

1. BUSINESS UNDERSTANDING

A **Avenue** é uma **corretora de valores** americana com a **sede em Miami e com escritório em São Paulo**, mas com o objetivo **100% focada** em **clientes que** **moram no Brasil** com o perfil de **varejo de alta renda** (pessoa que já investe em bolsa de valores, mas que desejam começar a investir no mercado financeiro americano) e com a **maior lista de opção de ativos**.

A empresa vem crescendo em um ritmo acelerado e já tem mais de 600 mil clientes investidores. A **plataforma** que funciona tanto na versão Web/Desktop quanto na versão mobile (APP) foi pioneira em investir em **recursos que melhoram o acesso do investidor comum à bolsa de valores norte-americana**.

1.1 Objetivos do Negócio

A Avenue deseja que, através deste projeto, seus serviços continuem auxiliando seus clientes a obter produtos de investimentos, mais especificamente Shares (ações) e ETF’s (fundos de investimento), através de um algoritmo que realize recomendação destes produtos, baseado no perfil do cliente.

Esta indicação de produto deve se assemelhar ao que o Netflix oferece atualmente, onde o cliente acessa a plataforma e recebe sugestões de novos produtos (ação ou ETF’s) baseado no seu perfil.

A recomendação não será exclusiva e personalizada para cada cliente individualmente, mas por cluster de clientes semelhantes (usuários que tiveram fortes ações com os mesmos produtos). Para montar os clusters, o algoritmo de Machine Learning buscará padrões entre os clientes, e apresentará uma lista de produtos sugeridos para cada cluster, ou seja, agrupamento por semelhança.

A métrica de sucesso para validar o algoritmo de recomendação é obter um modelo altamente personalizado para o cliente, tendo alta precisão. Porém, o Sponsor não sabe mensurar qual o valor ideal da precisão, tendo que alcançar o maior valor que conseguirmos.

E para concluir que o algoritmo alcançou esta expectativa, é utilizado uma parcela do dataset para treinar o modelo (aproximadamente 80% dos dados), e outra parte para testar a precisão do algoritmo (aproximadamente 20% dos dados). Para medir a assertividade, os clientes que estão na parcela do dataset de teste devem ser classificados em algum cluster, e observar se o investimento que o cliente realizou está na lista de recomendação do cluster.

O modelo de recomendação também será testado em produção pelo cliente/sponsor, que irá gerar feedbacks sobre o comportamento e performance do algoritmo implementado, utilizando a mesma metodologia de teste.

Os entregáveis do projeto devem ser:

* Um algoritmo de Machine Learning, implementado no Google Colab;
* Um Dashboard que apresente os resultados obtidos através da execução do algoritmo de forma simplificada, para apresentação aos stakeholders.

1.2 Avaliando a situação

Para atender a demanda do mercado, a Avenue possui cerca de 300 colaboradores, divididos nas seguintes áreas:

* Finance
* Staff
* Operations
* Business Development
* Executives
* IT
* Treasury

Atualmente, a Avenue conta com um pequeno time focado em análise de dados, formado por 3 colaboradores que atuam nas seguintes atividades:

* Rodrigo Petry – Head of data
* Fábio Carvalho – Engenheiro de dados
* Luiz Roberto – Analista de dados

Para auxiliar no projeto, teremos o acompanhamento direto de 1 (um) colaborador, com experiência júnior em Análise de dados e o 1 supervisor especializado em dados que oferecerá eventuais suporte e apoio para esclarecimento de dúvidas.

A solução da empresa é voltada para a recomendação de produtos financeiros, portanto utilizam termos específicos da área, como por exemplo share, ETF, fundos, ativos, custódia.

Relataram coletar os dados que serão disponibilizados para a implementação do projeto através de formulários utilizados no momento do cadastro do cliente na plataforma (web ou mobile), histórico de compra e venda de ativos.

O dataset disponibilizado contém informações reais de clientes, porém com informações sensíveis de clientes ocultas. É uma amostra que foi coletada através de algoritmo que sorteou aleatória e digitalmente os registros da população de registros da base da dados reais da Avenue, gerando uma tabela em Excel.

O arquivo no formato xls contendo a amostra dos dados será disponibilizado e mantido/atualizado através de uma conta do Google Drive que deve ser providenciada e disponibilizada pelos alunos deste time.

O cliente utilizará como premissa o acesso que disponibilizarmos para o compartilhamento dos datasets.

A Avenue possui um sistema de recomendação atualmente, mas não há registros de satisfação ou assertividade das recomendações dos produtos que são feitas hoje pela plataforma. Como o próprio cliente reforçou “é um sistema de recomendação muito simples que indica produtos apenas baseando-se no que o cliente preenche em seu perfil no momento do cadastro”. Não há métricas claras sobre a assertividade dessas recomendações, nem sobre a satisfação do cliente, nem se as indicações oferecidas foram aceitas (se gerou vendas a partir da indicação de um produto feita pela plataforma) ou não.

O que se tem é justamente a necessidade da implementação deste sistema de recomendação baseada em Machine Learning e que considere variáveis mais subjetivas e que possa ser aprimorada ao longo do tempo, sugerindo melhores produtos aos seus clientes.

Como é um produto que não possui precedentes nem métricas prévias, o cliente espera que este projeto traga a realidade sobre quão assertivo um sistema de recomendação com estas características possa alcançar. Portanto, o nível desejado de acuracidade é de +75%, mas temos total liberdade para que, durante o andamento do projeto, tragamos a realidade sobre o nível de assertividade possível, considerando o dataset e suas variáveis, período disponibilizado para realizar o aprendizado do ML, volume de dados, etc.

Os principais riscos encontrados até o momento no projeto são: o prazo, uma vez que temos a data limite de duração do curso para desenvolvimento, validação e implementação das soluções propostas; incerteza sobre a métrica de sucesso, uma vez que não há precedentes ou análises prévias para garantir um nível desejado específico de assertividade das predições; indicador de satisfação do cliente uma vez que não temos acesso a aplicação que o cliente fará no teste populacional; avaliação subjetiva de assertividade podendo mudar o escopo da análise e tratamento dos dados várias vezes ao longo de todo o processo de implementação. Além disso, como o produto do cliente pode ser influenciado por vários fatores externos (política, guerra, notícias, boatos, oscilações financeiras em todos o mundo, etc.), qualquer evento não previsto pode influenciar na predição e assertividade do modelo proposto.

1.3 Data Mining Goals

Para a Avenue, com objetivo de oferecer melhores soluções, ficou entendido que seria necessário entender melhor a preferência de seus clientes, podendo assim proporcionar melhores soluções para eles.

No momento, não há um algoritmo de recomendação de produtos baseado em perfil e comportamento de clientes similares, e sim uma recomendação global, em que não é possível avaliar uma taxa de acurácia na recomendação.

A proposta é separar os clientes em grupos, de acordo com seus perfis de movimentação e assim então direcioná-los para os produtos, ou seja, os produtos com maiores chances de serem adquiridos pelos clientes, serão apresentados para eles em uma espécie de vitrine, em destaque.

O indicador ideal tomado como base é de 70% de acurácia de recomendação, ou seja, se o cliente realmente comprou o que foi oferecido a ele, porém, ainda não há uma certeza no valor ideal do indicador, e é possível que esse valor seja inatingível, e então alterado conforme andamento do projeto.

1.4 Plano do Projeto

Para organização interna do grupo, utilizaremos o Whatsapp, Confluence e o Jira para comunicação e controle das atividades, nos reunimos diariamente para acompanhamento da evolução das atividades e semanalmente com o sponsor para acompanhamento da evolução das atividades.

Para a realização do projeto utilizaremos a metodologia CRISP-DM, onde seguiremos o seguinte cronograma para a realização das etapas:

Linha do tempo

Descrição gerada automaticamente

Figura 1: Etapas do CRISP-DM

1. DATA UNDERSTANDING

2.1 Coleta de Dados Inicial

Os arquivos estão em XLS/XLSX e foram disponibilizados no Sharepoint próprio da Avenue.

Utilizaremos o Google Colaboratory (Google Colab) para implementar a solução final solicitada pelo Sponsor, que é uma lista de recomendação de produtos para clientes.

Não há uma integração direta do Microsoft Sharepoint com o Google Drive ou Google Colab. Portanto, será necessário realizar o download dos arquivos que estão no Share Point e enviar os arquivos baixados para o Google Drive. Feito isso, as demais instruções de acesso através do Google Colab estão descritas no próprio notebook.

Cada planilha que está dentro do arquivo Dados DNC - Clientes, Ordens e Produtos\_2.xlsx equivale a 1 tabela do banco de dados da aplicação da Avenue que foi exportada pelo Sponsor e disponibilizada para o devteam poder trabalhar.

Os registros disponibilizados são referentes às movimentações de compra e venda de produtos financeiros dos clientes selecionados.

As planilhas contêm informações geradas pela aplicação da Avenue, em produção, reunindo registros de Clientes, Ordens e Produtos, do período de janeiro a outubro de 2022.

A planilha de Ordens diz respeito a todas as movimentações que ocorreram no período. A planilha de Clientes informa dados de quem fez essas transações, e a planilha de Produtos descreve os produtos que foram comprados/vendidos nesse período.

Nas planilhas não estão disponibilizados todos os registros da aplicação, pois foi coletada uma amostra aleatória deste período.

Os dados e as informações sensíveis de clientes foram ocultados.

Para evitar que registros fantasmas viessem no dataset disponibilizado, o Sponsor garantiu que a base principal da importação se concentra no histórico de ordens (compras e vendas) da aplicação e, a partir desta amostra, os dados de clientes e produtos relacionados a esta movimentação foram importados. Isso garante que todos os registros de ordens possuem um vínculo de clientes e produtos que estão disponíveis no dataset.

Não há uma periodicidade programada para atualizar o dataset. Basta que o devteam solicite e o Sponsor disponibilizará um novo arquivo contendo as alterações que forem pedidas, por exemplo, ampliar ou diminuir o range do período de importação, incluir mais produtos ou clientes, etc.

Todos os dados disponibilizados no dataset são fundamentais para iniciar o desenvolvimento do projeto. Caso haja necessidade ou entendimento de inserir novas informações ou fontes, basta solicitar ao Sponsor ou, se for fontes externas ao negócio do cliente (ex: histórico de variação cambial, registro de bolsa de valores, etc.) que podem ser obtidas por APIs disponíveis na internet, temos a liberdade de solicitar ou vincular o que for necessário para o correto desenvolvimento da solução.

Foram disponibilizados 2 arquivos no formato XLSX, em que um arquivo contém o dataset e o outro a descrição dos dados.

No arquivo de dataset existem 3 planilhas:

* Clientes
* Produtos
* Ordens

Com estas planilhas foi possível extrair 3 datasets de registros distintos que contém os registros que serão analisados.

No arquivo de descrição dos dados existem 3 planilhas:

* Perfil de Clientes
* Produtos
* Ordens

Com estas planilhas foi possível extrair 3 datasets distintos que contém as informações sobre os tipos de dados de cada campo dos datasets de registros. Não será utilizado para fins de análise, mas será incluído como automatizador que auxiliará a converter os tipos de dados para os seus equivalentes na linguagem Python.

Na planilha que contém os tipos e a descrição de cada coluna dos datasets de registros, estão bem definidos qual o tipo de dados esperado para cada variável do dataset. No entanto, diferem do padrão de tipos de dados aceito pelo Python.

Para tratar isso, foi criado um dicionário de conversões que relaciona o tipo descrito na planilha e seu equivalente na linguagem Python, bem como toda a lógica necessária para realizar as conversões do tipo automaticamente durante o processo de importação do dataset para o colab.

A seguir estão listados todos os datasets e a descrição das variáveis que compõem o entendimento das informações que serão trabalhadas durante o desenvolvimento do projeto.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tabela** | **Campo** | **Tipo** | **Descrição** |
| Perfil de Clientes | **AccountID** | STRING | Campo de identificação do clientes e de relacionamento com a tabela de ordens |
| Perfil de Clientes | **Function** | STRING | Profissão do cliente |
| Perfil de Clientes | **MonthlyIncome** | NUMERIC | Renda mensal em reais |
| Perfil de Clientes | **NetWorth** | NUMERIC | Patrimônio líquido que o cliente declara em reais |
| Perfil de Clientes | **TotalInvested** | NUMERIC | Valor total que o cliente possui de investimentos em reais |
| Perfil de Clientes | **SubscriptionPlan** | STRING | Plano que o cliente possui dentro da avenue. |
| Perfil de Clientes | **InvestmentExperience** | STRING | Experiência do cliente com investimento. |
| Perfil de Clientes | **RiskTolerance** | STRING | Risco que o cliente |
| Perfil de Clientes | **InvestmentObjective** | STRING | Objetivo de investimento do cliente. |
| Perfil de Clientes | **AvenueRiskProfile** | STRING | Perfil de risco do cliente (Conservador, Moderado, Arrojado, etc.) |
| Perfil de Clientes | **TargetAudience** | STRING | Público alvo, definido através de características no cadastro do cliente. |
| Perfil de Clientes | **Age** | INTEGER | Idade do cliente |
| Perfil de Clientes | **MaritalStatus** | STRING | Estado civil do cliente |
| Perfil de Clientes | **PersonMailingState** | STRING | Estado do cliente |
| Perfil de Clientes | **YearBirthDate** | INTEGER | Ano de Aniversário |
| Perfil de Clientes | **Geracao** | STRING | Geração que cliente se encaixa de acordo com seu ano de aniversário. |

Tabela 1: Descrição das colunas da planilha de Clientes

|  |  |  |  |
| --- | --- | --- | --- |
| **Tabela** | **Campo** | **Tipo** | **Descrição** |
| Ordens | Date | DATE | Data da transação |
| Ordens | AccountID | BYTES | Campo de identificação do cliente e de relacionamento com a tabela de perfil de clientes |
| Ordens | ProductCategory | STRING | Categoria do produto (Stocks ou ETF's) |
| Ordens | Symbol | STRING | Ticker do produto, símbolo. E campo de relacionamento com a tabela de produtos . |
| Ordens | Group | STRING | Grupo que o produto pertence. |
| Ordens | Industry | STRING | Indústria a que o produto pertence. |
| Ordens | Sector | STRING | Setor que o produto pertence |
| Ordens | BuySell | STRING | Tipo de ordem ( B = Buy / S = Sell ) |
| Ordens | TotalExecutedQuantity | NUMERIC | Quantidade executada |
| Ordens | TotalExecutedVolume | NUMERIC | Valor total executado em dólares |
| Ordens | PercCustody | NUMERIC | Percentual que o valor da ordem executada representa sobre a custódia total do cliente no dia da ordem. |

Tabela 2: Descrição das colunas da planilha Ordens

|  |  |  |  |
| --- | --- | --- | --- |
| **Tabela** | **Campo** | **Tipo** | **Descrição** |
| Produtos | Symbol | STRING | Ticker do produto, símbolo. E campo de relacionamento com a tabela de ordens. |
| Produtos | Product | STRING | Nome da empresa/ descrição da Stock ou do ETF's |
| Produtos | Product\_N1 | STRING | Nível 1 de segmentação do produto (Renda varível) |
| Produtos | Product\_N2 | STRING | Nível 2 de segmentação do produto (Stocks e ETF's) |
| Produtos | Product\_N3 | STRING | Tipo de produto Stocks ou Etf's |
| Produtos | Sector | STRING | Setor que o produto pertence |
| Produtos | Industry | STRING | Indústria/segmento a que o produto pertence. |
| Produtos | Group | STRING | Grupo que o produto pertence. |

Tabela 3: Descrição das colunas da planilha de Produtos

As tabelas 1, 2 e 3 descrevem as colunas das planilhas Clientes, Ordens e Produtos, respectivamente. “Campo” contém o nome de cada coluna do dataset, enquanto que “Tipo” nos informa o tipo esperado da variável e “Descrição” explica o que cada coluna diz a respeito.

2.2 Descrição dos Dados

Uma vez que se tem acesso a todos os datasets importados e suas variáveis devidamente ajustadas e normalizadas, inicia-se a etapa de exploração dos dados com o intuito de compreender melhor sobre os registros e identificar se os insights gerados atendem às expectativas e objetivo do projeto.

Para iniciar a exploração, serão avaliados que os dados possuem registros nulos ou duplicados, além da proposição de ações corretivas para eventuais inconsistências do dataset. Entende-se que os modelos de **machine learning** (que é a proposta deste projeto) **não podem conter registros nulos**. Já os registros duplicados ou desnecessários para o objetivo do negócio precisam ser identificados e tratados para que o ML esteja o mais **otimizado** possível.

Inicia-se a exploração, portanto, verificando os datasets em busca de valores nulos:

|  |  |  |  |
| --- | --- | --- | --- |
| **Colunas** | **Faltantes** | **Total** | **Valores Faltantes (em %)** |
| **AccountID** | 0 | 30226 | 0,00% |
| **Function** | 0 | 30226 | 0,00% |
| **MonthlyIncome** | 0 | 30226 | 0,00% |
| **NetWorth** | 0 | 30226 | 0,00% |
| **TotalInvested** | 0 | 30226 | 0,00% |
| **SubscriptionPlan** | 0 | 30226 | 0,00% |
| **InvestmentExperience** | 0 | 30226 | 0,00% |
| **RiskTolerance** | 0 | 30226 | 0,00% |
| **InvestmentObjective** | 0 | 30226 | 0,00% |
| **AvenueRiskProfile** | 0 | 30226 | 0,00% |
| **TargetAudience** | 0 | 30226 | 0,00% |
| **Age** | 0 | 30226 | 0,00% |
| **MaritalStatus** | 0 | 30226 | 0,00% |
| **PersonMailingState** | 0 | 30226 | 0,00% |
| **YearBirthDate** | 0 | 30226 | 0,00% |
| **Geracao** | 0 | 30226 | 0,00% |

Tabela 4: Descrição dos dados da planilha de Clientes

|  |  |  |  |
| --- | --- | --- | --- |
| **Colunas** | **Faltantes** | **Total** | **Valores Faltantes (em %)** |
| **Group** | 17779 | 60423 | 29,42% |
| **AccountID** | 0 | 60423 | 0,00% |
| **BuySell** | 0 | 60423 | 0,00% |
| **Date** | 0 | 60423 | 0,00% |
| **Industry** | 0 | 60423 | 0,00% |
| **Sector** | 0 | 60423 | 0,00% |
| **Symbol** | 0 | 60423 | 0,00% |
| **ProductCategory** | 0 | 60423 | 0,00% |
| **TotalExecutedQuantity** | 0 | 60423 | 0,00% |
| **TotalExecutedVolume** | 0 | 60423 | 0,00% |
| **PercCustody** | 0 | 60423 | 0,00% |

Tabela 5: Descrição dos dados da planilha de Ordens

|  |  |  |  |
| --- | --- | --- | --- |
| **Colunas** | **Faltantes** | **Total** | **Valores Faltantes (em %)** |
| **Group** | 14605 | 35064 | 41,652% |
| **Industry** | 10021 | 35064 | 28,579% |
| **Sector** | 10021 | 35064 | 28,579% |
| **Product** | 633 | 35064 | 1,805% |
| **Symbol** | 1 | 35064 | 0,003% |
| **Product\_N1** | 0 | 35064 | 0,000% |
| **Product\_N2** | 0 | 35064 | 0,000% |
| **Product\_N3** | 0 | 35064 | 0,000% |

Tabela 6: Descrição dos dados da planilha de Produtos

As tabelas 4, 5 e 6 mostram o número de valores faltantes em cada coluna dos dataset de Clientes, Ordens e Produtos, respectivamente. Além disso, descreve o “Total” de linhas que cada coluna do dataset tem.

Desta forma, conclui-se que o **dataset de Clientes não possui valores faltantes (nulos).** Já os datasets **Ordens** e **Produtos** apresentam algumas colunas com **valores nulos**. As colunas que mais têm valores faltantes são “Group”, “Industry” e “Sector”.

Mais adiante serão analisadas cada uma dessas variáveis com valores nulos para compreender melhor seu comportamento.

Seguindo com o objetivo da exploração, foram analisados os datasets de Clientes, Ordens e Produtos e não foi constatado nenhum registro duplicado.

Após analisar os datasets separadamente, fizemos a união deles em um único dataset, utilizando como referência o dataset de Ordens. Com isso, podemos utilizar ferramentas da Estatística Descritiva para entendermos melhor a distribuição dos dados.

|  |  |  |  |
| --- | --- | --- | --- |
| **Descrição** | **TotalExecutedQuantity** | **TotalExecutedVolume** | **PercCustody** |
| **Quantidade** | 60423 | 60423 | 60423 |
| **Média** | 14,65 | 420,00 | 0,207 |
| **Desvio Padrão** | 267,31 | 3279,90 | 0,265 |
| **Min** | 0 | 0,01 | -2,292 |
| **25%** | 0,2 | 24,48 | 0,038 |
| **50%** | 1 | 81,08 | 0,099 |
| **75%** | 3 | 228 | 0,254 |
| **Máx** | 20000 | 404598,88 | 8,861 |

Tabela 7: Estatística descritiva dos dados numéricos em um único dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descrição** | **MonthlyIncome** | **NetWorth** | **TotalInvested** | **Age** | **YearBirthDate** |
| **Quantidade** | 60423 | 60423 | 60423 | 60423 | 60423 |
| **Média** | 35072,33 | 1287304,44 | 715672,55 | 37,04 | 1984,96 |
| **Desvio Padrão** | 297251,96 | 12701550,73 | 21625505,07 | 10,53 | 10,53 |
| **Min** | 0 | 0 | 0 | 18 | 1941 |
| **25%** | 3500 | 26000 | 2000 | 30 | 1979 |
| **50%** | 8000 | 200000 | 40000 | 36 | 1986 |
| **75%** | 19000 | 800000 | 222000 | 43 | 1992 |
| **Máx** | 20853255 | 1283156474 | 1440000000 | 81 | 2004 |

Tabela 8: 2ª parte da Estatística descritiva dos dados numéricos em um único dataset

Nas Tabelas 7 e 8 estão descritos os dados das colunas que são numéricas do dataset resultante da união dos 3 datasets que foram disponibilizados. Nas tabelas podemos observar que temos a quantidade de linhas (“Quantidade”), que, assim como o esperado, é igual a quantidade de linhas do dataset “Ordens”, pois foi usado o mesmo como referência. Além disso, temos acesso à “Média” e “Desvio Padrão” dos dados. Mínimo e máximo de cada coluna também pode ser observado, assim como os quartis.

Esses dados nos dão uma noção inicial de como estão suas distribuições, que podem ser melhor visualizadas na “Exploração dos Dados”.

2.3 Exploração dos Dados

Nesta etapa, iremos observar melhor a maneira como os dados estão dispostos e se existe algum tipo de correlação entre eles. Para isso, iremos utilizar os gráficos de mapa de calor (para visualizar melhor as correlações), boxplot (para observar como os dados numéricos estão distribuídos) e o histograma (para observar como os valores dos dados estão distribuídos)

A correlação mede a associação entre as variáveis numéricas duas a duas, sendo que quanto mais próximo de 0 menor é a correlação, e quanto mais próximo de 1 ou -1 maior a correlação entre as variáveis.

No gráfico Boxplot, podemos observar os limites inferior e superior, além dos quartis e da mediana, mostrando como os dados estão distribuídos. No Boxplot também podemos ver os outliers.

O histograma é um gráfico de barras que demonstra a distribuição de frequências.



Figura 2: Mapa de calor das correlações

Na Figura 2 podemos visualizar as correlações existentes entre as variáveis numéricas do dataset.

A correlação entre “Age” e “YearBirthDate” ser -1 era esperada, pois quando temos idades maiores o ano de nascimento vai ser menor. Portanto representa uma forte associação inversamente proporcional.

As outras correlações são muito baixas (mais próximas de zero). A correlação entre “TotalExecutedQuantity” e “TotalExecutedVolume” é considerada baixa (0,21), porém devemos ficar atentos se houver multicolinearidade.

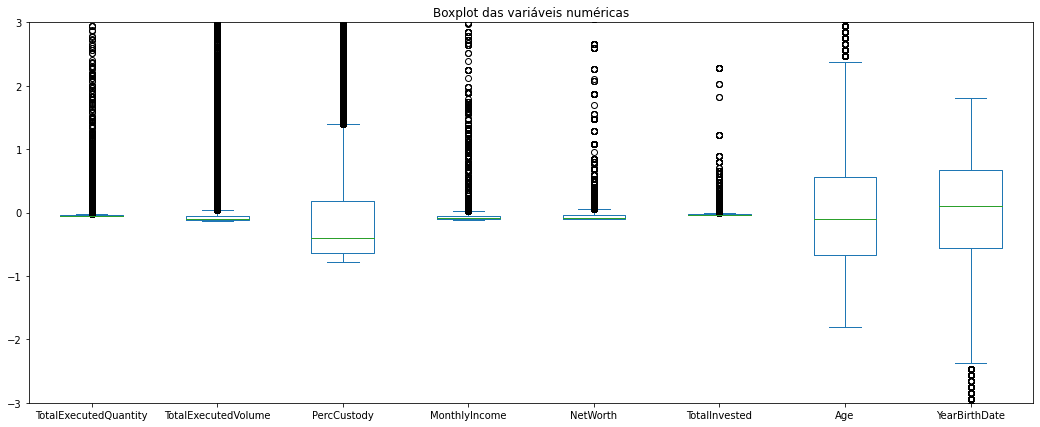


Figura 3: Boxplot das variáveis numéricas normalizadas

Na Figura 3 pode ser visualizado o Boxplot das variáveis numéricas do dataset normalizadas.

Analisando o Boxplot podemos observar que os dados de “TotalExecutedQuantity”, “TotalExecutedVolume”, “MonthlyIncome”, “NetWorth” e “TotalInvested” estão mais concentrados em relação aos demais. Isso porque a distância interquartil é muito pequena.

Já os dados de “PercCustody”, “Age” e “YearBirthDate” são menos concentrados.

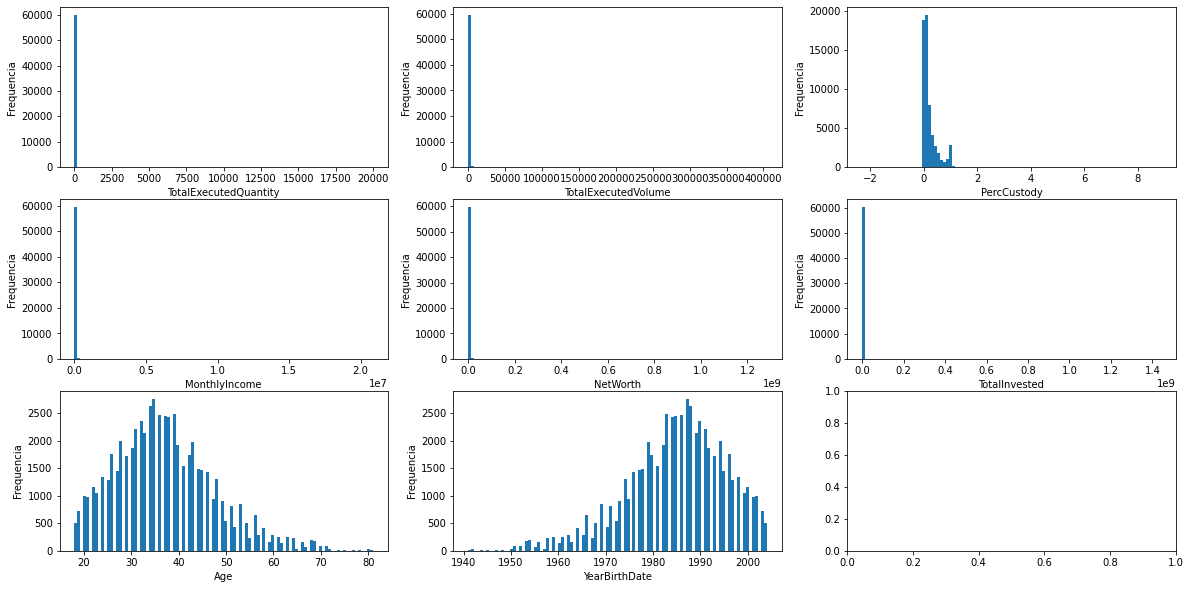


Figura 4: Histogramas das variáveis numéricas

Na Figura 4 podem ser visualizados os histogramas das variáveis numéricas. Nos gráficos podemos confirmar o que foi visto nos boxplots e ainda observar melhor como estão as distribuições dos dados. Também vemos a correlação de -1 das variáveis “Age” e “YearBirthDate”, pois os gráficos têm os mesmos formatos, porém invertidos.

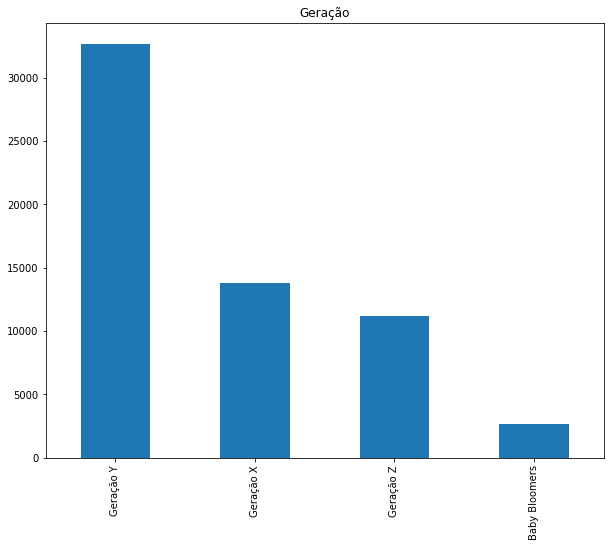


Figura 5: Geração dos Clientes

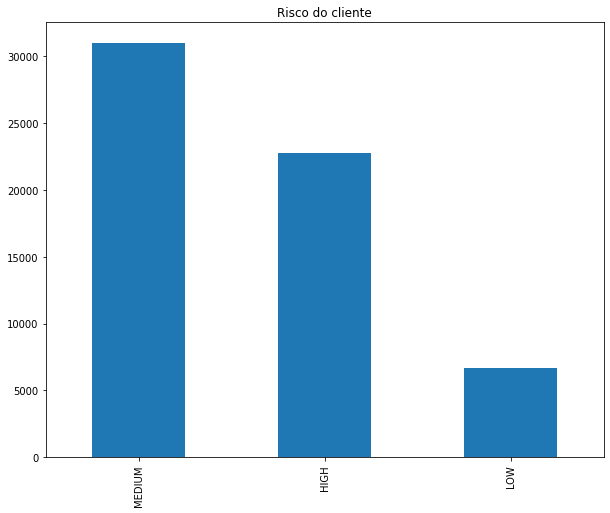


Figura 6: Risco dos Clientes

Pela Figura 5 e 6 conseguimos mapear um pouco do perfil dos clientes que mais temos no dataset. A maior parte dos clientes são da Geração Y (que tem entre 28 e 42 anos) e tem um perfil de risco médio/alto.

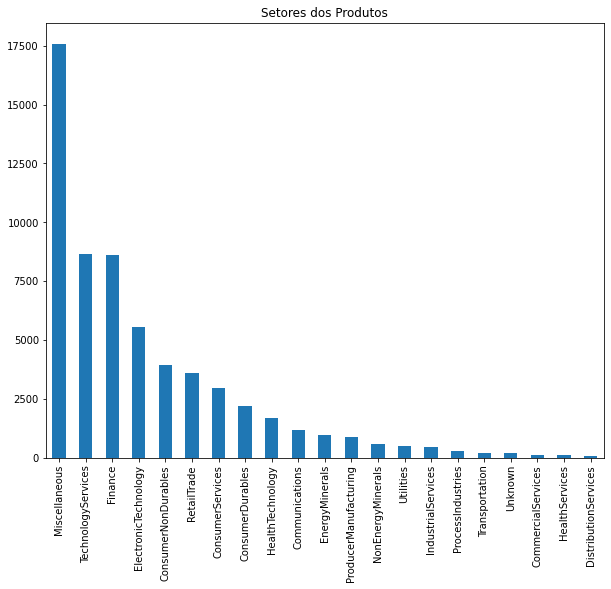


Figura 7: Setores dos Produtos

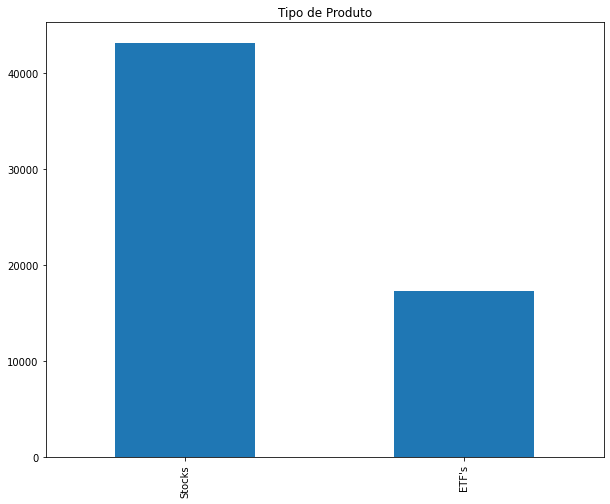


Figura 8: Tipo de produtos

Pelas Figuras 7 e 8 podemos analisar o tipo de produto que mais foi investido que está em nosso dataset. A maior parte investe em “Stocks” e os setores mais investidos são de Serviços de Tecnologia e Financeiro.

2.4 Qualidade dos Dados

Os dados apresentados no item 2.2 nos mostraram que a planilha de clientes não apresentou nenhum dado nulo. A planilha de ordens apresenta 29.42% de nulos na coluna ‘Group’. E a planilha de produtos apresenta 41,65%, 28,57%, 28,57%, 1,8% e 0,002% respectivamente para as colunas ‘Group’, ‘Industry’, ‘Sector’, ‘Product’ e ‘Symbol’.

Uma das possibilidades para esses valores faltantes foi ter perdido dados no momento de exportar o arquivo. A solução que sugerimos ao Sponsor é modificar o formato do arquivo que foi enviado de XLS/XLSX para CSV.

Alguns dados da tabela de Produtos foram disponibilizados com as iniciais “ZZ” antes do nome do produto. Isso também pode ter sido um erro no momento da exportação dos dados, visto que o Sponsor consultou em seu sistema e não existem os nomes dos produtos dessa maneira. Além disso, muitos desses dados iniciados com “ZZ” estão com valores faltantes, o que ajuda a pensarmos que foi algum erro ao exportar. Os arquivos, ao serem disponibilizados em CSV, provavelmente irão solucionar este problema.

Além disso, como algumas das informações dos produtos são coletadas de API’s, alguns valores faltantes podem ter sido coletados da API. Portanto, pode ser verificado se existe outra API que nos informa os dados de forma mais completa.

Se o problema persistir poderemos tratar de alguma outra forma, assim que entendermos juntamente com o Sponsor o que pode ter ocasionado os valores faltantes.

1. DATA PREPARATION

3.1 Seleção dos Dados

No dataset bruto disponibilizado pelo cliente, contém dados dos clientes, dos produtos e da transação realizada.

Quanto melhor o perfil do cliente for detalhado mais assertividade o modelo terá, visto que os clusters serão melhor separados e a recomendação terá uma acuracidade maior. Portanto, o dataset de “Perfil de clientes” está bem completo e não podemos prever quais variáveis o modelo dará maior relevância para fazer os clusters. Porém, as variáveis “Age” (idade) e “YearBirthDate” (ano de nascimento) estão se referindo a mesma coisa, podemos comprovar isso na Figura 2 contendo a correlação entre elas de -1.

Com isso, decidimos por excluir a variável “Age”, já que é uma coluna que depende do dia que o modelo foi gerado, já o “YearBirthDate” é um valor constante, evitando algum tipo de erro.

O dataset de “Produtos” é um detalhamento do produto que foi feita a transação. Portanto, quanto maior o nível de detalhamento melhor para a clusterização. Visto isso, as variáveis “Sector” (setor do produto), “Industry” (indústria do produto) e “Group” (Grupo do produto) são as mais relevantes para o modelo. A variável “Sector” deve ser a mais relevante por se tratar de um nível maior de detalhamento que as outras duas, como pode ser visto nas figuras abaixo.

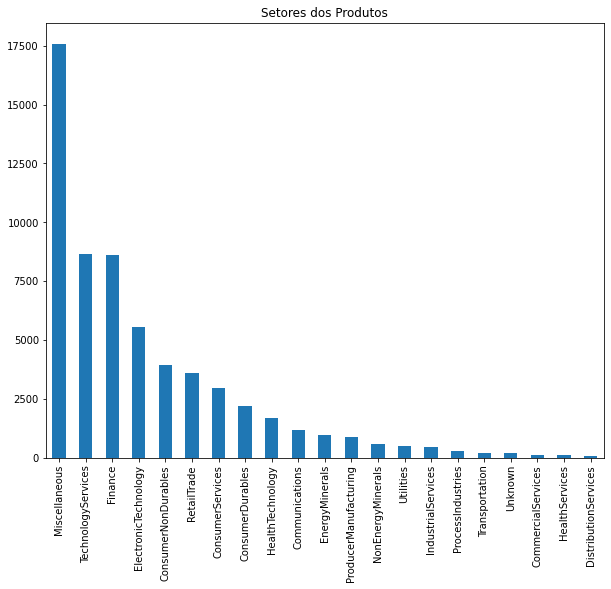


Figura 8: Setores dos Produtos

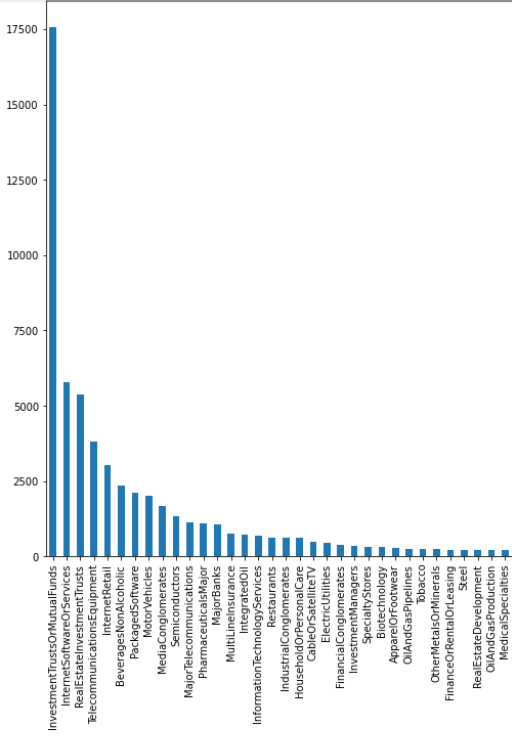


Figura 9: Principais Industrias dos produtos

Na Figura 9 estão representadas as principais industrias relacionadas aos produtos. Como há uma grande quantidade de industrias diferentes, foi plotado o gráfico de barras somente dos mais representativos, para uma melhor visualização.

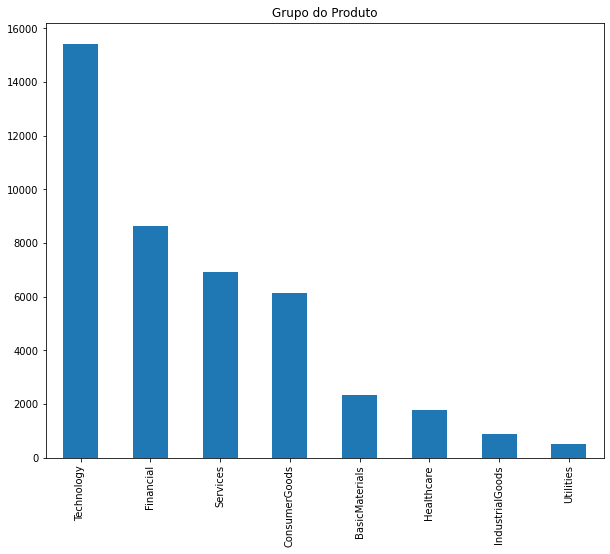


Figura 10: Grupos dos Produtos

O dataset de “Ordens” diz respeito aos dados da transação que foi realizada. A variável "PercCustody” (percentual do valor da ordem em relação a custódia total) é a mais relevante para o modelo.

|  |  |
| --- | --- |
| **Exclusão** | **Motivo** |
| Age | Não iremos utilizar a idade, pois já temos o ano de nascimento |

Tabela 9: Dados Excluídos

Os dados do dataset que foram excluídos estão na Tabela 9.

Além disso, o dataset foi filtrado para conter apenas registros de compra de produtos (são excluídos os dados de venda), visto que a recomendação será feita em relação às compras dos usuários.

3.2 Limpeza dos Dados

Os dados numéricos foram normalizados para uma melhor comparação entre as variáveis (mesma ordem de grandeza) e uma melhor visualização dos outliers. Foi definido, com o Sponsor, que os dados que estão acima de 3,5 vezes o desvio padrão serão excluídos inicialmente.

Uma outra possibilidade que temos é substituir os valores absurdos por dados do cluster que o cliente se encontra. Por exemplo a renda mensal de alguns clientes podem ter sido colocadas erradamente e alguns dados são considerados absurdos. Estes podem ser substituídos pela renda mensal dos outros clientes do seu cluster.

Esse método deve ser utilizado conforme o andamento do curso e estudarmos clusterização.

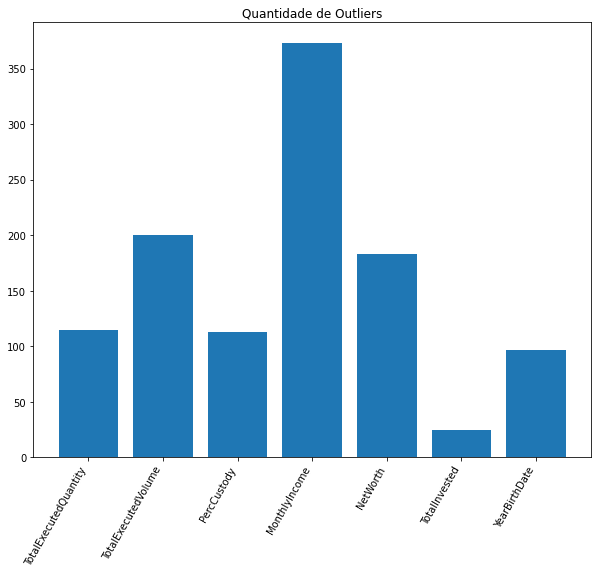


Figura 11: Quantidade de Outliers

A quantidade de dados excluídos por serem considerados outliers pode ser visualizada na Figura 12. Um total de 1106 dados outliers foram excluídos do dataset, o que representa 1,83% da massa de dados.

Espera-se que esse número reduza se substituir os dados de renda (“MonthlyIncome”) do cluster do cliente.

Como muitos dados da variável “Group” estão faltando e acreditamos que não seja necessário para o modelo, visto que a variável “Group” diz respeito à classificação mais grossa do produto, não detalhando muito o produto para fazer as recomendações. Com isso, foi decidido juntamente com o Sponsor que a variável “Group” será excluída.

3.3 Construção dos Dados

Os dados categóricos não serão tratados nesse momento, visto que não foi estudado o modelo de clusterização até o momento. Deixamos isso claro para o mentor e o Sponsor, pois alguns modelos podem ter variáveis categóricas, outros não. Além disso, alguns modelos são sensíveis ao se utilizar o OneHotEncoder nesse momento, outros não. Esta etapa, portanto, fica pendente nesse momento sendo extremamente necessária ser realizada posteriormente.

Uma variável que pode ser interessante para o modelo é atrelar a frequência que determinado cliente compra/venda o produto por grupo, setor e indústria. Isso pode ser interessante, visto que a recomendação para o cliente que compra/venda frequentemente determinado tipo de produto pode ser altamente impactante para os clusters.

|  |  |
| --- | --- |
| **Nova Coluna** | **Descrição** |
| FreqSector | Quantidade de compras que o cliente fez no Setor do produto dividido pelo total de compras, neste período |
| FreqIndustry | Quantidade de compras que o cliente fez no Industria do produto dividido pelo total de compras, neste período |
| FreqProduct | Quantidade de compras que o cliente fez do produto (contado pelo símbolo) dividido pelo total de compras, neste período |

Tabela 10: Novas variáveis criadas

Na Tabela 10 podemos observar as novas variáveis que foram criadas a partir de dados do próprio dataset. Elas dizem respeito da quantidade de compras/vendas que o cliente fez no setor, grupo ou indústria do produto que foi realizada a transação. Essa quantidade é em relação ao período de dados do dataset.

3.4 Integração dos Dados

O Sponsor solicitou que não houvesse integração de dados externos ao dataset. Ele julgou os dados suficientes para um bom modelo quando foi questionado sobre a integração dos dados.

3.5 Formatação dos Dados

Os dados disponibilizados pelo Sponsor precisam estar de acordo com a coluna “Tipo” das Tabelas 1, 2 e 3.

Com isso, inicialmente tivemos que modificar o formato data para que possa ser lido pelo Python como tal.

|  |
| --- |
| **Date** |
| 22 de jun. de 2022 |
| 26 de jul. de 2022 |
| 28 de jul. de 2022 |
| 20 de jan. de 2022 |
| 13 de jun. de 2022 |
| 5 de abr. de 2022 |
| 20 de jan. de 2022 |
| 5 de abr. de 2022 |
| 12 de set. de 2022 |
| 16 de fev. de 2022 |
| 27 de jan. de 2022 |

Tabela 11: Datas do dataset disponibilizados pelo Sponsor

|  |
| --- |
| **Date** |
| 2022-06-22 |
| 2022-07-26 |
| 2022-07-28 |
| 2022-01-20 |
| 2022-06-13 |
| 2022-04-05 |
| 2022-01-20 |
| 2022-04-05 |
| 2022-09-12 |
| 2022-02-16 |

Tabela 12: Datas com formato modificado

Na Tabela 11 podemos observar como as 10 primeiras datas foram disponibilizadas no dataset. Notamos que o mês está em português, “abr.” por exemplo teria que estar “apr” para que o Python consiga ler o mês. Portanto tivemos que fazer a tradução antes de modificar o formato da data para “datetime”

Além da variável “Date”, tivemos outras variáveis que precisaram ter seu tipo modificado.

|  |  |  |
| --- | --- | --- |
| **Variável** | **Tipo Antigo** | **Tipo Modificado** |
| MonthlyIncome | INT | FLOAT |
| NetWorth | INT | FLOAT |
| TotalInvested | INT | FLOAT |

Tabela 13: Tipos das variáveis

Na Tabela 13 vemos as transformações que foram necessárias para que as variáveis fiquem de acordo com a coluna “Tipo” das Tabelas 1, 2 e 3.

1. MODELING

4.1 Selecionando Técnicas de Modelagens

A técnica de modelagem que será utilizada nesse projeto é a de recomendação. Para usuários cold start estamos alinhando com o Sponsor, mas a tendência é que utilizaremos a clusterização dos clientes e posteriormente fazer uma recomendação dos top n itens.

Para fazer a recomendação, utilizaremos 5 técnicas para podermos comparar as métricas e escolher a melhor para o modelo. São elas: Content-Based, Top N consumidos, KNN, Co-visitation e SVD.

* Content-Based

É uma recomendação baseada em conteúdo, ou seja, tem foco nas features dos itens, encontrando itens similares.

É uma técnica que não deve ser boa para o projeto, visto que os produtos têm níveis de detalhamento muito baixo, sendo difícil ranquear os itens similares. Mesmo assim, rodaremos o modelo para confirmar nossa tese.

* Top N consumidos

Essa técnica leva em conta os top n itens consumidos no período estipulado, que neste caso será de 3 meses (tamanho do dataset disponibilizado).

É, portando, uma modelagem não-personalizada, ou seja, são recomendados os mesmos produtos para todos os usuários (não considera o perfil de consumo).

O Top N consumidos é uma ótima técnica de baseline para os modelos personalizados, isto é, um controle para saber se o modelo personalizado está performando bem

* KNN

O algorítmo item-KNN é de filtragem colaborativa baseado na similaridade de itens. Os algoritmos de filtragem colaborativa têm o foco nos feedbacks dos usuários.

É um modelo de parâmetros treináveis, ou seja, conseguimos separar o dataset em treino e teste para fazer a validação.

* Co-visitation

São recomendações em ofertas do tipo “Clientes que compraram o produto X também compraram y”. São caracterizados pela construção dos grafos entre os usuários e produtos.

* SVD (singular Value Decomposition)

É uma recomendação com fatoração de matrizes, procurando representar a matriz de preferências como uma multiplicação de diversas matrizes contendo fatores latentes dos usuários e itens.

Como em nosso projeto não temos feedback explícito (uma nota dada pelo usuário que mostraria seu sentimento sobre o produto) usaremos o feedback implícito de frequência de compra do usuário sobre cada produto. Ou seja, o feedback irá nos dar uma ideia do que cada usuário compra frequentemente, dando uma noção do que ele mais gosta de comprar, conseguindo fazer as recomendações.

Essa técnica foi necessária devido aos modelos de recomendação por filtragem colaborativa utilizarem o feedback dos usuários. Esses modelos tendem a ser mais interessantes para o nosso modelo, visto que para outros modelos deveríamos ter uma descrição melhor de cada produto.

4.2 Gerar design de Teste

Para testar o modelo, o dataset foi classificado por data e dividido em 2 datasets: 80% do para treinar o modelo; e 20% para teste. Ou seja, além da divisão dos dados, foi feita uma separação temporal.

O modelo não tem contato com o dataset de teste, portanto a assertividade que terá com o teste será uma medida válida da qualidade do modelo

4.3 Modelo de Construção

A recomendação de cada modelo foi treinada e podemos ver o resultado para um n=10, ou seja os 10 itens que serão recomendados.

* Content Based

Na tabela 14 podemos observar as recomendações do modelo contente based para a ação da Apple, ou seja, os produtos que são mais similares à Apple, de acordo com o detalhamento dos produtos.

Podemos observar que não um ranqueamento das similaridades, isso porque como a ação da Apple é um produto classificado como indústria “TelecommunicationsEquipment”, o modelo irá recomendar com o mesmo score (1,0) todos os produtos classificados como essa indústria, como esperado.

Como é uma recomendação que não é efetiva, não utilizaremos a Content-Based como modelo do projeto.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **id** | **score** | **Product** | **Product\_N3** | **Sector** | **Industry** |
| **SONM** | 1,00 | **SONIM TECHNOLOGIES INC** | Stocks | ElectronicTechnology | TelecommunicationsEquipment |
| **PWFL** | 1,00 | **PowerFleet Inc** | Stocks | ElectronicTechnology | TelecommunicationsEquipment |
| **AIRG** | 1,00 | **Airgain Inc** | Stocks | ElectronicTechnology | TelecommunicationsEquipment |
| **COMM** | 1,00 | **CommScope Holding Company, Inc.** | Stocks | ElectronicTechnology | TelecommunicationsEquipment |
| **CRNT** | 1,00 | **Ceragon Networks Ltd** | Stocks | ElectronicTechnology | TelecommunicationsEquipment |
| **AKTS** | 1,00 | **Akoustis Technologies, Inc.** | Stocks | ElectronicTechnology | TelecommunicationsEquipment |
| **SWIR** | 1,00 | **Sierra Wireless Inc.** | Stocks | ElectronicTechnology | TelecommunicationsEquipment |
| **CIEN** | 1,00 | **Ciena Corporation** | Stocks | ElectronicTechnology | TelecommunicationsEquipment |
| **AVYA** | 1,00 | **Avaya Holdings Corp** | Stocks | **ElectronicTechnology** | TelecommunicationsEquipment |
| **GILT** | 1,00 | **Gilat Satellite Networks Ltd** | Stocks | **ElectronicTechnology** | TelecommunicationsEquipment |

Tabela 14: Recomendações Content-Based

* Top N consumidos

Como n=10, serão recomendados os 10 produtos mais consumidos no período, como pode ser visto na tabela 15.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **item\_id** | **score** | **Product** | **Product\_N3** | **Sector** | **Industry** |
| **AAPL** | 2829 | **Apple, Inc.** | Stocks | ElectronicTechnology | TelecommunicationsEquipment |
| **AMZN** | 2065 | **Amazon.com Inc.** | Stocks | RetailTrade | InternetRetail |
| **GOOGL** | 1767 | **Alphabet Inc. - Class A Shares** | Stocks | TechnologyServices | InternetSoftwareOrServices |
| **VOO** | 1724 | **S&P 500 Vanguard ETF** | ETF's | Miscellaneous | InvestmentTrustsOrMutualFunds |
| **KO** | 1594 | **Coca-Cola Company, The** | Stocks | ConsumerNonDurables | BeveragesNonAlcoholic |
| **DIS** | 1418 | **Walt Disney Company, The** | Stocks | ConsumerServices | MediaConglomerates |
| **TSLA** | 1280 | **Tesla, Inc.** | Stocks | ConsumerDurables | MotorVehicles |
| **VNQ** | 1279 | **REIT Vanguard ETF** | ETF's | Miscellaneous | InvestmentTrustsOrMutualFunds |
| **MSFT** | 1231 | **Microsoft Corporation** | Stocks | TechnologyServices | PackagedSoftware |
| **FB** | 1005 | **Facebook, Inc.** | Stocks | TechnologyServices | InternetSoftwareOrServices |

Tabela 15: Recomendação Top 10 itens consumidos

* KNN

Na tabela 16, podemos observar as recomendações do modelo KNN para a ação da Apple. Neste caso, há recomendações de diferentes industrias e também variam o setor o tipo do produto. Isso porque o modelo KNN utiliza para recomendar os 10 vizinhos mais próximos da ação Apple, levando em conta os feedbacks dos usuários.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **id** | **Product** | **Product\_N3** | **Sector** | **Industry** |
| **BABA** | **Alibaba Group** | Stocks | RetailTrade | InternetRetail |
| **BXP** | **Boston Properties Inc.** | Stocks | Finance | RealEstateInvestmentTrusts |
| **IVV** | **Core S&P 500 iShares ETF** | ETF's | Miscellaneous | InvestmentTrustsOrMutualFunds |
| **MJ** | **ETFMG Alternative Harvest ETF** | ETF's | Miscellaneous | InvestmentTrustsOrMutualFunds |
| **VHT** | **Health Care Vanguard ETF** | ETF's | Miscellaneous | InvestmentTrustsOrMutualFunds |
| **PFE** | **Pfizer Inc.** | Stocks | HealthTechnology | PharmaceuticalsMajor |
| **QQQ** | **PowerShares QQQ** | ETF's | Miscellaneous | InvestmentTrustsOrMutualFunds |
| **O** | **Realty Income Corporation** | Stocks | Finance | RealEstateInvestmentTrusts |
| **SONY** | **Sony Corporation** | Stocks | ConsumerDurables | ElectronicsOrAppliances |
| **VT** | **Total World Stock Vanguard ETF** | ETF's | Miscellaneous | InvestmentTrustsOrMutualFunds |

Tabela 16: Recomendação KNN

Na tabela 17 temos os parâmetros utilizados no modelo KNN.

|  |  |
| --- | --- |
| **Parâmetros** | |
| **k** | 40 |
| **similarity** | pearson\_baseline |
| **user\_based** | False |

Tabela 17: Parâmetros do modelo KNN

* Co-visitation

Na tabela 18 podemos observar as recomendações do modelo Co-visitation para a ação da Apple. São produtos que “Quem comprou Apple, também comprou”. Os parâmetros utilizados, inicialmente, foram os default.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **id** | **Product** | **Product\_N3** | **Sector** | **Industry** |
| **GOOGL** | **Alphabet Inc. – Class A Shares** | Stocks | TechnologyServices | InternetSoftwareOrServices |
| **AMZN** | **Amazon.com Inc.** | Stocks | RetailTrade | InternetRetail |
| **AAPL** | **Apple, Inc.** | Stocks | ElectronicTechnology | TelecommunicationsEquipment |
| **KO** | **Coca-Cola Company, The** | Stocks | ConsumerNonDurables | BeveragesNonAlcoholic |
| **FB** | **Facebook, Inc.** | Stocks | TechnologyServices | InternetSoftwareOrServices |
| **MSFT** | **Microsoft Corporation** | Stocks | TechnologyServices | PackagedSoftware |
| **VOO** | **S&P 500 Vanguard ETF** | ETF’s | Miscellaneous | InvestmentTrustsOrMutualFunds |
| **TSLA** | **Tesla, Inc.** | Stocks | ConsumerDurables | MotorVehicles |
| **DIS** | **Walt Disney Company, The** | Stocks | ConsumerServices | MediaConglomerates |
| **XP** | **XP Inc.** | Stocks | TechnologyServices | InternetSoftwareOrServices |

Tabela 18: Recomendação Co-visitation

* SVD

Na tabela 19 podemos observar as recomendações do modelo SVD para o usuário “hvlfElsJbyBWfR9J7S08+g==”.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **user\_id** | **score** | **Product** | **Product\_N3** |
| **CDNS** | hvlfElsJbyBWfR9J7S08+g== | 0,459 | **Cadence Design Systems Inc.** | Stocks |
| **TFLO** | hvlfElsJbyBWfR9J7S08+g== | 0,436 | **iShares Treasury Floating Rate Bond ETF** | ETF's |
| **CCJ** | hvlfElsJbyBWfR9J7S08+g== | 0,435 | **Cameco Corporation** | Stocks |
| **IGR** | hvlfElsJbyBWfR9J7S08+g== | 0,429 | **CBRE Clarion Global Real Estate Income Fund** | Stocks |
| **UVXY** | hvlfElsJbyBWfR9J7S08+g== | 0,413 | **ProShares Ultra VIX Short-Term Futures ETF** | ETF's |
| **VIXY** | hvlfElsJbyBWfR9J7S08+g== | 0,395 | **VIX Short-Term Futures ProShares** | ETF's |
| **STNE** | hvlfElsJbyBWfR9J7S08+g== | 0,393 | **StoneCo Ltd.** | Stocks |
| **ISVL** | hvlfElsJbyBWfR9J7S08+g== | 0,383 | **iShares International Developed Small Cap Value Factor** | ETF's |
| **VT** | hvlfElsJbyBWfR9J7S08+g== | 0,379 | **Total World Stock Vanguard ETF** | ETF's |
| **SHOP** | hvlfElsJbyBWfR9J7S08+g== | 0,373 | **Shopify Inc.** | Stocks |

Tabela 19: Recomendação SVD

Na tabela 20 temos os parâmetros utilizados no modelo SVD.

|  |  |
| --- | --- |
| **Parâmetros** | |
| **lr** | 0,001 |
| **reg** | 0,005 |
| **n\_epochs** | 100 |
| **n\_factors** | 30 |
| **early\_stopping** | True |
| **min\_delta** | 0,0001 |
| **shuffle** | False |
| **min\_rating** | 0 |
| **max\_rating** | 1 |

Tabela 20: Parâmetros do modelo SVD

4.4 Avaliar Modelo

Para avaliar o modelo, vamos comparar os resultados obtidos pelos modelos SVD, KNN e Top N consumidos. O modelo de Top N consumidos serve como parâmetros para as métricas dos outros modelos.

As métricas que serão utilizadas para avaliar os modelos serão: Cobertura dos itens, Acurácia de Feedback, Acurácia de Classificação e Acurácia de Ranqueamento

* Cobertura dos itens

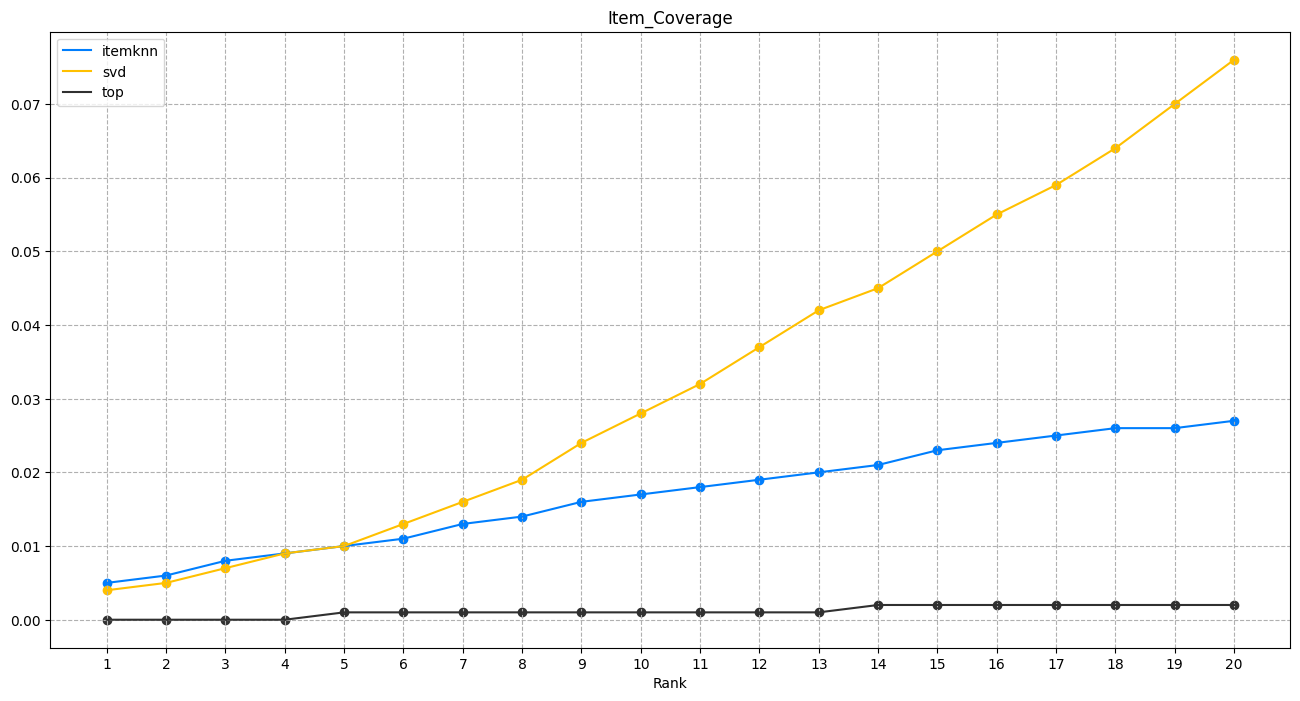


Figura 12: Cobertura dos itens

Pela Figura 12 podemos analisar a cobertura dos modelos para vários ranks (de 1 a 20), que é a quantidade de recomendações que serão disponibilizadas.

O valor da cobertura dos itens para o modelo Top é baixa devido ao fato de ser um modelo não-personalizado e todos os usuários recebem as mesmas recomendações. Os modelos SVD e KNN têm a cobertura aumentada conforme recomendamos mais itens, o que é natural, já que irá aumentar o número de produtos do catálogo que será recomendado.

* Acurácia de Ranqueamento

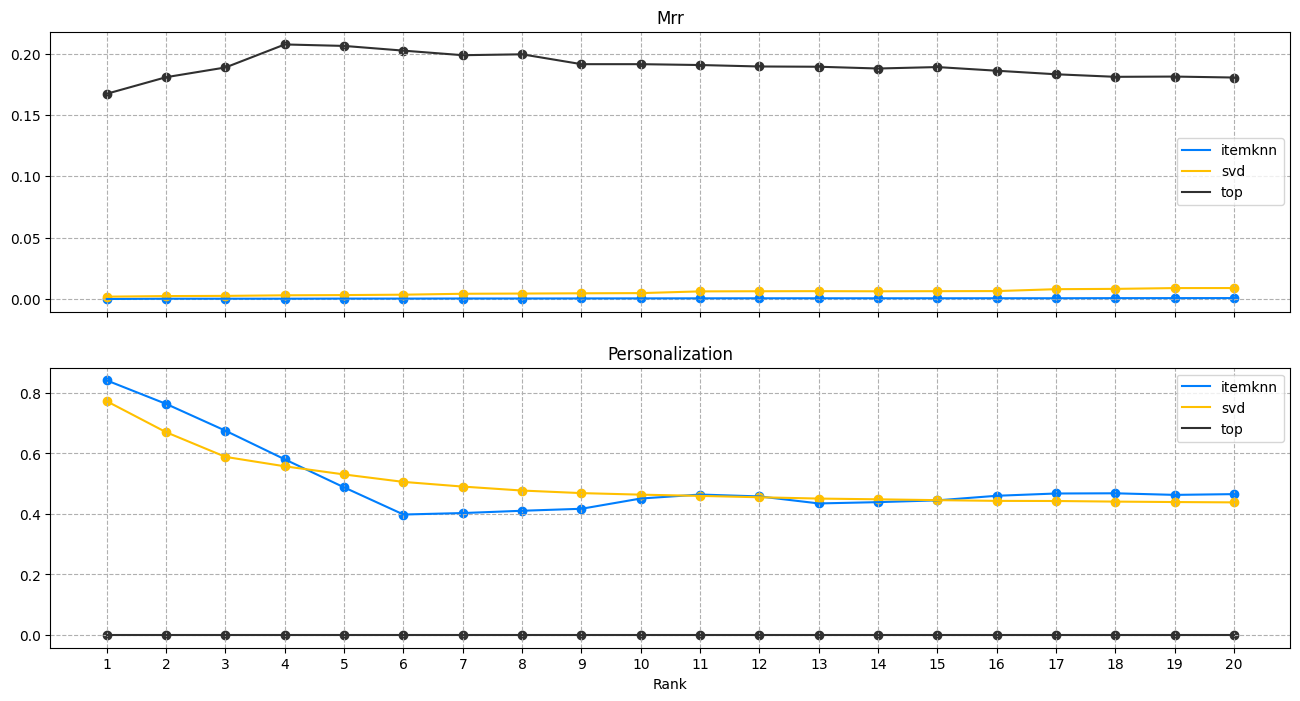


Figura 13: Acurácia de Ranqueamento

No gráfico superior da Figura 13 temos o MRR (mean reciprocal rank) onde ele varia de 0 a 1. Um valor igual a 1 significa que todos os itens na primeira posição de cada lista são relevantes; um valor igual a zero significa que nenhum item relevante está na lista de resultados.

O modelo top ter um MRR maior é natural por conter itens, na maioria, mais conhecidos, então podem ser mais procurados. Porém, consegue-se observar que a personalização do modelo top é quase nula, enquanto que os modelos SVD e KNN têm uma personalização alta, ressaltando a importância de se ter ofertas personalizadas para os usuários.

* Acurácia de Classificação

Nessa métrica, os modelos de recomendação são vistos como se fossem problemas de classificação, ou seja, montando uma matriz de confusão.

Alguns problemas dessa métrica são:

* + Não considera o ranking do item na lista, ou seja, se acertar um item na primeira ou vigésima posição tem a mesma efetividade
  + É uma métrica não derivável (não se consegue otimizar)

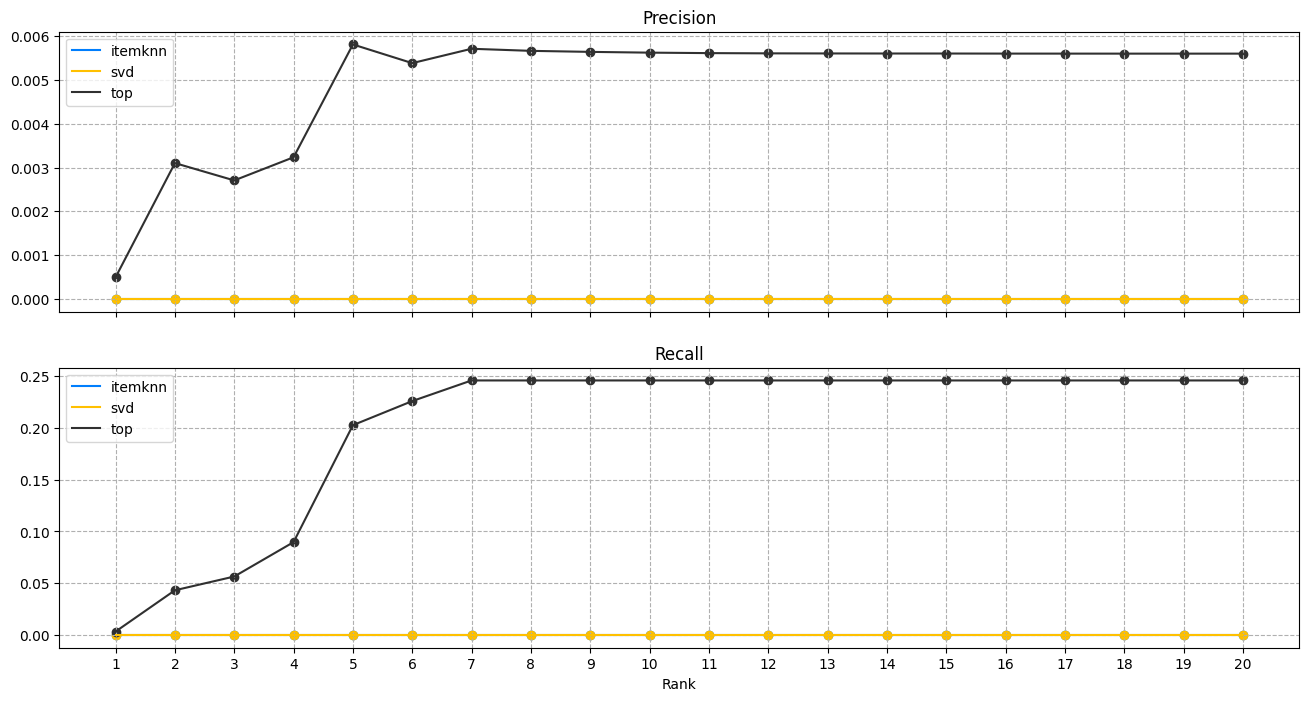


Figura 14: Acurácia de Classificação

Nessa métrica, podemos ver que os modelos SVD e KNN têm uma precisão muito baixa, e o modelo top tem uma precisão maior, por conta de ter na lista produtos mais conhecidos.

* Acurácia de Feedback

A Acurácia de Feedback mede o quanto os modelos estão acertando para um feedback que um usuário tem para um determinado item. Portanto, só serão considerados itens que o usuário comprou.

Alguns problemas dessa métrica são:

* + Não considerar o tamanho da lista recomendada
  + Não considera o ranking do item na lista recomendada
  + Depende do Feedback explícito do usuário, que é difícil de se obter

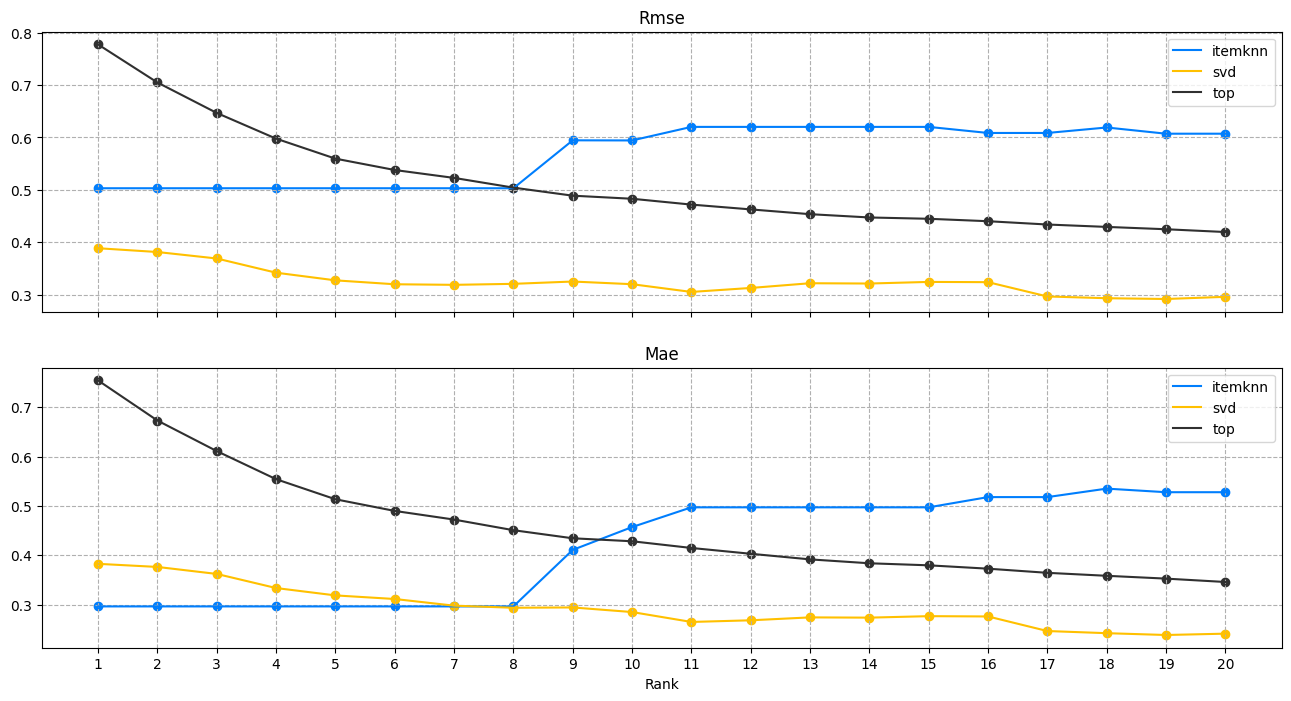


Figura 15: Acurácia de Feedback

Pelos gráficos da Figura 15, podemos concluir que os modelo SVD erra menos que o modelo top e KNN em relação ao feedback do usuário.

Pelas métricas que analisamos, podemos concluir que a precisão está muito abaixo do esperado. Isso pode ser devido a um erro de premissa adotado, como na construção dos feedbacks implícitos, o que será passado para o Sponsor e decidir o que pode ser feito a respeito.

Alterar alguns parâmetros dos modelos podem resultar em métricas melhores também.

Para o modelo Covisitation vamos utilizar a assertividade como métrica, ou seja, se o usuário comprou ou não o que está na lista de recomendados.

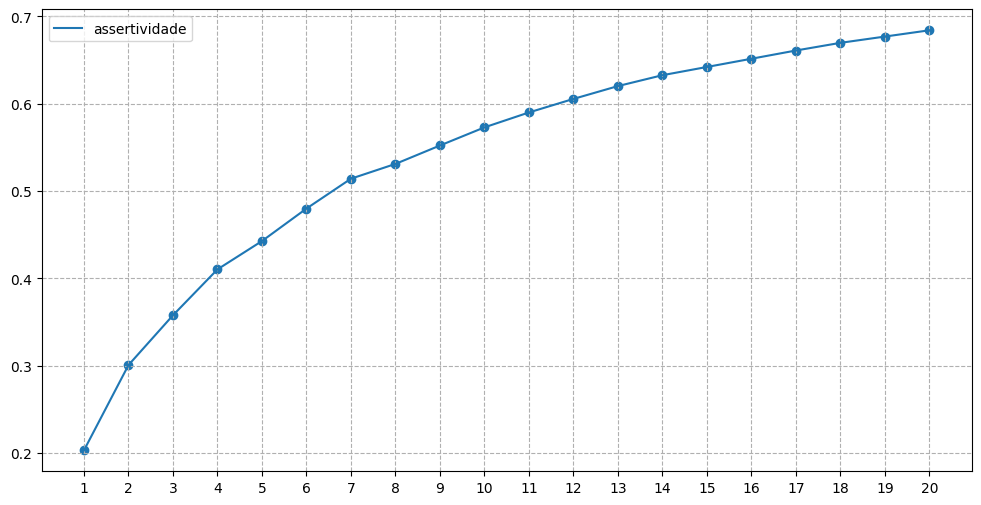


Figura 16: Assertividade do modelo Covisitation

Na figura 16 podemos observar a assertividade do modelo Covisitation variando o N, que é o número de recomendações. Portanto para um N = 10, temos uma assertividade de 58%. Isso quer dizer que para 58% dos usuários o modelo acerta qual o item que ele comprou.

Outra métrica que podemos estudar neste modelo é a cobertura do catálogo.

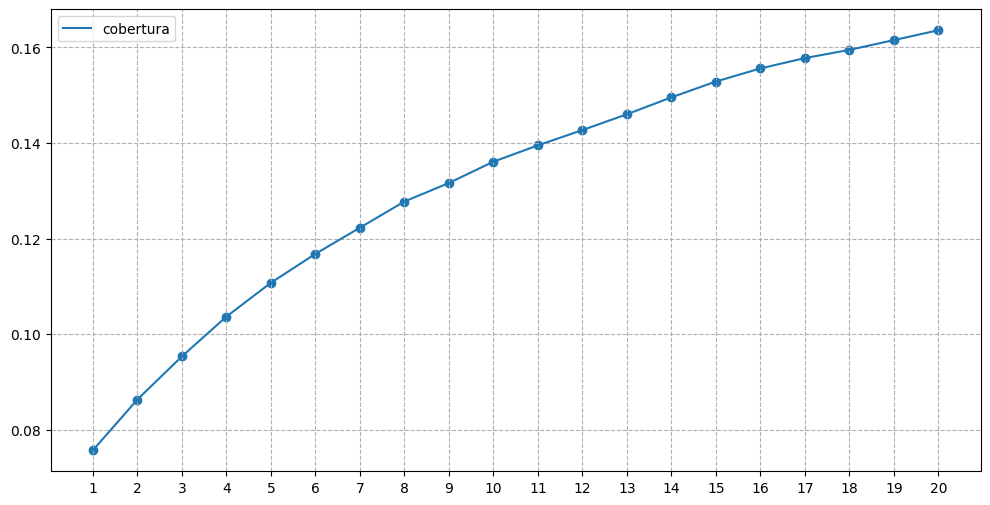


Figura 17: Cobertura do catálogo

Podemos observar na Figura 17 como varia a cobertura com o número de recomendações. Ou seja, para um N = 10 temos uma cobertura de 13% do catálogo. Isso quer dizer que 13% dos produtos estão contidos nas listas de recomendação dos usuários.

Portanto, o modelo Covisitation foi o que tivemos melhores resultados e é o que seguiremos no projeto.

1. EVALUATION

5.1 Avaliação dos Resultados

Os modelos de recomendação gerados atendem bem ao critério de sucesso de personalização, visto que os modelos gerados ficaram bem personalizados de acordo com o cliente. Para o critério de precisão, os valores foram baixos se comparados ao que estávamos esperando no início do projeto. Porém, como o próprio Sponsor nos disse, eles não têm noção de qual valor é considerada uma precisão alta ou baixa.

Para o modelo Content Based, o cliente fica numa bolha de produtos que são muito parecidos com os itens que ele comprou recentemente. Isso em alguns casos pode ser interessante, mas para nosso projeto não é, já que não queremos prender o cliente na bolha, e sim disponibilizar outros produtos que possam agradá-lo, de acordo com seu perfil. Além disso, o dataset tem um nível de detalhamento dos produtos muito baixo, dificultando o trabalho deste modelo. Uma saída seria buscar mais dados sobre o produto externamente, mas para esse projeto o Sponsor solicitou que trabalhássemos apenas com os dados que foram nos passados.

O modelo Top N consumidos é interessante mantermos na plataforma, mesmo não sendo um modelo personalizado. O modelo serve bem para os novos usuários, a quem não temos dados o suficiente ainda para fazer uma recomendação personalizada (cold-start de usuário)

Os modelos KNN e SVD alcançam uma alta personalização e cobre uma boa parcela do catálogo, mas as precisões foram extremamente baixas. Porém ainda estamos tentando melhorar os modelos com hiperparametrizações.

O modelo Co-visitation é o mais interessante para o negócio, visto que ele é uma recomendação do tipo “Quem comprou, também comprou”, nos dando uma alta personalização e também conseguimos cobrir uma alta porcentagem do catálogo de produtos.

5.2 Revisão do Processo

Algumas features que criamos no projeto não foram necessárias, sendo excluídas. São elas: freqIndustria e freqSetor. Elas não foram utilizadas porque para o modelo é mais interessante utilizar a frequência que o cliente comprar determinado produto (freqProduto)

A etapa de hiperparametrização ainda está sendo realizada, com a biblioteca Hyperopt, tanto para o modelo SVD quanto para o modelo KNN. Isso está sendo realizado afim de melhorar as métricas para esses modelos e fazer uma melhor escolha de qual modelo seguiremos no projeto. O Sponsor gostou do modelo Co-visitiation, mas temos a intenção de melhor os outros modelos para compará-los.

Um ponto que precisa ser refeito é sobre os outliers relacionados aos dados dos clientes. Eles não devem influenciar o modelo, visto que decidimos utilizar como feedback implícito a frequência que o usuário compra determinado produto. Portanto, não há a necessidade de retirá-los do modelo, devemos voltar a etapa de Data Preparation e modificar essa implementação.

Os processos em geral estão ocorrendo de forma correta para que tenhamos os melhores resultados possíveis para o modelo de recomendação, desde o entendimento do negócio, passando pelo entendimento e preparação dos dados, até a implementação e escolha do melhor modelo para o projeto.

5.3 Determinação dos Próximos Passos

Os próximos passos do projeto são:

* Retornar ao Data Preparation e tratar novamente os outliers
* Terminar a hiperparametrização para os modelos SVD e KNN, utilizando a biblioteca Hyperopt
* Escolher o melhor modelo de recomendação, que será utilizado no projeto

Com isso, o projeto ainda não está pronto para a implementação. As escolhas de parâmetros corretos fazem total diferença nas métricas e, com a hiperparametrização, nos guiará para a escolha do melhor modelo de recomendação e teremos um projeto para implementar em produção.

1. DEPLOYMENT

6.1 Plano de Implementação

A primeira etapa da implementação será para apresentação e testes internos do time de dados da Avenue.

O modelo de ML será aplicado em uma aplicação piloto desenvolvida com o auxilio do Streamlit. Neste ambiente o usuário terá a sua disposição 3 telas:

- Upload do dataset de produtos

- Upload do dataset de ordens

- Dashboard dinâmico que listará os clientes e permitirá executar a recomendação para o cliente selecionado. Como retorno, será exibida a listagem com as recomendações para aquele usuário específico.

Neste momento a eficácia do modelo será validada pelo time de dados conferindo a consistência das recomendações de forma manual.

A segunda etapa é servir um microsserviço de API REST, contendo os endpoints:

- Upload do dataset de produtos

- Upload do dataset de ordens

- Listagem de clientes

- Listagem de produtos

- Listagem de ordens

- Recomendador

Este serviço poderá rodar independente da aplicação principal da Avenue e poderá ser implementada no futuro pelo time desenvolvedores do APP da avenue, integrando o serviço de recomendação à interface do APP através da API REST.

Os resultados serão medidos baseados na aderência e aceitação do cliente, adquirindo algum dos produtos recomendados pelo modelo.

6.2 Plano de Monitoramento e Manutenção

O modelo deve ser integrado à interface do APP da Avenue conforme o desejo e estratégia da empresa para que o cliente final tenha acesso às recomendações.

Sugestão de implementação:

- Publicar uma aplicação/serviço que torne disponível em uma rede segura a API REST que expoe os endpoints do modelo.

- Com o modelo publicado e acessível, os desenvolvedores do APP da Avenue poderão implementar à interface do APP a melhor maneira de distribuição e acesso dos clientes aos resultados de recomendação.

- O cliente poderá, então, selecionar nesta interface o produto desejado, adquirindo-o em seu portfolio.

- Neste momento, recomenda-se registrar as compras realizadas nesta interface para que, posteriormente o time de dados possa coletar e analisar o comportamento de compra destes usuários, cruzando as informações de recomendação com as compras realizadas após as sugestão.

- A análise das compras não devem ficar restritas apenas à tela de recomendação, mas todas as compras posteriores à recomendação, uma vez que o cliente pode não comprar no primeiro momento, mas entende-se que houve uma “influência” anterior que pode ter levado à compra tardia. No entanto, deve-se dar uma nota ou gerar um registro diferenciado (como uma tag) que ajude a identificar se a compra foi realizada na tela de recomendação ou fora dela.

É fundamental monitorar o comportamento de compra dos usuários após as recomendações. Isso trará insumos para validar a assertividade (acuracidade) do sistema de recomendação e entender o quanto as recomendações estão sendo relevantes para a decisão de compra do cliente final.

Obtendo tais índices é possível medir e propor melhorias no modelo de recomendação.

Manutenção:

A manutenção do modelo de recomendação deve ser feita diretamente no código-fonte disponibilizado via GIT/GITHUB, mantendo o histórico de versão e controlando as alterações realizadas.

Pipeline de deploy:

É possível automatizar o processo de deploy do modelo em ambiente de testes (Streamlit). Tanto o Github quanto o Bitbucket possui um automatizador de CI/CD capaz de gerar o pkl e instanciá-lo na aplicação Streamlit. Isso acelerá o processo de validação das alterações antes de colocá-lo em produção.

Gitflow e CRISP-DM:

Para as futuras melhorias do modelo de recomendação é recomendável serguir o fluxo de implementação do CRISP-DM e para controle do versionamento e publicação do modelo utilizar o GITFlow. Desta forma, tanto a fase de pesquisa (CRISP-DM) quanto de modelagem e deploy( (GITFLOW) terão procssos claros e bem definidos que garanter a evolução segura do serviço.

Produção

Para o deploy da aplicação recomenda-se o uso do Amazon Lambda ou Google Cloud Platform (GCP) para servir o serviço da API Rest. Como medida de segurança, o acesso à API deve ser provida através de uma VPN onde apenas a rede de serviços da Avenue tenha acesso, uma vez que a API não possui nenhum serviço de autenticação entre as aplicações, sendo assim, a maneira mais eficiente de utilizar o serviço imediatamente, com o nível adequado de segurança é através da VPN privada.

6.3 Relatório Final

O projeto teve como objetivo fazer um sistema de recomendação para os clientes, de maneira personalizada, ou seja, cada cliente terá a sua própria recomendação de acordo com os gostos do cliente.

Inicialmente tanto nós quanto o Sponsor não tínhamos uma noção exata de um critério de sucesso, mas foi estipulado um valor de assertividade de aproximadamente 70%, mas que poderia ser alterado de acordo com a realidade encontrada.

Foi disponibilizado um dataset com dados dos clientes, produtos e das transações no período de 3 meses. Com isso, conseguimos ter uma noção de qual tipo de ativo era mais negociado e quais os perfis dos clientes da empresa.

Utilizamos alguns modelos diferentes para comparar de decidir qual será utilizado no projeto, entre eles: ItemKNN, SVD, Content Based e Covisitation. O modelo quer performou melhor foi o de Covisitation, que é do tipo “Quem comprou X também comprou Y” e foi o que decidimos utilizar no projeto.

Com ele, obtivemos uma assertividade de 58% e uma cobertura do catálogo de 13%, utilizando 10 recomendações por usuário.

Portanto, obtivemos uma assertividade alta e uma porcentagem do catálogo sendo recomendada relativamente alta. A cobertura é algo extremamente importante para esse negócio, visto que quanto maior mais chance o usuário tem de conhecer produtos novos e negociá-los.

6.4 Revisão do Projeto

O projeto foi bem sucedido, com os resultados que superaram as expectativas, tanto do Sponsor quanto do grupo. Inicialmente não tínhamos noção (nem o grupo, e nem o Sponsor) de quanto teríamos de assertividade, e conseguimos um valor muito bom.

Com o andamento do projeto vimos que não só a assertividade era um fator importante, mas também a cobertura do catálogo ser grande. Isso faz com que os usuários tenham contato com outros produtos que anteriormente não eram conhecidos.

O modelo obteve um aumento significativo da cobertura, se comparado com o sistema de recomendação adotado anteriormente pela empresa.

O que pode ser melhorado no modelo é a maneira com que se constrói o feedback implícito do usuário. Testamos várias maneiras diferentes e selecionamos a melhor, mas pode ser ainda mais aprimorado. Outro ponto de melhoria do projeto é o usuário que ainda não realizou nenhuma compra com a empresa, ou seja, novos usuários. Pode ser personalizado de acordo com o perfil do cliente.

Aprendemos muito com o projeto, desde a implementação em si, até a forma de lidar com problemas que surgiram no andamento e comunicação com o Sponsor. Acreditamos que o projeto foi de extrema importância para nós e para o Sponsor também, que sempre nos deixou claro a intenção de aprender sobre Modelo de Recomendação junto conosco.