然災害リスクのところに持ち込んでいきたいと感じることもできました。

もう一点、別の切り口で話をさせていただきたいと思います。﨑村さんから、コモンタスクフレームワークの強みを感じたということを説明していただきました。同じ課題を、みんなで解決していこうというようなことで、ディスカッションボードが活発だったことは、とても驚きでした。これまでの時代だと、１人の天才が何かしら新しいものを作れば、それにみんな乗っかってくるということだったかもしれません。今の時代は知識が飽和してきており、みんなで知識を持ち寄って、より良い精度のモデルをどんどん、スピーディーに作っていくというような文化が、機械学習の中ではすでにあるということを感じることができました。自然災害リスクでも、産学官が連携して、地震や台風に対する対策を進めていこうとはしていますけれども、民間企業だけでも、コモンタスクフレームワークのような課題を一緒に解決していこうと

いう文化を作ることができれば、より良い世の中がよりスピーディーに展開できるのではないかということを考えるきっかけになった点でも、コンペティションへの参加はとても有意義だったと思っております。

藤田　コモンタスクフレームワークのお話まで、ありがとうございます。近年、ディープラーニングや機械学習の名前をあちこちで聞くことになったきっかけの一つが、コモンタスクフレームワークというものなのですね。また、久下様が最初におっしゃっていましたが、普段、われわれが漠然と持っている知識も、コンペティションを通じて、実際に使ってみて、メリット、デメリットというものを肌で感じることができるという点も参加する意義ではないかと思います。ありがとうございます。



では、三つ目の質問に移ります。「コンペに参加するメリットは何だと思いますか。今後も参加する予定はありますか。また、コンペに興味はあるけれどもなかなか一歩が踏み出せない方々に向けてコメントをお願いします」ということで、逆順で﨑村様からお願いいたします。

﨑村　コンペに参加するメリットは、皆さんおっしゃっていましたが、実際に手を動かしてみて、ここのモデルは、このような場合にはうまく使えないのだということや、この組み合わせはとてもうまくいくということを、本当に身に染みて感じられるところかと思っていまして、今後も参加できるものがあれば、是非参加していきたいと思っています。
なかなか一歩が踏み出せない方に向けては、先ほど、知識やプログラミングへの不安が大きくて参加できないという方が一番多かったかと思うのですが、コモンタスクフレームワークという文化があることをお伝えしたいと思います。例えば、一般的なコンペにはディスカッションボードがあり、私も今回、モデルの提出の仕方など詰まってしまった部分をディスカッションボードで質問しました。皆様、レスポンスもとても早くて、ここをこのように変えたらいいのではないかということを、すぐに教えてくださるような環境でした。初心者でも分からないこ

とがあれば、すぐ聞いてしまうという形で参加ができると考えれば、一歩踏み出しやすいかと思います。

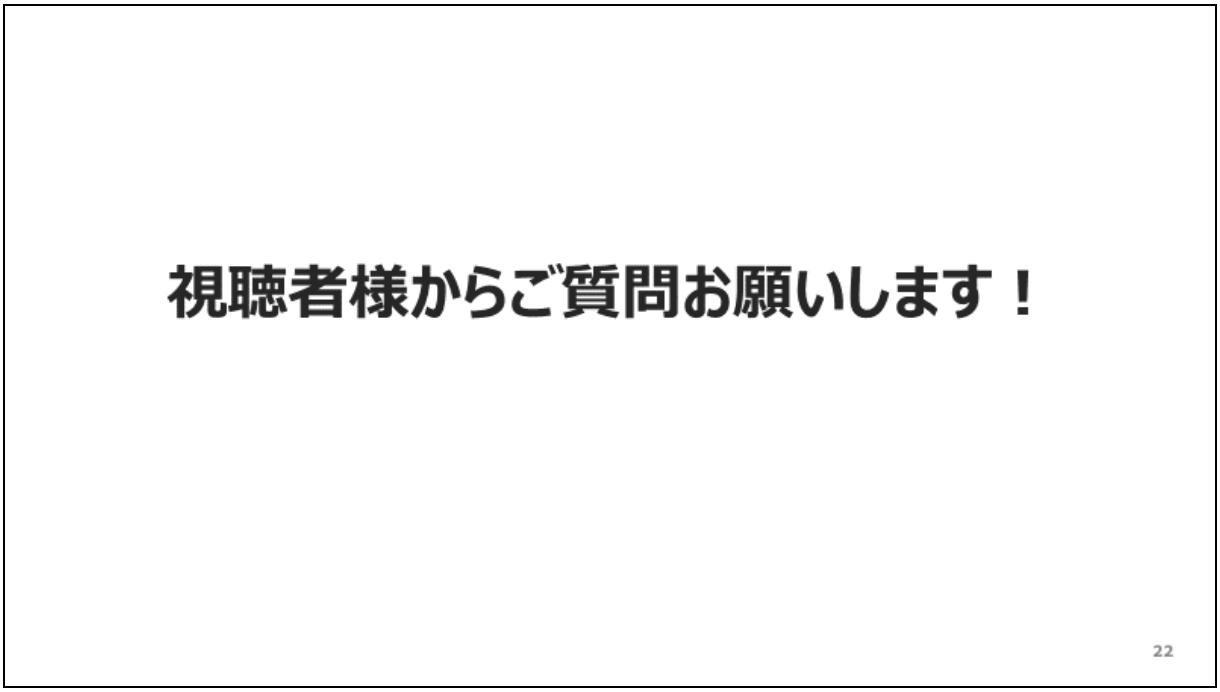
藤田　ありがとうございます。伊藤様、よろしくお願いします。

伊藤　私も今回参加してみて、実際にまず手を動かしてみるということが、とても大事だと思いました。教科書的な知識があったとしても、実際にやってみて初めて分かる部分というものも多くあるのだと改めて感じました。また、データを触ってみると、単純に、「こういうデータと、こういうデータって、このような関係があるんだ」という知見や、プログラムのパッケージの動き方、コンペのスコアを向上させていく中での発見などもたくさんありました。今後も面白いコンペがあれば、是非、参加しようと思っております。

藤田　ありがとうございます。荒井様、いかがでしょうか。

荒井　はい。まず、コンペに参加するメリットとしては、何の制約もなく、楽しく知識や経験が得られる点だと思います。やはり、業務であれば、会社の戦略や法令上の規制など、いろいろな制約がある中進める必要がありますが、コンペであれば、完全に自由です。今回でいえば、保険料が料率三原則に従う必要はないので、普段ではあり得ない条件の中で、自由に分析・料率設定できるところが、なかなか面白かったところです。

次に、なかなか一歩踏み出せない方々に向けてのコメントですが、私自身まだ１回目の参加なので、アドバイスできるようなことはありません。ただ、一つあるとすると、やはり興味のある分野のコンペに参加することが一番ではないかと考えています。例えば、データサイエンスの勉強を始めようとすると、最初はタイタニックやアヤメなどから始めるのでしょうけれども、そこで興味がないといずれ飽きが来ると思います。自分の興味がある分野であれば、ここをこうしようなど、どんどんひらめきがあり、深堀することができるのではないかと思いました。以上です。



藤田　ありがとうございます。次は視聴者の方から頂いた質問です。「このコンペに向けてのデータサイエンスの専門知識は、いつ、どのように勉強しましたか」、やはり、アンケートでハードルとして挙げられていた一番は、プログラミングや専門知識とのことでしたので、こちらについてお伺いしたいと思います。荒井様、いかがでしょうか。

荒井　はい。私の場合は、アクチュアリー会のムーンライトセミナーで、AGLM もそうですし、データ分析のコンペなどもあったので、そこで知識を得られたと考えています。

藤田　ありがとうございます。伊藤様、いかがですか。

伊藤　私の場合は、書籍やインターネットが中心ですが、数年前データサイエンスがはやり出した頃にちょうど大学にいたこともあって、大学の講座や、あとは社内の講座なども受講して、知識を得ることができました。

藤田　社内でも、何かやられているのですか。

伊藤　そうですね。弊社にはデータサイエンス研修のようなものがたくさんあって、大変、勉強になっています。

藤田　結構、細かいモデルの話などもあるのですか。

伊藤　はい、なかなかレベルが高くて、大変、勉強になっています。

藤田　ありがとうございます。﨑村様、いかがですか。

崎村　プログラミング自体は、学校で学習する機会がありましたが、社会人になる頃には正直使い方も忘れてしまっており、再勉強が必要でした。伊藤さんと同じように社内のトレーニングがあったり、オーガナイザーの藤田さんに引っ張っていただけたことも恵まれていました。周りで一緒に興味がある人を探してということは、とてもいい方法ではないかと思います。

藤田　ありがとうございます。会社によっては、独自で機械学習部やデータサイエンス部のような、サークル活動のようなものを作って、勉強会や輪読会を開催したりするという話は、たまに聞きますね。久下様、いかがですか。

久下　先ほど、崎村さんからも話がありましたとおり、Ｒや Python に関しては、実は業務でかなり使っていることもあり、特にハードルはなかったです。一方で機械学習に関しては、学ぶ機会がとても少なく感じていまして、自然災害に是非使っていかなければいけないという意識のもと、藤田さん、崎村さんを巻き込んで、機械学習について学ぼうというようなチームを作り、週次か隔週で勉強会を行う体制が取ることができていました。会社の中で興味がある方が集まって勉強会をするというケースもありますし、大手の企業だと、講座のようなものもあるようですね。そのような場に積極的に参加することが良いのではないかと思っております。

藤田　ありがとうございます。続いて「現在開催中、もしくは近いうちに開催予定のデータ分析コンペティション（アクチュアリー向け)」とのことですが、久下様もいらっしゃいますので、自然災害モデラー向けも付け加えさせてください。「そのようなコンペを探すには、どうすればいいでしょうか」、久下様から、いかがでしょうか。

久下　話が少し重複してしまうかもしれませんが、先ほどのチームの勉強会の中でコンペを探すこともしました。方法の 1 つとしては、Kaggle や他のプラットフォームのメールマガジンに登録するとよいです。そうすると、コンペティションの案内が配信されますので、その中から興味のあるコンペを選んで参加するとよいのではないでしょうか。

藤田　ありがとうございます。崎村様、いかがですか。

崎村　久下さんと同じように、メールマガジンに登録していくとＥメールにお知らせが来ますので、気になったら見てみるというスタンスが、とても良いのではないかと思っています。また、個人的には LinkedIn にも、データサイエンスが好きなグループもあるので、そのようなグループに入っていると、ちらちらと見ているだけで「このコンペに今、参加しています」という情報が流れてくるので、それもいい情報源になりそうです。

藤田　では、伊藤様、いかがでしょうか。

伊藤　そうですね、私もＳＮＳやメールの配信がメインになるかと思っております。

藤田　はい。ありがとうございます。荒井様、いかがでしょうか。

荒井　全く同じです。

藤田　Twitter、LinkedIn等のＳＮＳに登録して、メールマガジンの配信等から、興味を持ったものをピックアップしていくというアプローチがいいのではないのでしょうか、ということになりそうですね。
　全体向けの質問です。「他のチームの手法、戦略で興味深かったものはありますか」、挙手制にしましょうか。

﨑村　伊藤様のモデルでNGBoostを使われていて、正直なところ、私たちのチームはぶれの部分まで予測するという余裕がなかったので、次の機会があったら是非使ってみたい手法だと思いました。

藤田　他の方々はいかがですか。

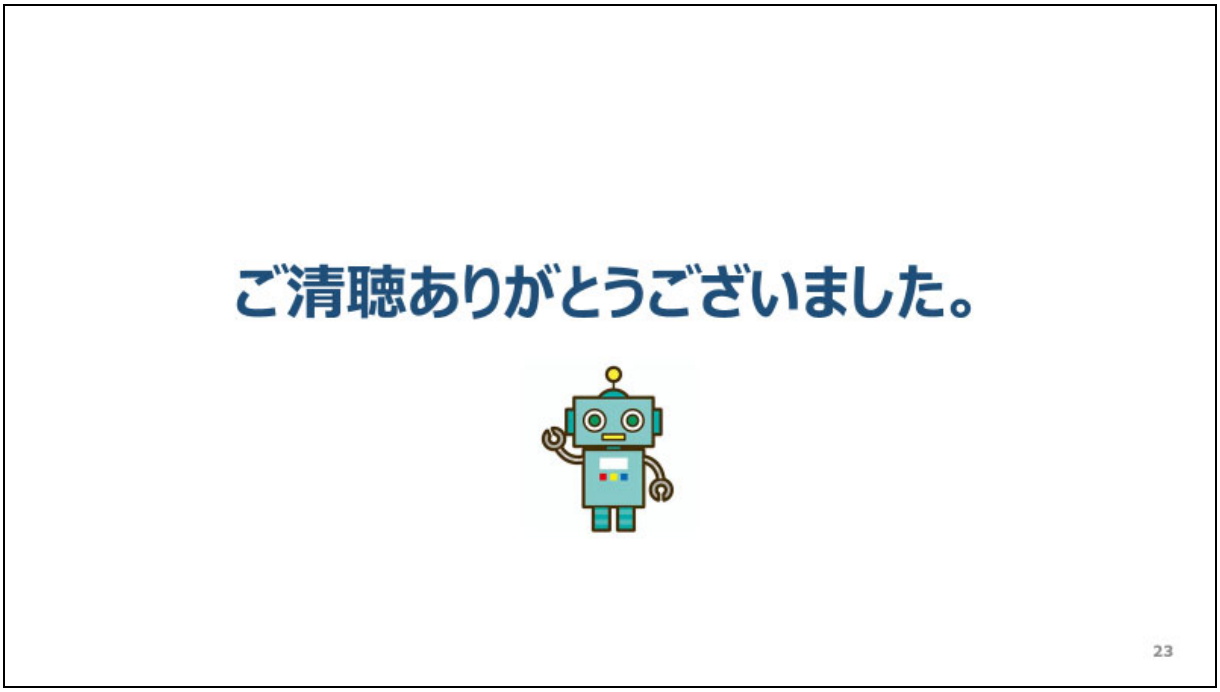
伊藤　私は、ＦＤ法のような形で、頻度や件数を分けて分析するという、業務ではやったことがあるのに、今回のコンペではすっかり頭から抜け落ちていたアプローチを聞いて、なるほどと思いました。

藤田　ありがとうございます。荒井様、お願いします。

荒井　﨑村様チームの保険料設定のところで、小口ロスに対してはマージンを一切設けないという、その戦略があまりにも斬新で、そういった戦略を採るチームがあったのであれば、料率競争で勝てないだろうというところが、反省点というか、面白かったところです。

藤田　ありがとうございます。次の質問は﨑村様に対してですね。「ロスモデルを大口とそれ以外に分けて作成することは理解できますが、提供されているデータから保険金の金額で、何らかしらのバーを設けたということでしょうか。もしくは、上位何％を取ったという感じでしょうか」。

﨑村　はい。今回は、バーを設けました。練習用のデータの平均値より大きいか、小さいかで大口、小口というように分けています。ただ、結果的にはそこに落ち着いたのですが、そこに至るまで、バーをどこに設定するかということを何パターンも試しております。



藤田　ありがとうございます。他にも多くのご質問を頂いておりますが、一旦、ここで締め切らせていただきます。時間の都合上、申し訳ありません。

クロージングということで、私から一言コメントしたいと思います。今回、AIcrowdが運営した保険系コンペを取り上げて、初心者としての立場からコンペの経験談や構築したモデルについて、パネルディスカッションを行いました。問題設定に関しては、あくまでもゲームなので、現実のビジネスとは多少のギャップがありますけれども、データの前処理からモデリング、結果の解釈など、実務プロセスの一連を深く学べたのではないかと思います。また、本コンペでは、予測精度の高さを競うだけではなく、他のチームの動きを踏まえた利益競争という観点が組み込まれていて、アクチュアリー的にも大変興味深かったのではないでしょうか。

近年の機械学習系ソフトウェアの発展、普及によって、機械学習手法というものが誰もが簡単に使える時代になってきておりますが、やはり、第一義的には、利用目的を明確にすることが重要ではないかと思います。われわれが現実に従事しているビジネス、保険金融業界を例に取りますと、不確実性の評価やモデルの解釈可能性、あるいは、モデルガバナンスといったことが強く要求されていて、単に予測精度を追求したモデルを適用することは難しい点があろうかと思います。

一方で、説明や解釈が難しいから機械学習は、やはり使えないと断定はせずに、今後は、保険業界なりの機械学習手法やフレームワークが必要になると考えられます。そのためにも、このようなコンペティションをはじめとする種々の活動を通した連携を加速していくことも重要になるだろうと思います。最後になりますが、本セッションを機に、多くのアクチュアリーの方々がデータ分析コンペティションに少しでもご興味を持っていただければ幸いです。ご清聴ありがとうございました。